

Countering Rumours in Online Social Media

Ebrahimi Fard, A.

DOI

[10.4233/uuid:bf835c87-da7b-4dd7-bfad-41fd1bb537c0](https://doi.org/10.4233/uuid:bf835c87-da7b-4dd7-bfad-41fd1bb537c0)

Publication date

2021

Document Version

Final published version

Citation (APA)

Ebrahimi Fard, A. (2021). *Countering Rumours in Online Social Media*. [Dissertation (TU Delft), Delft University of Technology]. <https://doi.org/10.4233/uuid:bf835c87-da7b-4dd7-bfad-41fd1bb537c0>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

COUNTERING RUMOURS IN ONLINE SOCIAL MEDIA

COUNTERING RUMOURS IN ONLINE SOCIAL MEDIA

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, Prof. dr. ir. T. H. J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Tuesday 9 March 2021 at 15:00 o'clock

by

Amir EBRAHIMI FARD

Master of Science in Management and Economics,
Sharif University of Technology, Iran
born in Qom, Iran.

This dissertation has been approved by the promotor.

Composition of the doctoral committee:

Rector Magnificus, Prof. dr. B.A. van de Walle	Chairperson Delft University of Technology and UNU-MERIT, promotor
Prof. dr. D. Helbing	Delft University of Technology and ETH Zurich, promotor
Dr. ir. T. Verma	Delft University of Technology, copromotor

Independent members:

Prof. dr. M.J. van den Hoven	Delft University of Technology
Prof. dr. ir. A. Bozzon	Delft University of Technology
Prof. dr. H. Alani	The Open University
Dr. ir. I. Lefter	Delft University of Technology



Keywords: Rumours, social media, recommender systems, counter-strategies, one-class classification, social manipulation.

Printed by: Gilderprint

Cover design: Amir Ebrahimi Fard (Based on a design first published in the "Die Karikatur und Satire in der Medizin: Medico-Kunsthistorische Studie von Professor Dr. Eugen Holländer, 2nd edn (Stuttgart:Ferdinand Enke, 1921), fig. 79 (p. 171)." (The original design is in the public domain))

© Copyright Amir Ebrahimi Fard, 2021. All rights reserved.

ISBN 978-94-6419-147-9

An electronic version of this dissertation is available at
<http://repository.tudelft.nl/>.

The research in this dissertation was funded by Delft University of Technology.

To Maman o Baba

ACKNOWLEDGEMENTS

*This book has come to an end
(but) the story yet remains*

Sa'di

Although my name is written as the author of this manuscript, my PhD was a #collective_effort, and a stellar group of people 🙌🏆🏆 tremendously supported me.

My deepest gratitude goes to @Maman, @Baba, and @Reza, three gems in my life. Thank you for your eternal encouragement and endless support in everything I do. You always pushed me to become the best version of myself, remain #hopeful, and face challenges. I learned #patience, #perseverance, and #dedication from you. Thank you for always being there for me ❤️💛💜.

To @Bartel, @Dirk, @Trivik, and @Scott, I was privileged to do my PhD under your supervision. You lit up my research path and helped me to #grow in academia and become an #independent_researcher. You gave me the #freedom to #explore from Coase theorem and industrial districts to rumour spreading. You always encouraged me to choose a research topic that I am really attached to. I would not stand where I am now without your support. My sincere and heartfelt gratitude and appreciation to you for providing me with the guidance and counsel I needed to succeed in my PhD 😊🙏. @Bartel, you were not only my promotor, you were also my #role_model and the source of #inspiration during my PhD journey, I will miss being your student. @Dirk, every time you came up with an idea, you blew my mind. I wish my PhD days had more than 24 hours to spend on those #brilliant_ideas. @Trivik, I cannot emphasise how much I learned from you; whether about #technical or #soft_subjects. I am very grateful for all this support. @Scott, you helped me to start my PhD and formulate my research. You also provided me with several great #collaboration_opportunities. Thank you so much for all of them.

How could I possibly finish my PhD without my #fantastic_friends 😊 @Natasa, @João, @Annebeth, @Majid, @Javanshir, @Vivian, @Farzam, @Sharlene, @Shantanu, @Arthur, @Vittorio, @Ioanna, I was extremely fortunate that my path crossed yours. @Natasa and @Annebeth, you are the yardstick of #friendship. I am so glad that we shared an office, a neighbourhood, and a PhD graduation team 😊. @Majid, I am very lucky to be a friend of a bright yet humble person like you. I truly appreciate all your support during my PhD journey.

Perhaps if I want to continue this acknowledgement letter, it will not end anytime soon, as many people kindly supported me in this journey. By all means, thank you 😊.

Amir Ebrahimi Fard

Delft, March 2021

CONTENTS

Acknowledgements	vii
List of Tables	xiii
List of Figures	xv
Summary	xix
Samenvatting	xxi
1 Introduction	1
1.1 An Overview on the Phenomenon of Rumour Spreading	2
1.2 Research Objective and Research Questions	3
1.3 Contributions and Guide to Readers	6
1.4 Engineering Social Technologies for a Responsible Digital Future.	8
2 Conceptualisation of False and Unverified Information	11
2.1 Introduction	12
2.2 The variations of false and unverified information	12
2.2.1 Rumour	12
2.2.2 Gossip	13
2.2.3 Legend.	14
2.2.4 Propaganda	14
2.2.5 Conspiracy Theory.	15
2.2.6 Fake-news	15
2.2.7 Pseudoscience	16
2.2.8 Misinformation	17
2.2.9 The Comparison of False and Unverified Information	17
2.3 Process-based Perspective	17
2.4 What has to be curbed?	19
2.5 Conclusion	20
3 The Landscape of Rumour Spreading	21
3.1 Introduction	22
3.2 Communication and Rumour Spreading	22
3.3 The Role of Communication technologies in rumour spreading	24
3.4 Recommendation systems under Scrutiny	26
3.4.1 Methodology and Data Collection	26
3.4.2 Analysis	29
3.5 Conclusion	38

4	Countering rumours	39
4.1	Introduction	40
4.2	Counter Rumour Strategies	40
4.2.1	Senders Strategies	42
4.2.2	Channel Strategies	42
4.2.3	Receivers Strategies	45
4.3	Evaluation of Strategies	48
4.3.1	Evaluation framework	48
4.3.2	Strategies effectiveness.	49
4.4	Conclusion	50
5	An Assessment of Academic Efforts regarding Rumour Confrontation	53
5.1	Introduction	54
5.2	Scientific emergence	54
5.3	Method	55
5.3.1	Data collection.	56
5.3.2	Emergence operationalization	60
5.4	Results	60
5.4.1	Novelty.	60
5.4.2	Growth.	65
5.4.3	Coherence	69
5.4.4	Impact	77
5.5	Discussion	81
5.6	Conclusion	83
6	Computational Rumour Detection using One-Class Classification	85
6.1	Introduction	86
6.2	Computational Rumour detection	88
6.3	Data.	90
6.3.1	Building the datasets.	90
6.3.2	Available datasets	93
6.4	Feature Extraction	95
6.4.1	Linguistic & content features.	98
6.4.2	User features.	101
6.4.3	Meta-message features.	102
6.5	Classification	102
6.5.1	Problem Statement	102
6.5.2	One-class Classification Approach	106
6.5.3	Experiments	110
6.6	Conclusion	120
7	Modelling Rumour Campaigns: A Proactive Approach	123
7.1	Introduction	124
7.2	Research Background	125
7.2.1	Deliberate Rumour Spreading as a Means of Information Operation.	125
7.2.2	Misinformation Machine	126
7.2.3	Cyber Operations Model	127

7.3	Model development.	127
7.3.1	The Block Diagram.	130
7.3.2	The Operationalisation of the Model.	130
7.3.3	The Data Model	137
7.4	Model Evaluation	140
7.4.1	Expert-based evaluation	140
7.4.2	Evaluation through exemplification	141
7.5	Conclusion	148
8	Discussion & Conclusion	151
8.1	Societal Relevance	154
8.2	Reflection and Future Research.	154
A	Appendix	159
A.1	Chapter 5	159
A.2	Chapter 7	164
A.2.1	Data Model	164
A.2.2	Model Evaluation	173
A.2.3	Interview Setting.	176
	List of Publications	197

LIST OF TABLES

2.1	Comparison between different forms of false and unverified information ([7]).	18
3.1	Representative titles from all six categories.	30
3.2	Left: $p(\text{Max} \mid \text{Group}) / p(\text{Max})$, bounded at 1. Y-axis shows top 5 topic words. Right: Longer representations of topics.	33
3.3	Summary statistics for each topic. All figures report averages (means). . .	34
3.4	Ratio of conspiratorial clips to rated clips for each category.	36
4.1	Analysis of the quelling strategies against epidemic control framework. . .	51
5.1	Technology emergence dimensions [188]	55
5.2	Comparison between three major databases of indexing bibliometrics data [194].	58
5.3	Queries for data collection from Web of Science	60
5.4	Operationalisation of emergence framework. The new criteria are marked with †.	61
5.5	Special issues in the field of rumour studies between 2000 and 2018.	75
5.6	The conferences in the field of rumour studies between 2000 and 2018. . .	76
5.7	The schematic dynamic of emergence dimensions in the field of rumour studies	81
6.1	The statistical information regarding Zubiaga [204] and Kwon [202] datasets.	95
6.2	The PoS tags and their description.	99
6.3	The NER tags and their description.	100
6.4	Comparison between multi-class classification and one-class classification. . .	106
6.5	Confusion matrix for one-class classification [212].	111
6.6	Baseline analysis on the Zubiagaset and Kwonset [204, 202]. We could not apply SVDD on the whole Kwonset since the standard solver of SVDD does not suit the large-scale datasets. We tackled this problem by subsampling the training set and experiment with a subset of the original dataset. . . .	112
6.7	The classifiers hyper-parameters and their valid range.	115
7.1	The explanations of MCOM classes [244].	128
7.2	Graphical representations of the model.	140
7.3	Model exemplification	144
A.1	Queries for data collection from Web of Science	160

A.2	Classes of the model (E).	165
A.3	Attributes/Data properties of the model (D).	167
A.4	Relations/Object properties of the model (R).	169
A.5	The list of news outlets that published Heshmat Alavi's articles.	174

LIST OF FIGURES

3.1	Communication process [85].	23
3.2	The schematic flow of data collection.	27
3.3	The YouTube recommendation tree when all the recommendations are distinct. In the case of the same recommended videos, the structure will be a directed graph.	28
3.4	Similarity of recommendations across topics.	31
3.5	Similarity of recommendations across search terms.	32
3.6	Distribution of conspiracy theories among the most-recommended clips from each topic. 1 = no conspiracy theory, 2 = mild conspiracy theory, 3 = severe conspiracy theory, x = clip no longer available at time of coding. . .	35
3.7	Fraction of top-recommended videos discovered at each stage of data collection.	37
4.1	Rumour counter strategies overview	41
4.2	The quelling strategies for a rumour responsible party [170].	46
5.1	Different phases and attributes of emergence [188].	56
5.2	The method of assessing the readiness of the academia in the field of rumour studies.	57
5.3	Data collection and data filtering steps	59
5.4	Schematic view of comparison table	63
5.5	Change in novelty level in the field of rumour studies	64
5.6	Growth and its composition in the field of rumour studies	66
5.7	The composition of different communities in the underpinning disciplines network within 1900 to 2018.	67
5.8	The yearly contribution of newcomers to the field of rumour studies . . .	68
5.9	The composition of research communities in five periods of 1900 ~ 1980, 1980 ~ 1990, 1990 ~ 2000, 2000 ~ 2010, and 2010 ~ 2018. Every bar denotes one community and different colours in each bar represent contributing research areas in the corresponding community. The length of each bar displays the number of subject categories in its community.	72
5.10	Co-occurrence of disciplines in the field of rumour studies	74
5.11	Theme significance of different research areas. To save the space, the following abbreviations are used: LS = Life Sciences, AH = Art and Humanities, CPH = Clinical, Pre-Clinical and Health, SS = Social Sciences, ET = Engineering and Technology, PS = Physical Sciences.	75
5.12	The growth of densification in author level	77

5.13	Assessment of impact using academic disciplines contribution to the field of rumour studies	79
5.14	Expectation analysis for funding acknowledgements and funding agencies in the field of rumour studies	80
5.15	The community formation in a research field.	82
6.1	The research flow of rumour detection with one-class classification approach.	87
6.2	Rumour resolution system has four modules: (i) rumour detection for identifying rumour related information; (ii) rumour tracking for collecting the posts discussing the rumour; (iii) stance classification for determines posts' orientations toward rumours' veracity; and (iv) veracity classification for verification of truth behind the rumour [206].	88
6.3	The methodology of building a dataset for computational rumour detection.	90
6.4	The number of publications regarding 11 popular social media platforms from 2008 to 2018 based on Scopus data. This figure illustrates the growing trend of using social media data by researchers. As the figure shows, scholars tend to work with Twitter data more than other platforms.	92
6.5	Categorisation of features for computational rumour detection.	97
6.6	The dependency tree of "I just tried cooking popcorn with 4 mobile phones its a lie I tell you A LIE".	98
6.7	Schematic description of two primary perspectives toward non-rumour. In both diagrams, squares with border show different events. Also, yellow and blue are denote rumour and non-rumour area respectively. In this figure, size does not mean anything and cannot be a basis for comparison.	104
6.8	Chain of the reasoning behind the problematic consequences of non-rumour in binary classification. It starts with a lack of sufficient theoretical background for the concept of non-rumour. It leads to the emergence of ambiguous and contradictory definitions of non-rumour. Lack of clear definitions causes data annotation to be done arbitrarily, which makes the rumour classifier unreliable (it is not clear, what it separates) and incomparable (it is not possible to compare the results of different classifiers).	105
6.9	Categorisation of one-class classification algorithms.	107
6.10	The impact of training sample size on the performance of classifiers in the Zubiagaset and Kwonset. The horizontal axis displays different classifiers, and the vertical axis shows their F1 score.	113
6.11	The impact of hyper-parameters on models performance in the Zubiagaset and Kwonset.	117
6.12	The classifiers performance in different feature categories in the Zubiagaset and Kwonset.	119
6.13	The execution time of classifiers in the Zubiagaset and Kwonset.	120
7.1	The flow of building DRSM. The bidirectional arrows show iteration between the phases.	125
7.2	The Misinformation machine model.	126
7.3	Maathuis Cyber Operation Model (MCOM) [244].	129

7.4	The block diagram of DRSM.	130
7.5	The component of goal definition and target selection.	132
7.6	The sub-component of capability development (the highlighted part). . .	133
7.7	The sub-component of message implantation.	134
7.8	The sub-component of impact assessment.	135
7.9	Model architecture design – logical flow of the deliberate rumour spreading in social media.	136
7.10	Global view of the model.	139
7.11	Data global view for Alavi rumour campaign.	148

SUMMARY

The phenomenon of rumour spreading refers to a collective process where people participate in the transmission of unverified and relevant information to make sense of the ambiguous, dangerous, or threatening situation. The dissemination of rumours on a large scale no matter with what purpose could precipitate catastrophic repercussions. This research aims at addressing this challenge systematically. More in detail, the primary research objective of this dissertation is

To systematically study the rumour confrontation within online social media.

To accomplish this objective, six steps are taken. At first, the conceptualisation of the main construct in this research is investigated. There are myriad of concepts in English language implying false or unverified information. However, despite years of academic research, there is no consensus regarding their conceptualisation, and they are often used interchangeably or conflated into one idea. This problem could become an obstacle to countering the surge of false information by creating confusion, distracting the community's attention, and draining their efforts. In the first step, this dissertation addresses this challenge by providing a process-based reading of false and unverified information. This view argues that although the genesis of such information might be deliberate or inadvertent and with different purposes, they primarily disseminate on the basis of similar motives and follow the same process.

After settling the conceptualisation problem, the next step investigates the role of communication mediums and especially online social media in the spread of rumours. Although the phenomenon of rumour dissemination has drawn much attention over the past few years, it is an ancient phenomenon. The rumours used to circulate through primitive forms of communications such as word of mouth or letters; however, the technological development, particularly social media, escalated the scale, speed, and scope of this phenomenon. This step aims to pinpoint the features privy to social media that facilitate the emergence and the spread of rumours. Especially, an exclusive automation mechanism of recommendation systems in social media is closely examined through a set of experiments based on YouTube data.

The third step in this study investigates the constellation of past counter-rumour strategies. Although rumour spreading and its potentially destructive effects have been taken into account since ancient times, it was only less than a century ago that the first systematic efforts against the mass spread of rumours began. Since then, a series of strategies have been practised by various entities; nevertheless, the massive waves of rumours are still sweeping over individuals, organisations, and societal institutions. In order to develop an effective and comprehensive plan to quell rumours, it is crucial to be aware of the past counter strategies and their potential capabilities, shortcomings and flaws. In this step, we collect the counter strategies over the past century and set them

in the epidemic control framework. This framework helps to analyse the purpose of the strategies which could be (i) exposure minimisation, (ii) immunisation or vaccination, and (iii) reducing the transmission rate. The result of the analysis allows us to understand, what aspects of confrontation with rumour have been targeted extensively and what aspects are highly neglected.

Following the discussion on the epidemic framework, one of the most effective approaches to rumour confrontation is the immunisation which is primarily driven by academia. The fourth step investigates the readiness of academia in this subject domain. When we do not know the readiness level in a particular subject, we either overestimate or underestimate our ability in that subject. Both of these misjudgements are incorrect and lead to decisions irrelevant to the existing circumstance. To tackle this challenge, the technology emergence framework is deployed to measure academia's readiness level in the topic of rumour circulation. In this framework, we study four dimensions of emergence (novelty, growth, coherence and impact) over more than 21,000 scientific articles, to see the level of readiness in each dimension. The results show an organic growth which is not sufficiently promising due to the surge of rumours in social media. This challenge could be tackled by creating exclusive venues that lead to the formation of a stable community and realisation of an active field for rumour studies.

The other aspect of the epidemic framework involves exposure minimisation and transmission rate reduction, which are addressed in the fifth step by an artificial intelligence based solution. The drastic increase in the volume, velocity, and the variety of rumours entails automated solutions for the inspection of circulating contents in social media. In this vein, binary classification is a dominant computational approach; however, it suffers from non-rumour pitfall, which makes the classifier unreliable and inconsistent. To address this issue a novel classification approach is utilised which only uses one rather than multiple classes for the training phase. The experimentation of this approach on two major datasets shows a promising classifier that can recognise rumours with a high level of F1-score.

The last step of this manuscript approaches the topic of rumour confrontation from a pro-active perspective. The epidemic framework helps to develop solutions to control rumour dissemination; however, they mostly adopt a passive approach which is reactive and after-the-fact. This step introduces an ontology model that can capture the underlying mechanisms of social manipulation operations. This model takes a proactive stance against social manipulation and provides us with an opportunity of developing preemptive measures. The model is evaluated by the experts and through exemplification on three notoriously famous social manipulation campaigns.

SAMENVATTING

De verspreiding van geruchten is een collectief proces waarbij mensen meedoen aan de overdracht van ongecontroleerde en relevante informatie om een ambigue, gevaarlijke of bedreigende situatie te begrijpen. Verspreiding van geruchten op grote schaal, met welk doel dan ook, kan catastrofale gevolgen hebben. Dit onderzoek heeft tot doel om dit probleem systematisch aan te pakken. Preciezer gezegd, het belangrijkste onderzoeksdoel van dit proefschrift is

Systematische bestudering van geruchtenbestrijding binnen online sociale media.

Hiertoe hebben we zes stappen gezet. Eerst onderzoeken we de conceptualisering van het hoofdconcept van het onderzoek. In het Engels zijn er vele manieren om uit te drukken dat informatie onjuist of ongecontroleerd is.

Ondanks jarenlang academisch onderzoek is er echter geen consensus over de conceptualisering ervan en worden de termen vaak door elkaar gebruikt of op één hoop gegooid. Dit kan een probleem zijn als we de vloed van valse informatie willen tegengaan door verwarring te creëren, de aandacht van de gemeenschap af te leiden en hun inspanningen teniet te doen. In de eerste stap pakken we in dit proefschrift dit probleem aan door valse en ongecontroleerde informatie procesmatig te interpreteren, op basis van het idee dat dergelijke informatie weliswaar zowel weloverwogen als per ongeluk de wereld in wordt gebracht, en met verschillende doeleinden, maar dat de verspreiding voornamelijk op basis van vergelijkbare motieven plaatsvindt, volgens hetzelfde proces.

Na het conceptualiseringsprobleem onderzoeken we in de volgende stap de rol van communicatiemiddelen en met name online sociale media bij de verspreiding van geruchten. Hoewel de verspreiding van geruchten de laatste jaren veel aandacht trekt, is het een oeroud fenomeen. Vroeger deden geruchten de ronde via primitieve communicatievormen, zoals mondeling contact of brieven; door technologische ontwikkelingen, met name sociale media, zijn de schaal, snelheid en omvang van dit fenomeen inmiddels enorm toegenomen. In deze stap willen we de specifieke kenmerken van sociale media aanwijzen die bijdragen aan het ontstaan en de verspreiding van geruchten. In het bijzonder onderzoeken we nauwgezet een exclusief automatisch mechanisme van aanbevelingssystemen in sociale media door middel van een reeks experimenten op basis van YouTube-data.

In de derde stap onderzoeken we welke strategieën men vroeger hanteerde in de bestrijding van geruchten. Hoewel men al sinds de oudheid onderkent dat de verspreiding van geruchten vernietigende effecten kan hebben, begon men pas een kleine eeuw geleden voor het eerst systematisch iets te doen tegen de massale verspreiding van geruchten. Sindsdien zijn er door verschillende instanties diverse strategieën in praktijk gebracht, maar nog steeds worden personen, organisaties en maatschappelijke instellingen geteisterd door grote golven van geruchten. Als we een effectief en breed toepasbaar

plan willen ontwikkelen om geruchten de kop in te drukken, is het essentieel dat we op de hoogte zijn van tegenstrategieën uit het verleden en dat we weten wat deze wel en niet wisten te bewerkstelligen. In deze stap verzamelen we de tegenstrategieën van de afgelopen eeuw en vergelijken we deze met de bestrijding van een epidemie. Dit kader helpt bij de analyse van drie mogelijke strategieën: (i) minimalisering van de blootstelling, (ii) immunisatie of vaccinatie en (iii) vermindering van de overdrachtssnelheid. Door deze analyse zien we op welke aspecten van geruchtenbestrijding men zich vooral heeft gericht en welke aspecten sterk verwaarloosd zijn.

In de terminologie van een epidemie is immunisatie een van de effectiefste methodes van geruchtenbestrijding; deze komt voornamelijk tot stand vanuit de wetenschap. In de vierde stap onderzoeken we in hoeverre de wetenschap in staat is om deze kwestie het hoofd te bieden. Als we voor een bepaalde kwestie dit 'paraatheidsniveau' niet kennen, overschatten of onderschatten we onze capaciteiten. In beide gevallen kan dat leiden tot besluiten die niet werken. Om dit probleem aan te pakken hebben we het ontstaan van een gerucht vergeleken met de opkomst van een technologie. Met dit kader hebben we het paraatheidsniveau van de wetenschap op het gebied van geruchtencirculatie gemeten. In dit kader bestuderen we vier dimensies van ontstaan (nieuwheid, groei, samenhang en impact) aan de hand van ruim 21.000 wetenschappelijke artikelen, om het paraatheidsniveau op elke dimensie te zien. Uit de resultaten blijkt een organische groei van paraetheid die onvoldoende is om opgewassen te zijn tegen de vloed van geruchten in de sociale media. Dit probleem kan worden aangepakt door exclusieve podia te creëren die leiden tot de vorming van een stabiele gemeenschap, en door de realisatie van een actief vakgebied voor de bestudering van geruchten.

Het andere aspect van het epidemiekader betreft minimalisering van de blootstelling en vermindering van de overdrachtssnelheid. Deze aspecten worden in de vijfde stap behandeld door middel van een methode uit de kunstmatige intelligentie. De drastische toename in volume, snelheid en de verscheidenheid aan geruchten betekent dat er automatische oplossingen nodig zijn om content te inspecteren die op de sociale media circuleert. Hiervoor is binaire classificatie de meest gebruikte rekenmethode; een nadeel hiervan is echter de non-rumour-valkuil, waardoor de classificatiefunctie onbetrouwbaar en inconsequent wordt. Daarom gebruiken we een nieuwe classificatiemethode, die voor de trainingsfase slechts één in plaats van meerdere klassen gebruikt. Uit de experimenten met deze methode op twee grote datasets komt een veelbelovende classificatiefunctie naar voren, die geruchten met een hoge F1-score kan herkennen.

In de laatste stap van het onderzoek bekijken we geruchtenbestrijding vanuit een proactief perspectief. Het epidemiekader helpt oplossingen te ontwikkelen om de verspreiding van geruchten onder controle te houden; maar meestal betreft dit passieve methoden, reactief en achteraf. In deze stap introduceren we een ontologisch model waarin de onderliggende mechanismen van sociale manipulatie kunnen worden beschreven. Dit model neemt een proactieve houding tegen sociale manipulatie aan en biedt ons de mogelijkheid om preventieve maatregelen te ontwikkelen. Het model wordt geëvalueerd door deskundigen en naast drie beruchte voorbeelden van sociale manipulatiecampagnes gelegd.

1

INTRODUCTION

Our lives begin to end the day we become silent about things that matter.

Martin Luther King Jr

1.1. AN OVERVIEW ON THE PHENOMENON OF RUMOUR SPREADING

In one of the most famous Shakespeare's plays - Henry the Fourth, Part II - he writes "rumour is a pipe, blown by surmises, jealousies, conjectures, and of so easy and so plain a stop, that the blunt monster with uncounted heads, the still-discordant wavering multitude, can play upon it". Shakespeare's words elegantly express how easy, widespread, and vicious the emergence and circulation of rumours could be. Since the play was first written, the phenomenon of rumour spreading is exacerbated and turned into a far-reaching phenomenon to the extent that the World Economic Forum ranked the spread of misinformation as one of the top risks facing the world today [1], and Oxford dictionary picked fake-news as the term of the year in 2016 [2].

Although the rumour spreading is mostly associated with political contexts owing to the excessive use of rumours by political figures to disparage their rivals and critics, the scope of this phenomenon is much bigger than politics [3, 4, 5, 6]. It is, in fact, a domain-agnostic phenomenon that arises in any circumstance in which meanings are uncertain, questions are unsettled, information is missing, and lines of communications are absent [6]. From the content perspective, rumours are unverified statements about instrumentally important topics. Thus any incident -no matter if it is political or not- could be a subject of rumour-mongering. People engage in the rumour process since it attributes a ready-made justification to unexplained events. It increases the comprehension and understanding of the situation by offering details and reasons as well as meanings and clarifications. Rumours might also be initiated deliberately as a psychological tool for strategic purposes such as character assassination, influence operations, and financial benefits [7, 8, 3]. The dissemination of rumours, whether being intentional or inadvertent, may feed on hate, create fear, and raise false hopes [9]. It may tarnish reputation of individuals [4], organisation [3], or even countries [10], provoke riot and unrest [5], shake financial markets [11], influence decision-making process [12], and disrupt aid operations [13, 7].

The rumour is a collective process whose existence is contingent on the circulation [14]. In this vein, the role of media is crucial as it streamlines the communication and increases the rate of reach to the audience [15] (and subsequent exposure to the rumour). Traditionally word-of-mouth and letter were the primary means of communication and rumour spreading [16]. The advent of the technologies such as printing press and radio for the mass communication profoundly affected the rumour spreading [15]. Particularly, the sudden rise of social media in the last decade of the twentieth century has provided a nurturing environment for rumours [17] to thrive and circulate in an unprecedented scale, speed, and scope [18, 19]. The size and diversity of social networks [20] as well as automation mechanisms [21, 22, 23, 24, 25] play a central role in the degree of rumour dissemination. Besides, other factors, such as a lack of media literacy [26], minimal supervision [17], low barrier to entry [27], and the lack of social media regulation [28] facilitate the creation and circulation of rumours.

The escalation in the rumour diffusion may lead to severe consequences that can influence political, economic, and social well-being [13]. For instance, on April 23 of 2013, the Associated Press Twitter account released a tweet saying "Breaking: Two explosions

in the White House and Barack Obama has been injured.” This tweet went viral by 4000 tweets in less than 5 minutes. The spread of this false news precipitated a big drop (with the value of 140 billion dollars) in the market in a single day. In fact, automated trading algorithms immediately began trading based on the potentials and consequences of the explosion in the white house and the death or injury of U.S. president [11]. This example just shows one case of rumour spreading with dire consequences. There are plenty of rumour dissemination cases in other domains such as elections [29, 30, 31, 32, 33, 34, 35], business issues [12, 6], and healthcare [36, 21] which lead into severe outcomes.

In response to the detrimental effects of rumour propagation, a series of confrontation strategies has been devised. Although taking the potential danger of rumour spreading into account and countering this phenomenon was an important action, it was often an intermittent effort with ephemeral impacts. There was no long-term plan behind the confrontation strategies. Whenever a major incident happened or was about to happen, rumours started to thrive and then countering techniques were proposed and practised [5, 37, 4, 7]. While this approach might have worked previously, it could not keep up with the rate of rumour supply and circulation due to the sudden growth of social media in the past decade. Therefore, the countering methods also began changing to the extent that variety of stakeholders such as social media platforms, governments, academia, and media organisations started to collaborate and developed new solutions. Although a constellation of counter-rumour strategies has been proposed and practised in different levels especially in the past few years, the massive waves of rumours are still sweeping over individuals, organisations, and societal institutions [13]. This is an alarming trend that has to be controlled; otherwise, due to the potential of social media rumours the repercussions might be catastrophic.

1.2. RESEARCH OBJECTIVE AND RESEARCH QUESTIONS

The primary motivation of this dissertation is to tackle the wild spread of rumours in online social media. A clear problem definition is the first and foremost prerequisite to this goal. Albert Einstein once said, “If I were given one hour to save the planet, I would spend 59 minutes defining the problem and one minute resolving it”. It is crucial to obtain a good understanding of the problem before taking any action; otherwise, the proposed solution would be inaccurate and irrelevant. The problem definition, in this case, entails determining what exactly has to be curbed and controlled. There are different variations of false and unverified information (e.g., fake-news, disinformation, misinformation, conspiracy theory, etc.) which are recognised by the scholars as similarly harmful phenomena. However, it has to be clarified what is and what is not in the focal point of this thesis. Besides, a full understanding happens when the targeted phenomenon is studied within the context (i.e., social media). It helps to understand whether and to what extent social media features facilitate the spread of rumours.

By defining and demarcation of the problem, it would be clear what has to be tackled. Because of the relatively long period of rumour confrontation in the societies, it is indispensable to obtain an overview of the past counter-rumour strategies. It would provide information about the strengths and weaknesses of the rumour responses in the past. Those information could be utilised later in the development of confrontation plan against rumour dissemination. After going through the past strategies and investigating

them, it is time to act and tackle rumours. There are two broad paradigms of passive and pro-active confrontations. Despite the clear advantages of pro-active approach, the current landscape of rumour confrontation is extensively dominated by passive approach. For an effective and feasible confrontation plan, both paradigms should be therefore presented in the rumour response agenda.

The passive paradigm consists of two major strands of short- and long-term strategies. Short-term strategies often aim to filter rumours using a machine learning technique called binary classification. It is a supervised learning technique in which a model is trained with existing samples of rumours and non-rumours in order to flag unforeseen rumour messages. The other set of strategies are the ones with the goal of creating long-term immunity. Those strategies tend to create a resilient society by training people to be more careful and critical about the information they receive. In this vein, the role of academia to assess the effectiveness of training methods or to develop new methods is crucial. Despite a great deal of research in this arena, the amount of progress by academia is not clear yet. This may lead to misjudgements about the performance of the research topic, which can ultimately result in wrong science policies regarding academic efforts for quelling rumours. The other confrontation paradigm is pro-active, which aims to take measure before a rumour begins to spread. This approach has not practised yet; thus, it is essential to take the preliminary steps and develop an early rumour confrontation model with a pro-active perspective.

To address the above-mentioned gaps, the principal objective of this dissertation is defined as follows:

To systematically study the rumour confrontation within online social media.

To accomplish this objective, it is required to look into four nearly-independent topics, which are described in the following:

- First, the main construct of this study, namely rumour needs to be scrutinised. It is like an underlying substrate that glues down different pieces of this manuscript together. The notion of rumour refers to a complex phenomenon with a controversial conceptualisation which makes its identification rather difficult among the closely related concepts. The lack of crystal clear understanding of rumour, would be like going to a war without knowing who the enemy is.
- Second, social media is a major medium for the emergence and the spread of rumours. In the post social media era, the spread of rumours scaled up, accelerated, and diversified. It is essential to understand the properties and mechanisms of this environment that facilitate the emergence and growth of rumours.
- Third, having an overarching view regarding the as-is situation of rumour confrontation, is essential to tackle rumour spreading in social media. This entails a comprehensive and critically analysed list of the past counter-rumour strategies juxtaposed in a common framework.
- Fourth, the addressing shortcomings of the past counter-rumour strategies is a prerequisite to the development of new confrontation strategies. Due to the long-standing vulnerability of human-being to rumours and the high rate of diffusion,

strategies based on short- and long-term approaches should be taken into account.

For each of the above topics, one or more research questions are raised, whose answers can help address the objective of the dissertation. The first question involves the main building-block of this research and take the rumour conceptualisation into account. The second one comprises the role of social media in the facilitation of rumour spreading. The third question tends to address the as-is situation of rumour confrontation by the analysis of the past counter-rumour strategies. The three remaining questions are about tackling rumours. In the following, the research questions are discussed in more detail.

RQ1. WHAT IS RUMOUR AND HOW IS IT DIFFERENTIATED FROM ITS CONCEPTUAL SIBLINGS?

This question is posed regarding the epistemic crisis of rumour and its conceptual siblings. There are plenty of concepts in the English language implying false or unverified information. However despite the years of academic research spent on those concepts, there is a considerable disagreement between the proposed definitions as they are often conflated into one idea or used interchangeably. The lack of consensus on the conceptualisation could become an obstacle to countering the surge of false information by creating confusion, distracting the community's attention, and draining their efforts.

RQ2. TO WHAT EXTENT SOCIAL MEDIA STREAMLINE THE SPREAD OF RUMOURS?

Rumour spreading is a long-standing phenomenon between human-beings. The development of communication technologies has facilitated the spread of rumours by introducing features such as synchronicity and distant mass communication. However, the emergence and radical growth of social media lead to a widespread hyper-connected network which provided a suitable environment for rumours to thrive and precipitate catastrophic consequences. This research question tends to investigate the anatomy of social media to understand the mechanisms and properties that could promote the spread of rumours.

RQ3. WHAT IS THE CURRENT STATUS OF RUMOUR RESPONSE STRATEGIES?

Despite tremendous efforts on the development of counter-rumour strategies, whenever a news-worthy incident occurs, social media flooded with rumours as if there is no mechanism to tackle this mischievous phenomenon. Here, the purpose of questioning the as-is situation is to shed light on the past efforts in countering rumours to ascertain the flaws and shortcomings of current control approaches. It works as a bird-eye view which allows understanding what aspects of rumour confrontation has been targeted extensively and what aspects are highly neglected.

RQ4. HOW READY IS THE ACADEMIA REGARDING THE SPREAD OF RUMOURS?

One of the important aspects of confrontation with rumour spreading that deserves special attention is to create immunity against the rumours. Academia is on the front-line

of developing immunity-based response; however, it is not known whether the past academic efforts could do justice to the significance of this confrontation approach. This can be problematic as it may lead to the overestimation or underestimation of academia regarding its competency in tackling rumours. What this research question is bringing up is to measure the readiness of academia regarding rumour spreading. This evidence-based approach prevents misjudgements and leads to decisions relevant to the reality and existing circumstance.

RQ5. HOW COULD WE IDENTIFY RUMOURS IN SOCIAL NETWORKS AUTOMATICALLY, CONSISTENTLY AND IN A TIMELY MANNER?

The massive flow of rumours in social media has made the manual inspection of the transmitted messages impossible. One of the alternative approaches that could be used is computational rumour detection which is scalable and fast. The dominant computational technique for the identification of rumours is the binary classification which tends to be inconsistent as it is highly dependent on annotators' volition in the annotation phase. This would call for a solution that benefits the scalability and speed of this approach and can address its inconsistency issue.

RQ6. HOW COULD WE TAKE PREEMPTIVE MEASURES REGARDING RUMOURS IN SOCIAL MEDIA?

Despite the discrepancies between the counter-rumour strategies, they share a similar confrontation style. They develop resilience against rumours in a retrospective manner. They tacitly assume the inflow of rumours always recycles the past rumourmongering techniques. Thus if the new rumours use novel techniques, it would be pretty hard to rein them. In order to address these issues, we could switch to the pro-active confrontation style, which simply means looking at the rumour process from rumourmonger eyes. This would allow to think like adversaries, discover their plans before execution, and develop preemptive measures.

1.3. CONTRIBUTIONS AND GUIDE TO READERS

In particular, this dissertation makes six different contributions to the field of rumour studies by a systematic study on the rumour confrontation within the social media. Chapter 2 contributes a comprehensive conceptualisation regarding the notion of rumour and its conceptual siblings. Chapter 3 investigates the role of social media in the rumour promotion by measuring the extent that recommendation systems streamline the spread of rumours. Chapter 4 evaluates the as-is situation of rumour confrontation by presenting past counter-rumour strategies, and then setting them in the epidemic control framework. Chapter 5 contributes to the rumour immunisation approach by measuring the readiness of the academia regarding rumour spreading through a bibliometric approach. Chapter 6 contributes to the mitigation of rumour transmission- and rumour exposure-rate by proposing a novel approach to computational rumour detection based on machine learning techniques. Finally, Chapter 7 proposes a pro-active approach to rumour confrontation by developing an operational level model that can capture the underlying mechanisms of rumour campaigns. The following outline puts forward the list

of contributions as well as their corresponding chapters and research questions.

- **Chapter 2** In response to RQ1, we examine the epistemic crisis between different variations of false and unverified information. We delve into the literature and infer that rumour, misinformation, disinformation, propaganda, conspiracy theory, pseudoscience, and fake-news belong to the same conceptual family as they follow a similar development process. The genesis of each concept might find its origins in different uses, but after the first generation of transmission, different variations start to look alike. This would help the scientific community to pool their knowledge and resources on confrontation with rumour spreading instead of endless discussions on the categorisation of false and unverified information. It is also discussed that gossip and legend do not belong to the rumour family no matter how similar their development process is. This would also brief the community to include different variation of rumours and leave out gossips and legends when they study rumours. Chapter 2 provides a more detailed explanation regarding this process-based view to rumour and its conceptual siblings.
- **Chapter 3** In response to the RQ2, namely the role of social media in the promotion of rumours, we investigate an exclusive social media automation mechanisms of recommendation systems as they are alleged to play a central role in the spread of rumours. To this end, we analysed 1,000 YouTube videos about conspiratorial topics. Our analysis along with a handful of studies in this domain show that the automation mechanisms in online social media platforms have a clear impact on the spread of rumours; however, this effect mediates by a variety of factors such as location, time, and rumour topic. In Chapter 3 data collection, experiments and results are thoroughly discussed.
- **Chapter 4** In response to RQ3, we collect, review, and analyse major counter-rumour strategies that were dispersed in the literature. Our focus is on the organisational and governmental response to tackle the rumours since the world war II. To understand why those strategies could not steadily rein in rumour spreading, we analyse them using epidemic control framework due to the strong similarity between the propagation of disease and information. We conclude that the ephemeral reactions, the absence of a comprehensive plan, and neglecting the immunisation-based solutions are amongst the reasons for the failure of response to rumour dissemination. Chapter 4 provides detailed explanations regarding the counter strategies and their analysis.
- **Chapter 5** In response to RQ4, we use the theory of emergence to assess the readiness of academia regarding rumour spreading. Based on this theory, five dimensions of novelty, growth, coherence, impact, uncertainty and ambiguity determine the status of an emerging phenomenon. In this research, we first need to quantify the academic efforts regarding rumour spreading to be able to measure it. To this end, we collect more than 21,000 scientific papers about rumours. The next step is the operationalisation of the emergence dimensions. After this phase, and measuring the degree of emergence in the topic of rumour spreading, we could observe an increasing trend for the growth, the coherence and the

impact and a decreasing trend for the novelty. To propel this research domain and encourage academia to contribute more to this arena, we propose an external push strategy meaning arranging dedicated publication venues such as journals and conferences for this field of research. In Chapter 5, data collection, analysis, and results are explained and discussed in detail.

- **Chapter 6** In response to RQ5, we first pose a major issue regarding the binary classification as the predominant approach in computational rumour detection. We argue that unlike rumour samples which are often annotated similarly, non-rumours get their labels arbitrarily based on annotators' volition. Because of that, binary classification may lead to unreliable outcomes. To tackle this issue, we propose to use a novel classification approach called one-class classification (OCC). Unlike the binary classification, the training in OCC is only based on one class. We apply seven once-class classifiers from three different learning paradigms and compare their performance. Our results show that this approach can recognise rumours with a high level of F1-score. Chapter 6 provides detailed explanations regarding data, features, and experiments.
- **Chapter 7** In response to RQ6, we propose a proactive rumour confrontation approach which provides us with an opportunity of looking at the rumour campaigns from an adversarial perspective and developing preemptive measures. We develop this model in a step by step manner. We start from a coarse-grained model (by combining the misinformation machine model and Maathuis Cyber Operation Model), then we operationalise it based on the literature, real cases, and expert interviews in an iterative manner. Finally, we give a formal presentation of the model using OWL. In Chapter 7 the model development and verification is explained in detail.

1.4. ENGINEERING SOCIAL TECHNOLOGIES FOR A RESPONSIBLE DIGITAL FUTURE

This section explains the relevance of this thesis with the TU Delft's research program of "Engineering Social Technologies for a Responsible Digital Future". The technological developments are accelerating across a large number of domains, from health to finance and communication [38]. This rapid development is like a double-edged sword which comes with perils and promises. Although on the surface, technologies often offer a lot to fix problems and improve humans life, but underneath they may lead to more troubles. Thus, there is an urgent need for the investigate those technologies (and the changes inflicted by them), in order to take appropriate measures before it gets too late.

One of those technologies with far-reaching implications on our lives is social media. It has removed the physical barriers and allows multilateral synchronous communication with long-distance locations. It also provides us with the opportunity of multi-media message transmission. However, all those features could also serve the mischievous function of rumour spreading which may lead to catastrophic repercussions. Hence it is of the utmost importance to protect and secure this technology by countering irresponsible usages. Due to the multidisciplinary nature and the large scale

of the problem, it should be addressed by social technologies which incorporate both social and computational aspects of the problem. This problem is addressed based on the principles mentioned above. It harvests a socio-technical approach to benefit social media without any concern regarding rumours.

2

CONCEPTUALISATION OF FALSE AND UNVERIFIED INFORMATION

Money may hire a rumor agent but it cannot forge a rumor chain

The Psychology of Rumor, Gordon Allport & Leo Postman

There are myriad of concepts in the English language, implying false or unverified information. Despite years of academic research, there is no consensus regarding their conceptualisation and they are often used interchangeably or conflated into one idea. This problem could become an obstacle to countering the surge of false information by creating confusion, distracting the community's attention, and draining their effort. To tackle this issue, the following research question is posed in this chapter:

- *What is the rumour and how is it differentiated from its conceptual siblings?*

To address this question, we identify and explain the various forms of false and unverified information, their relevance, and impact. In the next step, we argue that if we take the process-based view into account, most of those variations behave like rumour spreading. Based on this approach, although the genesis of such information might be deliberate or inadvertent and with different purposes, they primarily disseminate on the basis of similar motives and follow the same process¹.

¹This chapter is based on the following under review manuscript: Fard, A. E., & Verma, T. A Comprehensive Review on Countering Rumours in the Age of Online Social Media Platforms. In Causes and Symptoms of Socio-Cultural Polarization: Role of Information and Communication Technologies, Springer (Under Review).

2.1. INTRODUCTION

There are many concepts in the English language implying false or unverified information. Terms such as misinformation, disinformation, rumour, urban legend, fake-news, propaganda, and conspiracy theory are just a few of these concepts that intermittently appear in the scientific arena. What academia has experienced regarding the conceptualisation of those terms and their conceptual siblings is an epistemic crisis. Although there are plenty of studies exploring various kinds of false and unverified information from different angles, there is considerable disagreement between proposed definitions. They are often conflated or have been used interchangeably [39, 40, 41, 42, 43, 44, 17, 13, 45, 46, 47]. The lack of consensus on the conceptualisation would create confusion and drains the community's efforts in countering the surge of false information.

Despite the discrepancies in the definitions, if the dynamics of false and unverified information is taken into account, then much similarities would appear between many of their seemingly different variations. By the dynamics, it means taking the process (i.e. creation and dissemination) of false & unverified information into account. Although the genesis of such information might be deliberate or inadvertent and with different purposes, they primarily disseminate on the basis of similar motives and follow the same process [6, 8, 48].

Tackling this issue would help to understand what we are and what we are not going to confront. In other words, it would demarcate the boundary of this research. Additionally, addressing this issue would lead to a more valid and accurate plan to overcome the threat of misleading information.

2.2. THE VARIATIONS OF FALSE AND UNVERIFIED INFORMATION

This section investigates rumour, gossip, legend, propaganda, conspiracy theory, fake-news, pseudoscience, and misinformation as major variations of false and unverified information.

2.2.1. RUMOUR

The notion of rumour refers to unverified and instrumentally relevant information statements in circulation that arise in a situation of ambiguity, danger, threat, or change; and are passed along by the people attempting to make sense or to manage risk [6, 49, 8, 7]. In the following, the elements of this phenomenon is discussed in details.

First, rumours are declarative statements composed of nouns and verb statements that purport to inform, explain, predict and thus provide information [6, 50]. For example, the viral (false) rumours of "McDonald puts red worms in their hamburgers" [3], "Procter & Gamble has a connection with the church of Satan" [3], "The African AIDS pandemic occurred because the AIDS virus was created in a Western laboratory and tested on Africans" [8] are all declarative statements aiming to transfer (misleading) information to their readers. Second, the rumour is a collective process that arises in the collaboration of many. It involves a division of labour among participants, each of whom makes a different contribution [14]. Rumour existence is contingent on its circulation [50], and end of the communication activity equals to the death of rumour [14] therefore private thoughts, prejudices, beliefs, attitudes, or stereotypes held by an individual are

not deemed as a rumour, although each of which may be conveyed in a rumour [6]. Rumour is not considered as the transmission of a designated message, but as something that is shaped, reshaped, and reinforced in a sequence of message transmission [14]. Rumour can also be viewed as a meme that may adapt, survive, or die just like species in the nature that follow the same process [8].

Third, a rumour spreads if it relates to, affects, or threatens rumour participants in some way. The term “instrumental” emphasises on the purposeful function of rumour rather than being solely sociable, entertaining, and aimless. Although rumours could be a vehicle for the entertainment and sociability, they are not primarily meant to pass the time. Rumours tend to be about topics that people perceive relatively urgent, significant, or important [7]. In their seminal book “Psychology of Rumor”, Allport and Postman write, “... an American citizen is not likely to spread rumour concerning the market price for camels in Afghanistan because the subject has no importance for him” [4].

Fourth, rumours are unverified in some context, and they are not accompanied by substantial evidence for at least some group of people [8]. Being unverified does not equal to being false. In fact, an unverified piece of information can be true or false. Truthfulness refers to the correspondence with objective reality, while verification means correspondence with objective reality based on an external resource. Fifth, rumours tend to thrive amid situations that are ambiguous, confusing, uncertain, and threatening. Situations with uncertain meanings, unsettled questions, missing information, absent communication lines, and physical & mental impacts [6]. Rumours are predominantly associated with wars, natural/human-made disasters, elections, economic and political crises, and minority group antagonism since such contexts have a high level of ambiguity or pose a threat [3].

Sixth, rumours are circulating primarily as a sense-making or threat management mechanism. In order to understand rumour as a sense-making mechanism, we first need to understand how individuals make sense of things. Sense-making is similar to the task of explanation, which aims at increasing comprehension and understanding. It offers details and reasons as well as meanings, clarifications, and justifications. One of the forms of sense-making is threat management. As it is discussed earlier in this section, threat or potential threat is one of the contexts that rumours tend to emerge. In such situations, rumours operate as a coping strategy by neutralising the threats or encouraging people to deal with them through positive actions or simply feeling better about them. For example, denigrating the source of the threat or bolstering our position, cause, or group are typical stable causes that are posed by the rumours at the time of threats [6]. Although rumours primarily function as a sense-making and threat management mechanism, they serve other functions such as titillation and breaking the monotony [6, 3], alliance making, and enforcement of communal norms. None of these functions is mutually exclusive, in other words, in a rumour, there might be some people who find it entertaining or some others who use it to build alliance; however, the essence of rumours is sense-making and threat management while other functionalities are secondary [6, 3].

2.2.2. GOSSIP

Gossip is an evaluative social talk about an individual's personal life. It is a group level evolutionary phenomenon [51] which glues down groups and adjust people's relation-

ships. Gossiping can maintain group cohesiveness, and establish, change, and maintain group norms, group power structure, and group membership. It can also function as an entertainment mechanism [6, 7]. Gossiping is also an effective strategy when it comes to the intragroup competition [6, 52]. Gossips may get slanderous and be driven by nefarious self-serving motives. They may break groups apart or taint people's reputation. However, there are benevolent gossips that function as a warning against the harmful behaviour of particular individuals. Gossips may also regulate individuals' behaviour regarding the context. Gossip is a mechanism between friends, not those who do not know each other. It is the signal of affiliation, closeness, and camaraderie [6, 52].

2.2.3. LEGEND

Legends² are narratives with moral lessons about unusual, humorous, or horrible events [53, 7]. They are recounted for many years, and after a prior history of distortion and transformation, they converge to stable forms and become part of the folklore and verbal heritage of people³. Legends are immortal because they capture the universal aspects of human character. Legends are told in a storytelling framework. They have a setting, plot, characters, climax, and denouement. They function as a mechanism for entertainment and propagation of values and mores. Legends also make sense of the world by providing answers to the persistent riddles of life. Legends are about subjects which considered important for successive generations. If the legends are about primal forces, cosmology, or religious beliefs, then they are called myth [4, 7].

2.2.4. PROPAGANDA

Propaganda refers to persuasive tactics devised deliberately by governments or corporations to promote or challenge a particular viewpoint by manipulating symbolic representations [14, 42, 54]. Propaganda might be used in a variety of subject domains; however, two of them are more prevalent: politics and business. The former aims to spread pro-government or pro-party narratives; to attack the opposition or mount the smear campaigns; to distract or divert conversations or criticism away from important issues; to drive division and polarisation, and to suppress participation through personal attacks or harassment [55]. In the latter, the goal is to influence the beliefs and undermine reliable evidence by the corporates. During the second half of the twentieth century, tobacco companies organised campaigns to undermine scientific evidence demonstrating the link between lung cancer and smoking. They could successfully delay regulation to reduce smoking [39, 42].

Propaganda over online social media is called computational propaganda. It is described as "the use of algorithms, automation, and human curation to purposefully manage and distribute misleading information over social media networks" [25]. The computational setting in computational propaganda allows automation which brings scalability and anonymity. Many of the state and non-state actors use computational propa-

²The term "legend" refers to both traditional legends (about knighthood, ogres, witches, sleeping princesses, etc.) and modern or contemporary legends (about dating, technology, organ removal, etc.) (Modern/Contemporary legends are also called urban legends which is a misnomer because those narratives do not necessarily take place in urban environment).

³For urban legends, they take the context whereby they are recounted.

ganda to suppress their oppositions, to promote their viewpoints, to divert or destroy movements, and to create fake trends [25, 55, 56].

Propaganda may take three broad shapes on the basis of accuracy and source recognition. The white propaganda refers to relatively mild propaganda with an accurate message, identified source, and acknowledged sponsorship. In contrast, the black propaganda is credited to false sources and aims to harm its audience via lies, fabrications and deceptions. The third shape is situated between black and white propaganda, where the message accuracy is uncertain, and the source may or may not be identified [57, 58, 59].

Propaganda may take particular shapes. One of them is innuendo which functions as a character assassination technique to discredit the reputed individuals. For instance, since early times innuendos tarnished U.S. presidential elections by accusing of candidates with illicit sexual relations, racism, brutal treatment of wives, drunkenness and the alleged possession of certain blood types [4]. It may also serve as a projection technique to accuse another person of the same things that the accuser is guilty of [8]. One of the most notorious shapes of propaganda rumour is disinformation which was invented by KGB in 1923. It is black propaganda based on forgeries. Disinformation includes forged and fabricated narratives, letters, documents, photographs, reports, and press releases [60, 61, 62]. One of the kinds of forgery that is getting increasingly popular is audiovisual (AV) manipulation. It includes both the cutting edge AI-reliant technologies of deepfakes as well as cheap-fakes that are conventional techniques of audiovisual manipulation such as speeding, slowing, cutting, re-staging, or re-contextualising footage [63].

2.2.5. CONSPIRACY THEORY

Conspiracy theories are unverified information in circulation about events or incidents that are caused on deliberate hostile purposes by a coalition of actors operating in secret. A conspiracy theory assumes pre-designed patterns govern the universe, and there is no room for randomness and coincidence. That is why conspiracy theories try to randomly connect the dots and find the secret patterns [64, 65, 66]. Conspiracy theories may arise in a variety of subject domains such as scientific research (e.g. global warming is a hoax created by China [67]), sport (e.g. referee bribing conspiracy theory [64]), or the government (e.g. deep state conspiracy theory [68]). Among the commonly used conspiracy tactics are contradictory explanations, overriding suspicion, nefarious intent, something must be wrong, persecuted victim, immunity to evidence, and re-interpreting randomness [69].

2.2.6. FAKE-NEWS

The notion of fake-news⁴ is defined as “the fabricated information that mimics news media contents in form but not in the process and intent”. Fake-news outlets do not follow editorial norms and guidelines [17]. In such outlets, there is neither fact-checking

⁴In the current political climate, there is a major disagreement in academia regarding the consumption of the term “fake news” as it became a value-loaded term linked to particular political figures [13, 17]; however, due to the lack of an alternative term and to avoid adding further confusion to the existing fluid terminology, we have elected to retain the term “fake-news”.

nor source verification; articles are emotionally charged and written in narrative style; sometimes, articles have inconsistencies with the registration date [44]. Although fake-news articles have mostly arisen in a political context, there are plenty of cases in other domains such as vaccination, and stock values [17].

Since the early days of journalism, fake-news found its way into the news outlets. Fake-news articles could draw attention easier than real news as there is no constraint for fabrication, and we can be as creative as we want to develop appealing, attention-grabbing and memorable fake-news articles [70]. More attention means higher readership, which can lead to a more significant profit margin for the news outlets [71]. One of the earliest and most successful fake-news articles was the New York Sun's "Great Moon Hoax" of 1835 claimed there was an alien civilisation on the moon. This fabricated story drew much attention to New York Sun to the extent that its circulation reached from 8000 to 19000 copies, which meant overtaking Times of London as the world's bestselling daily newspaper [71, 72, 73].

2.2.7. PSEUDOSCIENCE

A statement is considered pseudoscientific if it satisfies three criteria of (i) scientific domain, (ii) unreliability, and (iii) deviant doctrine. The criterion of scientific domain entails a pseudoscientific statement to be about an issue within the domain of science. The term "science" implies science in a broad sense which comprises humanity as well. Based on the criterion of unreliability, a pseudoscientific statement suffers from a severe lack of reliability and trust. Besides, it can neither be used for knowledge production nor practical cases. The deviant doctrine criteria indicate the support of pseudoscientific statement proponents to represent that statement as the most reliable knowledge on the subject matter. In order to consider a statement pseudoscientific, all the three conditions require to be confirmed. For example, if a statement satisfies the first two criteria but not the third one, probably it is fraud in science or mistake in science, but not pseudoscience [74].

Pseudoscience can take two different forms of science denialism and pseudo-theory promotion [74]. Science denialism refers to "the rejection of empirically supported propositions despite scientific consensus and the effort to create the appearance of debate when there is none" [75]. Some typical examples are climate change denialism, Holocaust denialism, relativity theory denialism, AIDS denialism, vaccination denialism, and tobacco disease denialism [74]. Science denialists pursue certain types of techniques to present their arguments and persuade others. The FLICC framework collected those techniques and categorised them under five groups of fake-experts, logical fallacies, impossible expectations, cherry-picking, and conspiracy theory [69, 76].

The other category of pseudoscience is pseudo-theory promotion which is referred to the fabrication of a set of claims in order to advance the pseudoscientist's theory. Sometimes it leads to the rejection of parts of science. Some typical examples of pseudo-theories are astrology, homeopathy, iridology, Scientology, transcendental meditation, and ancient astronaut theories [74, 77]. Science denialism and pseudo-theory promotion are not mutually exclusive, and there are cases with shades of both categories. For instance, although Scientology is an exemplar case of pseudo-theory promotion, Scientologists attack science-based psychological treatments and disparage it, in order to

justify and promote their solutions [74].

2.2.8. MISINFORMATION

The other term that is often used in this domain is misinformation. This concept originates in cognitive psychology and developed by the scholars who were studying misinformation effects on memory formation, visual object classification, children's ability to infer the mental states of others, and performance on multiple-choice tests [39]. The misinformation effect refers to "the distorting effects of misleading post-event information on memory for words, faces, and details of witnessed events" [78]. Nevertheless, nowadays, the term has found a much broader yet loose meaning: any kind of deceptive message that might be harmful but spreads inadvertently [39]. It overlaps with many of the concepts that we discussed so far. Regardless of the intent behind the spread of information, if its truthfulness is unverified for an individual and s/he spreads it (without malicious intention), then that piece is considered as misinformation. This means those uninformed individuals who discuss different forms of rumours without strategic purposes are participating in misinformation circulation.

2.2.9. THE COMPARISON OF FALSE AND UNVERIFIED INFORMATION

This section compares different forms of false and unverified information using the information dimension scale (IDS) [7]. IDS is a framework to differentiate information structures. The first dimension specifies to what extent a piece of information is supported by an evidence. The second dimension shows to what extent the statement is important, is significant, and will be talked about seriously. The third dimension measures the extent to which the information discredits someone and is derogatory. The fourth and fifth dimension try to capture the original theme, structure, and function of the statement when it initiates. The last dimension rates the extent to which the statements is entertaining, amusing, and enjoyable. Table 2.1 compares the discussed concepts based on the IDS framework.

2.3. PROCESS-BASED PERSPECTIVE

This section investigates the dynamic of false and unverified information by explaining the emergence and dissemination phases. False and unverified information is initially shared to serve four broad purposes of (i) social manipulation, (ii) sense-making and threat management, (iii) social dynamics, and (iv) cultural dynamics. Social manipulation refers to "the purposeful, systematic generation and dissemination of information to produce harmful social, political, and economic outcomes in a target area by affecting beliefs, attitudes, and behaviour" [79]. Planting misleading information into public is a long-standing manipulation strategy when much is at stake (e.g. in wartime, elections, highly competitive markets). False and unverified information can also appear at the time of uncertainty and threat as a coping strategy and "to give meaning to our sensations and to put a context around them, so they gain significance and fit into an understanding that coheres" [6]. They may also function as a social mechanism to entertain, to supply social information, and to establish, change, or maintain group membership, group power structure, or group norms [7]. The cultural dynamics is the other purpose

Table 2.1: Comparison between different forms of false and unverified information ([7]).

	Evidentiary basis	Perceived importance	Content slanderous	Message theme & structure	Function	Entertaining	Citation
Rumour	Low	High	Maybe	The message is a declarative statement, consisting mostly nouns and verbs.	To make sense of ambiguity and to manage threat or potential threat	Maybe	[6, 7]
Gossip	Maybe	Low	High	The message is evaluative, informal and refers to individuals.	Allowing groups to become more cohesive and to define their membership, norms, and power structure	High	[6, 7]
Legend	Low	Low	Low	The narratives pertains to issues that are important for successive generations such as birth, marriage, and death. They have story like structures, including setting plot, characters, climax, and denouement.	To entertain and to convey mores, norms, and cultural truths	High	[4, 6, 7]
Propaganda	Low	High	High	The messages are about supporting or challenging particular viewpoints or ideologies mostly in politics and business. The message structure is composed of fabricated materials as well as manipulated and vivid images, symbols, and slogans.	To simultaneously induce psychological threats and to function as a sense-making and threat management mechanism.	Low	[4, 6, 7, 25, 55, 57, 58, 60, 61]
Conspiracy Theory	Low	High	High	The messages follow a narrative about covert and hostile activities of secret and powerful groups.	To cope with threats by providing alternative explanations for events and incidents.	Low	[48]
Fake-news	Low	High	High	The messages are emotionally charged and written in narrative style; They are not fact-checked and their source is not verified. In order to draw attentions, elements of (i) threat-related information, (ii) sexually related information, or (iii) elements associated to disgust, are incorporated.	To gain financial benefits by drawing eyeballs.	Maybe	[17, 44, 71, 70]
Pseudoscience	Low	High	Maybe	The message pertains to an issue within the domain of science in a broad sense.	Sense-making and threat management by rejecting the empirical studies and/or promoting fabricated claims	Low	[74, 77]
Misinformation	Low	High	Maybe		For sense-making	Maybe	[39]

of spreading false and unverified information to establish, maintain, or impart cultural mores, or values, and also provides answers to the persistent riddles of life [7, 4].

After creation of the message with either of those purposes (deliberately or inadvertently), it has to be conveyed to public. The generation and circulation of false or unverified information could be done through conventional (e.g., word of mouth) or modern (e.g., newspaper, television, or social media platform) means of communication. Sometimes they could also combine and create a hybrid environment for the information circulation.

After the false and unverified information releases, it would be extremely challenging to control its passage and keep it in check no matter for what purpose it has been created in the beginning. It forms spontaneously and its development depends upon fortuitous events, momentary emotional reactions, and the particular interest of those who make up the public [14]. When false and unverified information goes public, then it is driven by spontaneous discussions including different kinds of communication posture among the people who come across those information [7]. People often participate in those discussions in a collaborative manner by raising questions, providing information, indicating beliefs and opinions, expressing feelings, or suggesting a course of action no matter what is the type of false and unverified information [80]. Thus, whether it is a black propaganda to tarnish a presidential candidate, or an honest mistake about confusing the sound of a fire-cracker with an explosion, the same dynamic will happen. This dynamic process is exactly similar to what happens in rumour spreading when people engage in the shared sense-making process through interaction with others.

In fact, although the genesis of different variations of false and unverified information might be for different reasons, they pursue a similar dissemination dynamic which causes they look like rumour spreading after the first generation of the transmission. Propaganda and fake-news are planted into public deliberately to induce psychological threats and take advantage of the people; however, the audience treats them in the same way that they treat rumours. In fact, people circulate their impressions, interpretations, or reactions among themselves in order to make sense of those information [81, 8, 6]. Similarly, conspiracy theories and pseudoscience emerge as a coping strategy among a group of people to manage psychological threats in response to uncertain or threatening situations [48]. Nevertheless over the time they pursue the same path as rumours. Legends, myths, and urban legends are also very similar phenomenon; however, their life-cycle is much longer than rumours; hence we consider them as a separate phenomenon. Similarly, gossip is also a distinct phenomenon as it happens in the group level with a slightly different dynamic [7].

2.4. WHAT HAS TO BE CURBED?

This section explains why not every form of message in the constellation of false and unverified information needs to be curbed and controlled. As it is discussed before, rumours, legends, and gossips are three broad variations of false and unverified information. Amongst them, the least harmful one is a legend and its siblings, namely urban legend and myth. Although they had been rumour once, after years of transmission, it is a distinct phenomenon with key differences. The primary goal of legends is to share values and to provide answers to the riddles of life. Thus, it is highly unlikely that legends lead

to harmful consequences. The other one is gossip which mainly maintains group-level mechanisms such as cohesiveness, power structure, norms, and membership. However, it may take the slanderous shape and function based on nefarious self-serving motives. Nevertheless, it is highly unlikely that the impact of gossiping goes beyond the group boundary and reach to higher levels.

The other construct is rumour and its offspring. In a broad sense, rumour functions as a sense-making or threat management mechanism; however, depending on the form of rumour, both sense-making and threat management may take different shapes. In propaganda rumour, it mostly serves a pernicious function. Although there are different types of propaganda rumour, it is mostly used for malign purposes. The conspiracy rumours are also relatively harmful and harass their subjected groups by falsely accusing them. Fake-news rumours might become harmful by promoting appealing yet fabricated materials to lure individuals. The pseudoscientific rumour is a toxic phenomenon that attacks the institution of the science by tarnishing scientists, scientific evidence, and scientific methods. Misinformation does not inflict any harm wittingly; however, as it is discussed before, it may appear when uninformed individuals are engaged in rumours process. Besides, even if misinformation rumour does not take the shape of derogatory rumours and spread with a benign yet inadvertent motive, it may lead to harm.

Thus among the variations of false and unverified information, the rumour family operates in large-scale, and even if they start spreading unwittingly and without malign intent, they may still lead to severe consequences. Therefore it is of the utmost importance to take the variations of rumour into account and develop a solution to curb and control this phenomenon otherwise the repercussions would be inevitable and may influence political, economic, and social well-being.

2.5. CONCLUSION

This chapter provides a conceptualisation regarding a variety of false and unverified information as well as arguing why not every form needs to be taken down. It starts with raising the issue of epistemic crisis regarding different forms of false and unverified information. To address this issue, rumour, gossip, legend, propaganda, conspiracy theory, fake-news, misinformation, and pseudoscience are discussed and then analysed from the process based perspective. We infer that except gossip and legend which are fundamentally different regarding the context of emergence, content, and functionality, the rest of false and unverified information variations are in fact different forms of rumour. Finally, the last part of this chapter argues legends are highly unlikely to be harmful; gossips might be harmful but in small-scales; rumours might become extremely harmful in large-scale, therefore it is of utmost importance to dedicate all the resources and attention to confront them.

3

THE LANDSCAPE OF RUMOUR SPREADING

It is no longer enough to automate information flows about us; the goal now is to automate us.

Shoshana Zuboff, *The Age of Surveillance Capitalism*

After settling on the conceptualisation of false and unverified information, the next step is to understand features and mechanisms in one of the major rumour dissemination environments, namely online social media. To this end, this chapter brings up the following research question:

- *To what extent social media streamline the spread of rumours?*

To address this question, we first investigate major communication technologies and particularly online social media in the rumour spreading. After that, we focus on the recommender system as one of the exclusive AI-enabled mechanisms of online social media. Our goal is to understand whether and to what extent the recommender system tends to promote rumours. For this analysis, we chose YouTube. In the rest of this chapter, we explain our methodology and data collection procedure. After that, we present our results, which are consistent with the radicalisation hypothesis. Finally, we discuss our findings, as well as directions for future research and recommendations for users, industry, and policy-makers¹.

¹This chapter is based on the following publications:

- Alfano, M., Fard, A. E., Carter, J. A., Clutton, P., & Klein, C. Technologically scaffolded atypical cognition: The case of YouTube's recommender system. *Synthese*.
- Fard, A. E., & Verma, T. A Comprehensive Review on Countering Rumours in the Age of Online Social Media Platforms. In *Causes and Symptoms of Socio-Cultural Polarization: Role of Information and Communication Technologies*, Springer (Under Review).

3.1. INTRODUCTION

In the broadest term, rumour spreading is a form of communication. People use it to share their comments and feelings in specific contexts. The rumours used to circulate through word of mouth or letters; however, the technological development provided us with novel forms of communication. Although those technologies were developed to streamline the message transmission among people by removing the barriers of physical distance and opening up the possibilities of synchronous multi-lateral communication, they also facilitated the dissemination of rumours [18, 19].

This chapter aims to study rumours from the media perspective in order to understand the role of media in the prevalence of rumour circulation. In this vein, the transition between the pre and post social media era is of utmost importance because although the phenomenon of rumour spreading has been an invariable constant throughout history [4], it has been scaled up, diversified, and accelerated by the dramatic growth of social media. The goal is to recognise and consolidate the exclusive mechanisms privy to social media, that facilitate rumour circulation.

To this end, after elaborating upon the notion of rumours in Section 2, the rest of this chapter investigates the role of different communication technologies on rumour spreading and then inspect a distinctive mechanism of recommendation system in social media. A set of experiments is developed to assess the role of this mechanism. It investigates the role of YouTube recommendation system in the promotion of conspiracy theories and leading the users toward rumours rabbit holes.

3.2. COMMUNICATION AND RUMOUR SPREADING

Communication is an indispensable aspect of life; from the tiniest organisms to the biggest ones, they communicate with each other in order to serve a specific function. Viruses listen to their relatives when deciding how to attack their hosts [82], fish transmit signals to their rivals during aggressive displays [83], and giraffes make noise to locate each other [84]. Similarly, human beings also communicate with each other for a variety of reasons, such as emotional sharing, persuasion, and information exchange [85].

It is challenging to define the notion of communication satisfactorily as it comprises a broad range of activities. Communication is talking to each other; it is watching television; it is our outfit: this list is endless [86]. Nearly every book on communication studies offers its definition [87]; however, “arriving at a “best” definition has proved impossible and may not be very fruitful” [88]. In fact, Dance and Larsen [89] surveyed 126 definitions for this concept until 1976, and since then, even more definitions are formulated [87].

Defining communication as a process brings us closer to its complexity. It captures the dynamic aspects of communication and represents this process as a never-ending and ever-changing phenomenon [87]. One of the most influential process-based communication models is the one developed by Shannon and Weaver [90, 85].

In this model, the configuration of the communication may vary which means the sender and receiver might be singular or plural entities; therefore, the sender-receiver interaction is either one-to-one, one-to-many, many-to-one, or many-to-many. Besides, the scale and the demographic composition in the sender and/or receiver may vary. The

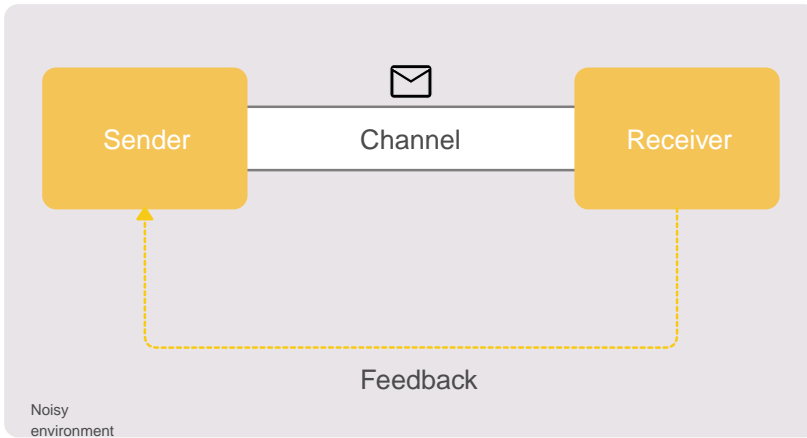


Figure 3.1: Communication process [85].

communication channel might be synchronous or asynchronous. In the synchronous channel, both sender(s) and receiver(s) are active and engaged. Conversely, when the channel is asynchronous, the sender initiates the message, and the recipient receives it sometime later [90]. Although the Shannon-Weaver model does not capture some aspects and nuances of rumour spreading such as the context of emergence and functionality, the broad reading of this model from the communication process allows us to tailor it for the phenomenon of rumour spreading. As it is discussed in chapter 2, rumours are the circulation of relevant and unconfirmed messages as sense-making and threat management mechanisms in response to uncertainty and situational ambiguity. By adaptation of Shannon-Weaver model, the rumour would be a sequence of unconfirmed and relevant message transmission among a group of sender/receiver individuals who participate in the rumour in order to make sense of the ambiguous/uncertain situation.

The way that the introduction of new mediums may affect the communication process and consequently rumour process is by expanding the modes of communication (e.g. changing the synchronicity or configuration of the communication). If the new mediums provide us with the opportunity of rapid access to a big and diverse audience across the world, this means rumours would be able to circulate with the scale, speed, and scope that have not seen before. The next section investigates the role of technological development in the rumour prevalence [18, 19].

3.3. THE ROLE OF COMMUNICATION TECHNOLOGIES IN RUMOUR SPREADING

In this section, we discuss how the introduction of new communication technologies could facilitate message transmission and thus rumour dissemination. The development of communication technologies is a complex and continues process. Here we slice up this developing process and take snapshots of few technologies with significant importance in rumour dissemination.

3

Our starting point is the pre-printing era when there was no synchronous and mass communication technology, and the possibility of long-distance message transmission was quite limited [91]. Within this period the communication was mostly coordinated locally and in small-scale. Thus, rumours were also often about local issues and remained within the communities. The invention of the printing press was a turning point in communication technologies as it made mass communication possible [92]. This technology increased the chance of exposure to rumours by allowing to share the same message among many people [93, 71, 72, 94, 95, 42].

Another crucial technology was the telephone which introduced synchronicity to the communication process. Telephone could accelerate and expand the rumour circulation by offering fast long-distance communication. The key distinctive features of the printing press and telephone technologies (i.e., synchronicity and mass communication) were later on incorporated into radio and television and created synchronous mass communication mediums. It was the first time in history that distant live communication with masses became possible. They could also draw more attention because of multimedia elements. Besides, compared to written media, the radio and television comprised a wider audience since even people without the ability to read could understand it. Due to the above-mentioned features, radio and television were extensively utilised in rumour spreading [96, 97, 98, 99, 100, 101, 102, 103, 104, 105].

The mass spread of rumours began by the emergence of the World Wide Web (WWW). Within this period, the distant synchronous/asynchronous multi-lateral communication was permitted [90]. Through the forums, chat rooms, and other WWW-based applications, people could communicate without even knowing each other. This would allow individuals to hide their identities or use anonymous or even fake avatar in their profiles [20, 106]. Minimal supervision is another feature of the WWW that fostered rumour spreading. Despite the majority of the communication mediums that had previously been available only to a marginal, self-selected group of people who were somehow linked to media outlets, WWW created a free venue for the people to express their thoughts and opinions (in the forums, chat-rooms, their blog or website) no matter who they are and how credible is their messages [17, 105, 107, 108, 109, 110, 111, 112].

The rise of online social media platforms was the landmark in the history of rumour spreading. This technology supports distant multi-lateral communications with different synchronicity modes. Social media is a complex phenomenon that offers novel features in three layers of social, institutional, and technological. From social perspective, it is an enormous hyper-connected network [20] pursuing power-law degree distribution [113, 114, 115, 116]. The large scale of social networks vastly increases the number of people who might be reached. It also increases the chance of creating communities with

similar values, beliefs, and interests [117, 118]. This turns social media to a suitable environment for the shaping and organising movements in a participatory, leaderless, horizontally coordinated, and ad-hoc manner which is an ideal setting for the conspiracy-based movements since it allows people to easily find like-minded peers and freely communicate their controversial thoughts. Besides, within social media, high-degree nodes or hubs which play the role of influencers and opinion leaders affects the small-world property in the network which eventually lead to the virality of information and rumours [116, 13, 119]. Furthermore, when the size of a network increases, the chance of diversity among the users improves and that would expand the scope of potential rumours [20].

From the institutional perspective, social media platforms allow their users to participate in information dissemination while leveraging anonymity. It specifically gives a safety margin to the initiators or moderators of inorganic rumours [18, 120] since they can “say whatever they want, whenever they want, and yet be shielded by anonymity” [121]. The platforms also enable the democratisation of the content by allowing individuals to consume, create, and distribute their content without governmental control [122, 123] and with minimal supervision [17]. This means people of different ages, education, and nationality are free to share their thoughts, and discuss their ideas about a variety of topics in politics, sport, or trivial daily-basis incidents, to name but a few. They can produce and share whatever content they want as long as it does not violate the platforms guidelines which are developed at a minimal level not to curtail freedom of speech [106, 124]. This policy would lower the barrier to entry for not only those who have not received any training on journalism but also for the ones who bluntly reject the journalistic norms of objectivity and balance [17]. Besides, any attempt to control such an extravagant system is perceived as the censorship, since social media platforms are considered as the manifestation of freedom of speech and whatever that restricts this sphere will be interpreted as a violation of freedom of speech. That is why codes of conduct, style guides, and journalistic guidelines in online social media platforms are in the minimum level [81]. This is an ideal setting for rumour spreading because the information is not verified before releasing in social media.

From technological perspective, social media are equipped to mechanisms such as recommendation systems and social bots that are alleged to facilitate the spread of rumours [25]. Recommendation system or web personalisation is a specific type of algorithms which is used to enhance user experience by reducing the information overload [125] and helping users to find compelling content in large corpora by personalised suggestions [126, 127, 128]. In addition, they benefit service providers by bringing business value to them as well. The recommendation systems are widely used by social media platforms to tailor the enormous amount of contents available in the platforms. The other automation mechanism in online social media platforms is social bots. They are computer programs that tend to emulate and alter human behaviour and produce content and interact with other humans [129]. Social media bots engage in commercial activities by facilitation B2C relations, including selling of products or services. They can also function as a notification machine and automatically capture breaking news, events, and incidents. The other functionality of social bots is the promotion of participation and engagement in social and civic activities.

Both social bots and recommendation systems are alleged to play central role in

the spread of rumours through amplification of the messages with strategic goals and leading people toward rumours rabbit holes, respectively. For the case of social bots, there are studies that show the significant role of bots in the circulation of rumours. It is claimed that they drive the diffusion of rumours by liking, sharing, and searching for information. Particularly, they are responsible for substantial amount of contents during political events such as the 2016 U.S. Presidential election and 2017 French election, to name but a few [17, 25, 24]. On the other hand, there are studies, including the largest research on the digital spread of rumours that show the insignificance of social bots compared to humans in the spread of rumours [13]. Although controversial on the surface, all those paradoxical results might be part of a bigger picture [81] which currently does not exist. Similarly, in the case of recommendation systems, some studies show the effectiveness of such systems in the circulation of rumours [130, 131, 132], while some others raise doubt about the role of recommendation systems in the spread of rumours [132].

3.4. RECOMMENDATION SYSTEMS UNDER SCRUTINY

In recent years, academic critics such as Zeynep Tufekci and technological whistleblowers such as Guillaume Chaslot have raised the alarm about technological scaffolds that have the potential to radicalize the people who interact with them [133, 23]. Anecdotal reports of YouTube recommending fake news, conspiracy theories, child pornography, and far right political content have cropped up in North America, South America, Europe, and elsewhere [134]. Ribeiro et al. examined alt-right recommendations on YouTube and found that there is a pathway from other types of content to these topics [135]. However, to date, these concerns have been speculative and anecdotal, as there have been no systematic studies of the promotion and amplification of conspiracy theories² via the YouTube recommender system. In this chapter, we fill that gap with a large-scale, pre-registered exploration of the YouTube recommender system. At first YouTube data collection and data annotation are explained in detail, then the results are reported and discussed.

3.4.1. METHODOLOGY AND DATA COLLECTION

On YouTube, when a user enters a search query, the YouTube search system returns the most relevant (i.e., most likely to be watched to the end by an account with this digital footprint) videos regarding the user's query. After the user chooses one of the results, YouTube launches two separate yet closely connected operations: (i) showing the video panel and meta information, and (ii) recommending further relevant videos. When the user clicks on one of the subsequently recommended videos or lets it automatically play, the same scenario repeats, and the requested title is displayed alongside still further recommended videos. If the video is watched to the end, the top-recommended clip is played next. Alternatively, the user may click on any of a list of recommended clips.

*** To access the codes and data used in YouTube project, please refer to https://osf.io/cjp96/?view_only

²As we discussed and elaborated in the previous chapter, the genesis of different variations of unverified information might be for different reasons; however, they pursue a similar dissemination dynamic which causes they look like rumour spreading after the first generation of the transmission. Therefore, although conspiracy theories may emerge as a coping mechanism, after the circulation and in the long-run it pursues the same dynamic of rumour spreading [48].

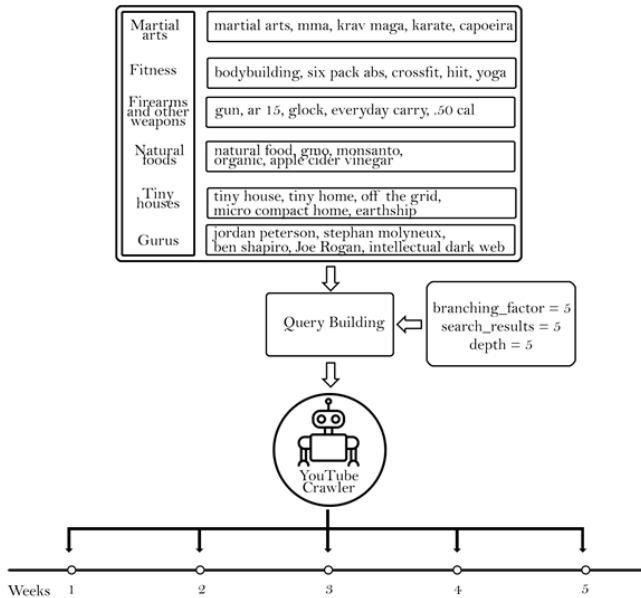


Figure 3.2: The schematic flow of data collection.

On a typical laptop or desktop screen, five or six recommendations are visible without scrolling.

In 2016 an ex-YouTube engineer Guillaume Chaslot developed a program to investigate the YouTube recommendation system during the U.S. presidential election 2016 [136]. This program simulates the behaviour of users on the YouTube platform by starting from a given search query and going through the recommended videos recursively until a predetermined level has been reached. The program has two modules: the crawler and the analyser. The crawler module collects the search results and recommended videos from YouTube, and the analyser stores, ranks, and visualises the results.

Figure 3.2 gives a schematic version of the crawler's operation, which is analogous to a breadth-first search (BFS). First, the user initialises the crawler by providing the search query (q), the number of search results from the search query to begin with (k), the branching factor (b), and the depth of the exploration (h). For the search terms, it has already been shown that particular terms are more likely than others to lead to conspiratorial contents [137]. To build the search terms, we looked at the case of Buckley Wolfe [138]. It suggests martial arts, fitness, firearms, and gurus as potential starting points toward conspiracy theories. All of these topics are stereotypically masculine³ and right-wing. We also wanted to add other potential search terms that are neither masculine nor right-wing. After discussion with the authors, we reached to two additional terms of natural foods and tiny houses which are associated with anti-capitalism and concerns about environmental and climate impact.

³All of the gurus are men, and many of them also have disproportionately male followings.

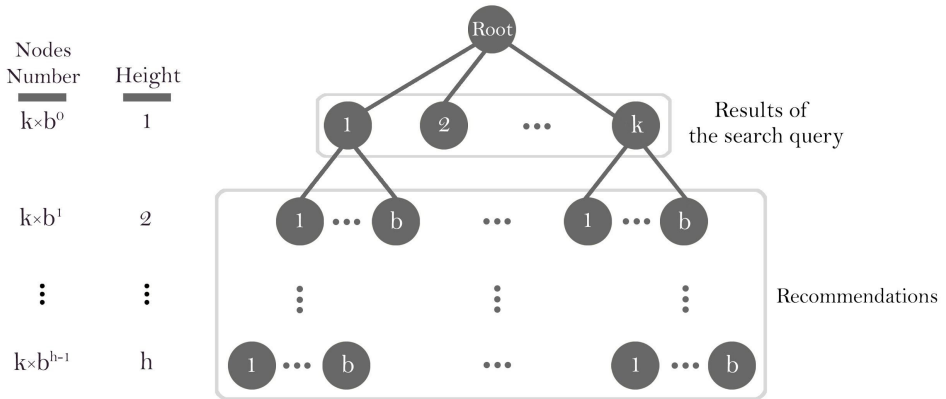


Figure 3.3: The YouTube recommendation tree when all the recommendations are distinct. In the case of the same recommended videos, the structure will be a directed graph.

Figure 3.3 illustrates a simplistic version of the crawler algorithm, where there are no duplicate recommendations. In such a situation, the robot starts with search query (q) results and obtains the first k videos that the YouTube search engine returns in response to the search query. Then, for every one of those videos, the robot collects the recommended videos and selects the top b recommendations recursively until it reaches the desired depth. In this case, the robot collects $S = \sum_{i=1}^h kb^{i-1}$ videos, including k initial videos from the search query and $S - k$ distinct recommended videos. In reality, many of the videos suggested by the YouTube recommendation system are the same, which makes the recommendation structure a potentially cyclic directed graph rather than a tree. In such a structure the number of recommended videos is $S < \sum_{i=1}^h kb^{i-1} - k = k \sum_{i=2}^h b^{i-1}$. After reaching the h^{th} level, the crawler module stops and the analyser module takes control. This module receives URLs for all the collected videos as well as their corresponding metadata and stores them in a predefined path.

For this project, we operationalise each of six topics with five search terms, meaning that we have a total of thirty seed searches. Then we launch the crawler five times for each seed search. The data collection took place over five weeks between August and September 2019.

To more faithfully replicate the conditions of someone like Buckey Wolfe, we also used a virtual private network (VPN) to simulate the searches and recommendations as if the user were based in the United States (in particular, in St. Louis, Missouri). The initial arguments are set to $k = 5$, $b = 5$, and $h = 5$. Each progressive stage of the process thus increases the number of returned videos by a factor of five, meaning that — for each search term — we end up collecting information about $5 + 5^2 + 5^3 + 5^4 + 5^5 = 3905$ videos. Since there are five search terms associated with each topic, this means we collect information about 19,525 URIs associated with each topic, for a total of 117,150.

Naturally, this is too many videos to code by hand, especially given that many of them are as long as three hours or more. For this reason, three independent coders evaluate the 100 most-recommended videos for each topic (600 videos total), assessing them on

the three-point scale described above. In this context, most-recommended status is determined by calculating PageRankIn for each clip [139]. PageRankIn represents the probability of landing on a particular node by following a random walk through the network, which means that it identifies the basins of attraction in the network of recommendations.

3.4.2. ANALYSIS

For the experimentation, we first measure the similarity across topics and search terms, then we build a topic model on the transcripts of coded videos. We then investigate statistics of collected videos for each topic. Afterwards, we do two major analysis of this research namely, (i) analysing the distribution of conspiracy theories among the most-recommended clips from each topic and (ii) measuring the fraction of top-recommended videos discovered at each stage of data collection.

Figure 3.4 is a similarity matrix that represents the overlap of recommendations across topics. This matrix is created based on the Jaccard similarity index. To calculate this metric, for every pair of the topics, the number of common videos in both topics is divided by all the videos associated with those topics (i.e., the intersection is divided by the union). The darker the box, the more overlap. Martial arts is more associated with fitness and firearms than with natural foods, tiny houses, or gurus. Fitness is most associated with natural foods. Firearms are somewhat oddly associated most with natural foods. Natural foods are most associated with tiny houses, and vice-versa. Gurus are somewhat more associated with natural foods than other topics. This may be due to the fact that Jordan Peterson tends to promote his medically contentious diet of eating only red meat⁴.

Figure 3.5 represents the similarity of recommendations across search terms, which suggests that some of our topics are more internally consistent than others⁵. Tiny homes and gurus are the most internally consistent, followed by firearms.

Table 3.1 lists some representative titles among the most-recommended clips in each category. As these examples indicate, many of the most highly-recommended clips have sensationalising titles that make use of all-caps, exclamation points, and other standard clickbait devices⁶.

We now turn from the full dataset to the 100 most-recommended clips for each topic. As an exploratory step, we built a topic model on the transcripts for the coded videos. Transcripts were retrieved using the YouTube API. Of 600 videos, 480 had transcripts available (some auto-generated, some user-entered). Transcripts were preprocessed to remove non-alphabetic material, common English stop words, words fewer than 3 characters, and descriptions of on-screen text. The resulting transcripts were then lemmatized using nltk [140]. Lemmatized transcripts were transformed into a tf/idf representation ($min_df = 0.05$, $max_df = 0.96$), and a range of topic models were built using

⁴See <https://www.theatlantic.com/health/archive/2018/08/the-peterson-family-meat-cleanse/567613/>.

⁵This matrix is created using the same method the same as the previous one. The only difference is the unit of analysis, which is more fine-grained in Figure 4. Here, instead of calculating the similarity between every pair of topics, we perform all the calculations at search-term level.

⁶For instance, they tease a revelation without giving enough details to form reasonable expectations. Which three common mistakes are made in street fights? What is the secret to mastering a handgun? What strange secret, Earl Nightingale? YouTube content creators share YouTube's interest in selling advertisements, so it is unsurprising that some of them are desperate to draw attention and curiosity with their video titles.

Topic	Example titles
Martial arts	<ul style="list-style-type: none"> • 20 MOST EMBARRASSING MOMENTS IN SPORTS • 3 Common Mistakes in a Street Fight - Bruce Lee's Jeet Kune Do • The Gracie UFC Conspiracy
Fitness	<ul style="list-style-type: none"> • WE TRIED KETO for 45 Days, Here's What Happened • The ONLY 3 Chest Exercises You Need for MASS (According to Science) • The mathematics of weight loss Ruben Meerman TEDxQUT (edited version)
Firearms	<ul style="list-style-type: none"> • Improvised Suppressors for .22 Rimfire • 223 -vs- 5.56: FACTS and MYTHS • The Secret to Mastering a Handgun
Natural foods	<ul style="list-style-type: none"> • Strange answers to the psychopath test Jon Ronson • Interstellar Travel: Approaching Light Speed • The Revelation of the Pyramids (Documentary)
Tiny houses	<ul style="list-style-type: none"> • The basics on a Speed square • Off-Grid Tiny House TOUR: Fy Nyth Nestled in Wyoming Mountains • Surprise! Awesome figured maple (I DID NOT EXPECT THIS!!!)
Gurus	<ul style="list-style-type: none"> • Jordan B. Peterson Full interview SVT/TV 2/Skavlan • Proven Biblical Money Principles - Dave Ramsey • The Strangest Secret in the World by Earl Nightingale full 1950

Table 3.1: Representative titles from all six categories.

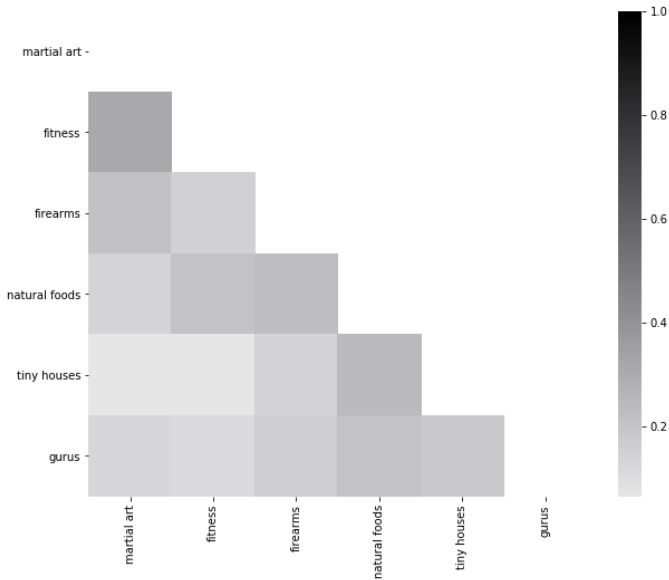


Figure 3.4: Similarity of recommendations across topics.

non-negative matrix factorization (NMF), one transcript per document. As this was an exploratory analysis on a relatively small number of documents, a 12-topic model was chosen by manual inspection as the solution that maximised discriminability while minimising intruders.

Table 3.2 presents the results of the topic model. On the right are longer representations of each topic. The heatmap on the left shows, for each pair of topic and group, the percentage of transcripts that had that topic as their maximum normalised loading compared to the overall percentage of documents that had that topic as the maximised loading (ratios below 1 are cut off to improve visibility). Intuitively, this shows the extent to which a topic is over-represented in a group relative to the whole set of transcripts.

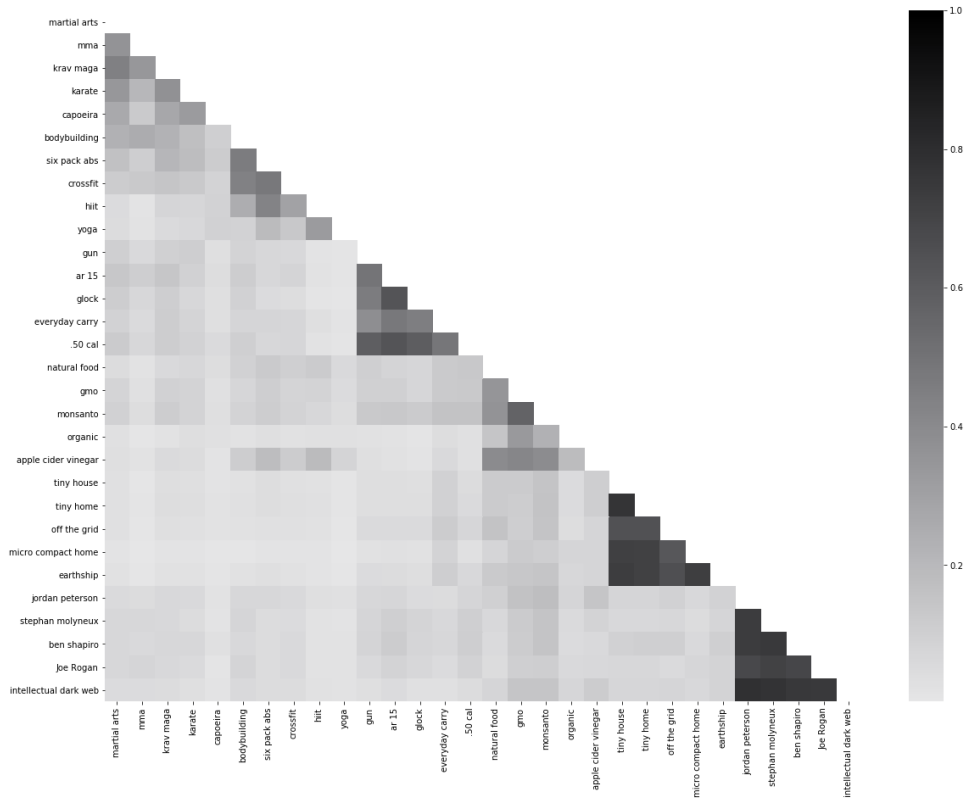


Figure 3.5: Similarity of recommendations across search terms.

Category	Views	Likes	Dislikes	Length in seconds
martial arts	11,094,353	81,015	6,381	908.3
fitness	6,029,494	67,272	3,883	1409.94
firearms	4,746,370	41,300	3,086	1185.96
natural foods	6,727,368	96,167	4,939	1253.47
tiny houses	3,505,384	40,158	2,348	1619.3
gurus	4,839,331	70,172	3,939	3398.25

Table 3.3: Summary statistics for each topic. All figures report averages (means).

Some of the results are unsurprising. Firearms, fitness, martial arts, and tiny houses each have a unique characteristic topic, one which loads on words that one would expect from those videos. This shows that the topic model was able to extract sensible patterns from the data. Natural foods also has a unique high-loading topic, but one which appears to emphasise a common core of fringe scientific ideas. This suggests that the popular videos in natural foods are not unified by their particular recommendations so much as their adherence to a loose set of beliefs that are used to justify their content.

The pattern seen with the guru videos differs from that of the other five. The two topics that guru videos load most heavily on are “rhetorical” topics which are characterised by a manner of speaking — one more congenial, the other more angry. In other words, what appears to be most characteristic of guru videos is not a specific content but a more general manner of speaking. Insofar as there are similarities, they are mostly with the fitness and martial arts categories, suggesting perhaps a rhetorical style more broadly associated with an exaggerated masculinity. This exploratory may be worth further investigation.

Table 3.3 provides summary details for each of the 100 most-recommended clips. The most-viewed topic was martial arts, followed by natural foods, fitness, gurus, firearms, and tiny houses. The most-liked topic was natural foods, followed by martial arts, gurus, fitness, firearms, and tiny houses. The most disliked topic was martial arts, followed by natural foods, gurus, fitness, firearms, and tiny houses. The longest videos were associated with gurus, followed by tiny houses, fitness, natural foods, firearms, and martial arts.

Each of these 600 clips was independently coded by three different coders according to the scheme described above. We observed adequate interrater reliability (Fleiss’s $k = .445$, $z = 27.5$, $p < .0001$). To arrive at finalised ratings, we used the following decision procedure. First, if all three raters agreed, then their consensus was entered as the final rating of the clip. Second, if two of three raters agreed but the third disagreed, then we entered the value agreed-upon by the majority as the final rating of the clip. Finally, if all three raters disagreed (meaning that the clip received scores of 1, 2, and 3), one member of the research team reviewed the clip a second time and came to a final conclusion. Such maximal disagreement occurred in just 14 out of 600 cases (2.3%). Figure 6 represents the severity of conspiracy theories among the 100 most-recommended clips from each topic.

As Figure 3.6 makes clear, the YouTube recommender system does indeed promote conspiracy theories from all six topics. However, the proportion and severity differ from

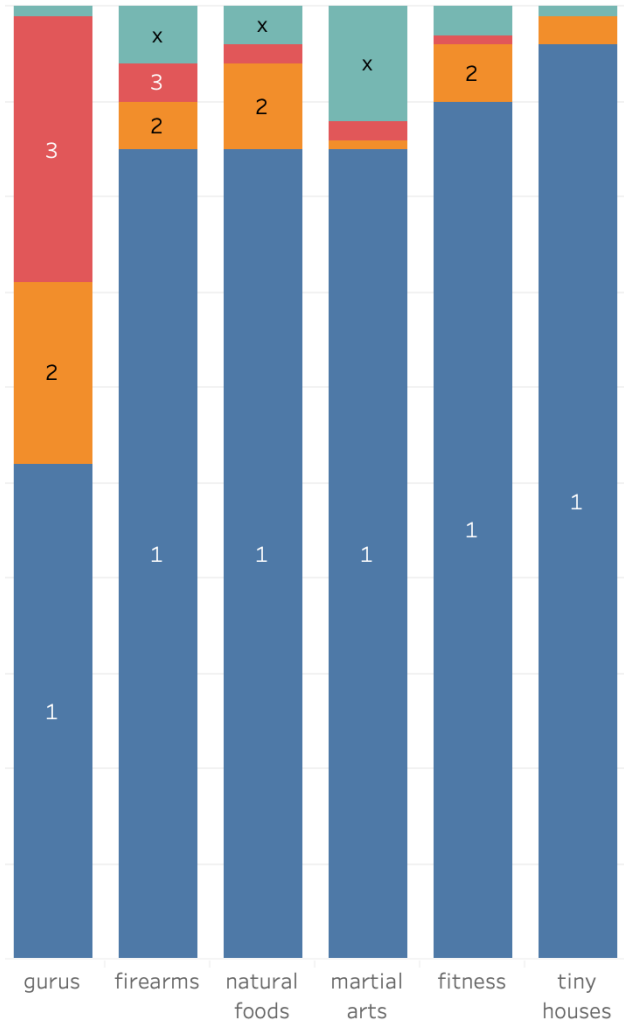


Figure 3.6: Distribution of conspiracy theories among the most-recommended clips from each topic. 1 = no conspiracy theory, 2 = mild conspiracy theory, 3 = severe conspiracy theory, x = clip no longer available at time of coding.

Topic	Ratio
gurus	.475
natural foods	.115
firearms	.096
fitness	.072
martial arts	.034
tiny houses	.030

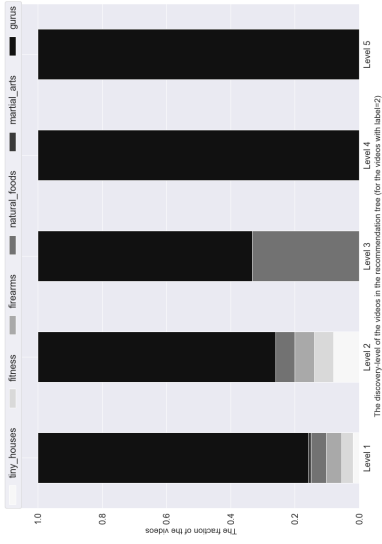
Table 3.4: Ratio of conspiratorial clips to rated clips for each category.

topic to topic. One of the questions that may raise is whether the recommender system promote more conspiracy theories than some sort of neutral baseline? We are unable to address this question in the current study because we have no way of ascertaining what a neutral baseline might be. It might be possible to compare one recommender system to another, or to compare this recommender system to an older version of the same recommender system. However, we lack access to these comparators. What we have established is that the YouTube recommender system does in fact push conspiracy theories, not that it pushes them harder than they would be pushed by an alternative. Thanks to an anonymous reviewer for raising this point.

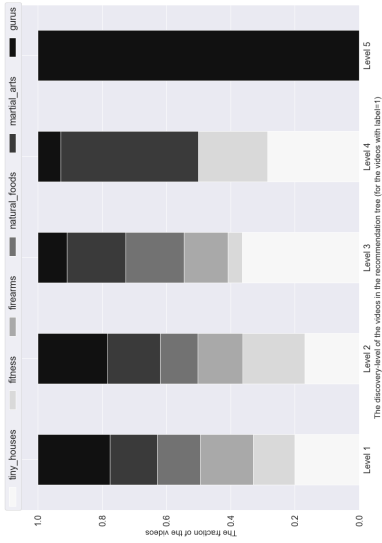
To test our main hypothesis about the proportion and severity of conspiracy theories, we calculate for each topic the ratio of the number of clips that received a rating of 2 or 3 to the number of clips that received a rating of 1, 2, or 3. This leaves out clips that received a rating of x . This analytic method may miss some information, but because the list of reasons that a clip might be unavailable is so diverse, we decided not to presume that unavailable clips were or were not conspiratorial. The resulting ratios are represented in Table 3.4.

Remarkably, nearly half of the visible most-recommended videos from the gurus topic were conspiratorial. The other topics seem less worrisome, though still problematic. Over 10% of the visible most-recommended videos from the natural foods topic were conspiratorial, as were nearly 10% of the videos from the firearms topic. Thus, the most conspiracy-heavy topic by far was associated with the political right, and the next two were split between natural foods and the right firearms. It is fairly clear that firearms are associated with the political right; natural foods might seem like a left-wing interest but is politically ambiguous, as the film *Dr. Strangelove* illustrates with its gag about “precious bodily fluids.”

Recall that our pre-registered hypothesis was that the proportion of conspiracy theories associated with different topics would be ordered as follows: gurus = firearms > natural foods > martial arts > fitness > tiny houses. This hypothesis is largely borne out. The actual ordering is gurus > natural foods > firearms > fitness > martial arts > tiny houses. In other words, the second- and third-ranked items as predicted turned out to be the third- and second-ranked items in the actual data, while the fourth- and fifth-ranked items as predicted turned out to be the fifth- and fourth-ranked items in the actual data. The top item (gurus) and the last item (tiny houses) were correctly predicted. As an exploratory analysis, we also examined the clips with each rating to see which stage of data collection they cropped up in. Figure 3.7 shows these results.



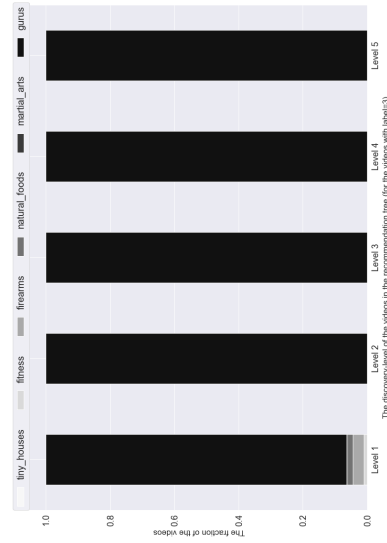
(a) Videos labeled 1



(b) Videos labeled 2



(c) Videos labeled 3



(d) Videos labeled 4

Figure 3.7: Fraction of top-recommended videos discovered at each stage of data collection.

As figure 3.7 shows, conspiratorial content can appear in any depth of the recommendation tree/graph. For some topics, they only appear in early levels, while for other topics conspiratorial contents keep showing up in deeper levels. Second, between mild and severe conspiracy theories, the earlier has a higher chance of appearance in deeper levels of the recommendation tree/graph. In Figure 3.7c, there are severe conspiracy theories from different topics only in the first level (and in other levels the conspiracy theories belong to one particular topic). By contrast, in Figure 3.7b, mild conspiratorial content belonging to different topics are available in the first three levels. Third, as we can observe in Figure 3.7a, until level 4 videos are coming from 4 different categories, which means even in deep levels of recommendation tree/graph we should expect to see non-conspiratorial contents. Finally, in general, the guru videos tend to be recommended more the deeper one goes into the tree. Level 5 (the final bar of each histogram) is almost all gurus for every single label. This suggests that the “rabbit hole” effect is especially pronounced in the case of guru videos.

3.5. CONCLUSION

This chapter started by raising an important question: how does social media streamline the dissemination of rumours. To address this question, the spread of rumours since the pre-printing press era till social media era is taken into account. The focus was especially on social media since the spread of rumours scaled-up, diversified, and accelerated during this period. Social media covers a variety of communication modes and also offers a set of unique features such as enormity, communality, specificity, and virality that facilitates the mass spread of rumours. One of the novel features of social media which is not the result of hyperconnectivity in social networks and is proposed and developed by social media platforms is the automation mechanism. One of the automation mechanisms is the recommendation systems.

It is alleged to play a central role in spread of rumours through amplification of the messages with strategic goals and leading people toward rumours rabbit holes, respectively. It is a controversial topic with conflicting results. Some scholars claim that they indeed reinforce the spread of rumours, while some others disagreed and showed opposing evidence. To tackle this issue, a set of experiments is set-up to investigate the role of recommendation systems in online social media platforms. In this study, the first large-scale, systematic, pre-registered attempt to establish whether and to what extent the recommender system tends to promote such content is carried out. The results show that the YouTube recommender system does indeed promote conspiracy theories. However, the proportion and severity differ from topic to topic. The recommendation systems in online social media platforms have small a clear impact on the spread of rumours; however, this effect may mediate by variety of factors such as location, time, and rumour topic.

4

COUNTERING RUMOURS

Whoever fights monsters should see to it that in the process he does not become a monster.

Friedrich Nietzsche

The first two chapters investigated the conceptualisation of false and unverified information and the role of social media as a major rumour spreading environment. This chapter aims at doing a comprehensive review over past counter-strategies against rumour spreading. To this end, the following research question is posed:

- *What is the current status of rumour response strategies?*

To address this question, we first collect counter-strategies pertaining to different phases of communication. For the next step, we introduce the epidemic framework in order to evaluate the efficacy of counter-strategies. We inspect past strategies employed in addressing rumour dissemination and use the framework to explore parallels between epidemic management and addressing rumour. We identify the highly neglected aspects of the current cumulative rumour response and factors that may be effectively targeted in the future. Our approach might support understanding social media's role in propagating rumours and devising active measures in quelling this epidemic¹.

¹This chapter is based on the following publication: Fard, A. E., & Lingeswaran, S. (2020, April). Misinformation Battle Revisited: Counter Strategies from Clinics to Artificial Intelligence. In Companion Proceedings of the Web Conference 2020 (pp. 510-519).

4.1. INTRODUCTION

In the previous chapters, the concept of rumour and the medium of spread were thoroughly investigated. From now on, the focus of this thesis would be on tackling the rumour circulation. This chapter tends to briefly review the past attempts to counter rumours and set them into a common framework to study their strengths and weaknesses easier.

Although rumour spreading and its potentially destructive effects have been taken into account since ancient times, it was only less than a century ago that the first systematic efforts against the mass spread of rumours began [4]. Since then, a variety of techniques have been exercised by the media organisations, academic institutions, and recently online social media platforms. Even governments stepped forward and developed strategies to counter rumour circulation.

Despite the scholarly field of rumour studies has seen tremendous advancement in the development of counter-strategies [4, 14, 3], the massive waves of rumours are still sweeping over individuals [8], organisations [7], and societal institutions [13]. One of the explanations for this situation is the massive increase in the scale, scope, and speed of spread of rumours on the one hand, and lack of a comprehensive plan to confront this phenomenon on the other hand. In order to develop an effective and comprehensive plan to quell rumours, in addition to come up with novel techniques, it is crucial to be aware of the past counter strategies and potential capabilities. To this end, we collect the counter strategies over the past century and analyse them using an epidemic control framework [141]. The result of the analysis would allow us to understand, within the past century, what aspects of confrontation with rumour have been targeted extensively and what aspects are highly neglected.

The rest of this chapter is organised as follows. The next section quickly reviews the counter rumour strategies from the communication perspective. After explaining the strategies, Section 4.3 evaluates them based on the epidemic control framework and determines how each of the strategies is confronting the rumours. In this section also possible weaknesses of the past counter-strategies are discussed.

4.2. COUNTER RUMOUR STRATEGIES

This section reviews the rumour control strategies from the communication perspective. As it is discussed in chapter 2, a communication process is composed of three major elements of sender, channel, and receiver in which senders transmit messages to receivers through communication channels. As Figure 4.1 represents, the past counter rumour strategies are classified based on their impacts on either of those elements. The first group of strategies takes those who initiate rumours into account and aims to restrain them (in case of deliberate rumour spreading), the second group tends to secure communication channels and minimise the likelihood of rumour emergence and circulation within the communication channels, and the purpose of the third group is to protect those who were targeted by the rumours. In the following, the strategies are explained in more detail.

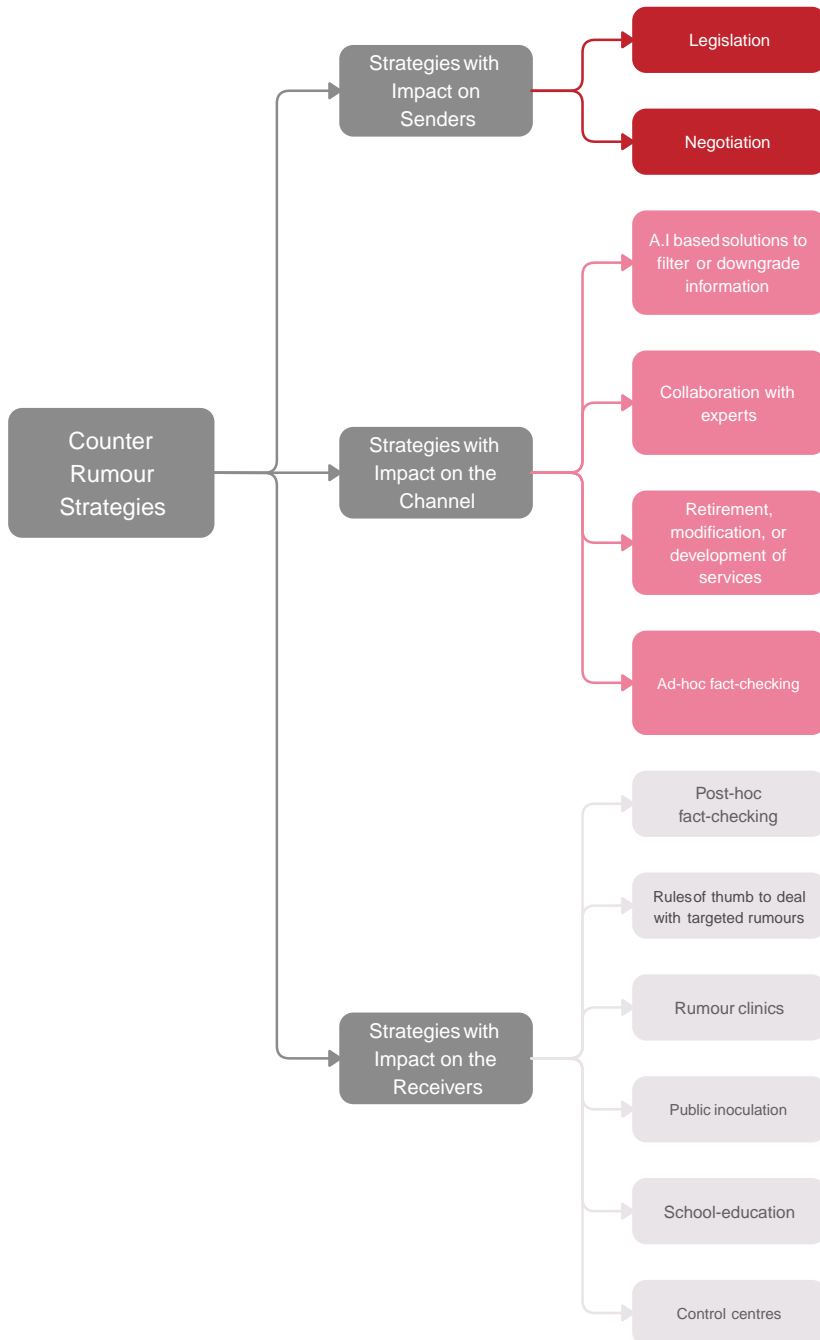


Figure 4.1: Rumour counter strategies overview

4.2.1. SENDERS STRATEGIES

This section tends to discuss counter-strategies that are developed to confront those who deliberately participate in the rumour process. To this end, two main approaches of legislation and political solutions are proposed.

LEGISLATION

The legislation is one of the oldest state-level strategies against those who willingly contribute to the spread of rumours. The earliest form of legislation in this arena is the defamation law which may cause lawsuit against defamers. The main purpose of defamation law is to protect the reputation of an entity against libel. Although defamation laws are well established and thus a very effective (and obvious) instrument to directly confront and discourage the spread of targeted rumours, they are also very costly and have a long time horizon. However, lawsuits can draw media attention and thus, it can strengthen public awareness by revealing the latent actors and their techniques in the diffusion of rumours. One of the important points that need to be taken into account is to protect those who pursue and win the lawsuits against negative repercussions from the defendant [12]. By the growth of online social media platforms, the expansion of legal actions from conventional mediums to online social media gradually started. Germany was among the first countries taking online environment into account [142, 143, 144, 145, 146, 147]; however, the number of countries with regulation regarding rumour spreading in online environment is increasing [148, 149, 150, 151].

NEGOTIATION

Social manipulation campaigns are becoming an important tool for information operations. On the one hand, different countries and organisations use this approach to influence public opinion. On the other hand, platforms and media organisations try to flag such operations and take down the accounts and messages linked to those operations [152]. This looks like a cat and mouse game since there are always new manipulation campaigns while platforms are ready to take them down. One of the strategies to confront or at least limit this endless cycle is to negotiate with the offenders and working toward an agreement with them. This is not always a feasible option due to lack of interest from either sides, unknown identity of the offenders, or unrealistic expectations; however, in case of a satisfactory situation, this is a promising approach with meager cost and high output [153].

4.2.2. CHANNEL STRATEGIES

In this section, the strategies concerning the protection of communication channels are discussed. Those strategies aim to make channels rumour-proof and minimise the likelihood of rumour emergence and circulation. They are proposed by either media organisations or social media companies. In the following, the strategies are discussed in detail.

ARTIFICIAL INTELLIGENCE BASED SOLUTIONS

Due to the number of active users and hyperconnectivity of the network between them, information spreads widely and rapidly throughout social media. One of the most efficient approaches to tackle rumours is Artificial Intelligence (AI) as it is fast and cheap.

Besides, it can tackle the spread of rumours at scale, across languages and time zones, without relying on reactive reports. The AI techniques are used for two main purposes of content filtering and downgrading.

Filtering Every minute, millions of messages are transmitted across online messaging applications, and thousands of posts are published on social network platforms. Among this massive flow of information, there are problematic contents which have to be filtered out before making public. Due to the massive inflow of information, manual inspection of all the contents seems impossible. What platforms often do is first to inspect the contents using machine learning models. In this step, if the algorithm can make a decision with very high confidence, there is no need for the second opinion, and the decision is deemed as final; otherwise, it is sent to the next step which is human judgement.

Downgrading We are living in an information-rich world, which wealth of information comes with a dearth of attention [154]. Rumours misuse this simple principle quite often and represent themselves as novel and appealing contents in order to catch eye-balls. In such a setting, the role of ranking systems is crucial. If rumours could manipulate such systems and elevate in search results or timelines, then they will most likely spread rapidly. However, if ranking systems could identify rumours and relegate them, their visibility and subsequently chance of spread reduces. Platforms have also taken this into account and tried to incorporate signals to their rankings systems in order to make it sensitive to rumours [155, 156, 157].

COLLABORATION BASED SOLUTION

In order to protect communication channels from rumours, the collaboration between experts from media organisations, social media platforms, and academia is essential. They all do have complementary expertise which together would help to reduce the chance of rumour emergence and circulation.

Collaboration with fact-checkers Truth finding is a crucial step to mitigate rumours in communication channels. Controversial articles are circulating in social media and news outlets, while users have no clue about their truthfulness. If those articles could get verified by evidence, they might not spread widely and would quell quickly. Fact-checking also helps AI solutions since it provides valuable training samples for the machine learning algorithms and makes them more accurate. Facebook is practicing the same strategy through collaboration with its userbase and independent third-party fact-checkers all around the world. If some shared contents need to be verified by the experts, independent third-party fact-checkers assess the flagged contents. The downside of such a fact-checking system is the lack of scalability. There are too many items that require verification, while the resources are limited. To address this problem, Facebook is going to expand its collaboration with fact-checking organisations and crowdsource it to the individual fact-checkers by giving the fact-checking privilege to some of its users [158].

Collaboration between news outlets and social media platforms Media literacy and digital journalism is an essential element in building resilient communication channels against rumours. Social media, along with major media organisations, have the potential to promote quality journalism in the digital era. To this end, major social media platforms and news outlets created partnership, initiated training programs, and even developed products and services to empower journalists [159, 160, 161, 162, 163].

Collaboration with academia In an academic-corporate relationship, collaboration is of the essence for both sides. It helps the academia to ensure industrial relevance in its research [164], and on the other hand, it provides the opportunity of knowledge complementary and risk-sharing with the corporates [165]. In the case of countering rumours, also both social media platforms and academia can benefit collaboration in multiple ways. Diffusion of rumours is a multifaceted phenomenon which originates in several scientific disciplines such as psychology, neuroscience, and computer science, to name but a few. Collaboration with academia allows platforms to have access to expert human capital in a wide range of disciplines. On the other hand, by access to unique datasets from social media platforms, academics will be able to test not only old social theories and hypotheses but also propose new ones.

4

DESIGN BASED SOLUTIONS

The design solution is another approach practised by social media platforms that aims to reduce the likelihood of platform misuse. For this approach, social media companies either retire their operational services, modify them, or develop new services.

Service retirement When the platform owners see more trouble than benefit coming from a service, they decide to retire that service. It is the most severe yet naive approach regarding a service. Although this approach eventually solves the platform misuse, it negatively affects many applications that their operations depend on the retired services. As an example of service retirement, in 2018 and after Cambridge Analytica scandal, Facebook started shutting down some of its APIs such as Events API, Search API, and App Insights API [166].

Service modification The more intelligent way of approaching platform misuse is service modification. Changing services in terms of closing back doors and making them more restricted reduces the chance of platform misuse. However, closing all the breaches and making a service abuse-proof takes enormous effort and much time. As an example of service modification, WhatsApp recently announced that based on their new policy, a message could be forwarded to the maximum five recipients [167].

New service development The newly designed services are essentially developed to reduce the chance of rumour emergence or propagation within the platform. The new services mostly aim to provide meta-information regarding the circulating messages. For instance, Facebook has recently developed a service called context bottom, which brings

more information about the articles that are shared on Facebook. These information include the publisher's Wikipedia entry, other recent articles they have published, information about how many times the article has been shared on Facebook, where it is has been shared, as well as an option to follow the publisher's page [155].

AD-HOC FACTCHECKING

Media organisations are in the front-line of combating rumours. They are the first provider of the news for their audience; thus, they have a huge responsibility in sharing accurate and impartial contents. To address that need, since the early 20th century, news outlets began a new practice to inspect the veracity of information before making public. Applying this procedure minimised appearance of rumours in news channels. Later on, this procedure is called ad-hoc fact-checking because it aims to eliminate errors before a piece goes live [168].

4.2.3. RECEIVERS STRATEGIES

This section tends to discuss strategies to protect rumour receivers. Here, the assumption is that rumours are out there and how the potential audience could confront it. In the following six counter-strategies are discussed.

POST-HOC FACT-CHECKING

In the wake of deceptive ads that populated 1988's US presidential race, a new procedure started to practise by news outlets in order to mitigate the consequences of rumours. It was called political fact-checking which was devoted to analysing the factual accuracy of politicians' statements. This journalistic practice is also called post-hoc fact-checking since it identifies and corrects errors after it goes public. Unlike the ad-hoc fact-checking, which aims to correct mistakes before making public, political fact-checking's goal is to correcting the rumours once it is already out there in the public sphere [169].

RULE OF THUMBS TO DEAL WITH TARGETED RUMOURS

When a rumour particularly targets an entity (e.g., an individual, an organisation, a country), then the spokesperson of the targeted entity should take a stance. Figure 4.2 displays possible ways of responding to the rumour. Initially, it should be decided whether to comment about the rumour or to ignore it. In case of commenting, then there are three avenues that could be followed: (i) Confirmation of the rumour and giving detailed information about it, (ii) Denial of the rumour and giving a rebuttal, or (iii) Withholding to comment about the rumour [170].

The other option is ignorance which is the weakest quelling strategy. It is highly unlikely that a rumour dies on its own because something that is incredible to someone may be deemed as plausible by another person. In other words, even if some people drop the rumour, some others are likely to pick it up.

Some rumours are fully or partially truthful. Confirming the truthful parts of the rumours reduce the chance of rumour generation [170]. One of the most common strategies to rein in the rumours is denial. There are a few factors that influence denial effectiveness. A denial should be based upon the truth. Dishonesty, especially a false denial is a dangerous strategy that may damage the credibility of the responsible party [170].

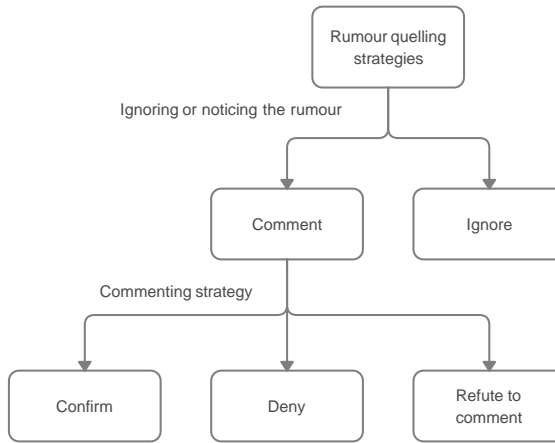


Figure 4.2: The quelling strategies for a rumour responsible party [170].

The denial effectiveness is enhanced when the denial source is trusted and is aligned with the rumour context. It is also very important to avoid repeating the rumour during the denial. The repetition fosters beliefs to the rumours. The denial should also be issued as soon as possible rather than after reaching a certain critical mass. Finally, a denial message should deliver a clear, detailed explanation with strong evidence indicating the rumour falsehood and convey to listeners a clear course of action about what they should do when they come across the rumour. A no comment response at best would work the same as ignorance strategy, and at worst gives more credence to the rumour. It reinforces the cover-up hypotheses and transfers the message that “we have something to hide” [7, 6, 170].

RUMOUR CLINICS

The rumour clinic was a response to the growing demand for a strategic solution to the problem of rumour dissemination during World War II. At its core, the rumour clinic is referred to a group of technical, representative, and prestige advisors to collect and analyse significant rumours and train the clinic participants about the tactics that rumours use to mislead people. The initial plan was to deploy and organise the clinics under governmental supervision; however, the fear of sharing the rumours with the public and the following inevitable repercussions make them extremely prudent in such a way that they ultimately withdrew from the project. Despite the government’s initial withdrawal from rumour clinics, they could not stand the idea of public rumour revelation and in less than a year launched a concrete effort to bury clinics. Their effort paid off in 1943 when the rumour clinic column no longer appeared in the newspapers [4, 171].

CONTROL CENTRES

One of the earliest systematic attempts against the mass spread of rumours was a telephone service called rumour control centre (RCC). It was, in fact, a follow-up on the rumour clinics which retired in 1943 [171]. RCCs were also recognised by other names such as “Rumour Central”, “Verification Centre”, or “Fact Factory” depending on their emergence location. They appeared amid the racial conflicts in the US in the 1960s to serve three purposes: riot control, riot prevention, and provision of information service to the general public. “Citizens were encouraged to call RCCs if they heard a rumor that suggested social tensions were increasing in their area. Working with local police and intelligence units, staff would try to locate the source of the rumor and test its veracity. The police could then take preemptive measures to address the unfolding situation by monitoring or arresting suspected agitators and spreading counter-information in areas of the city where the rumor was circulating” [172].

RCCs were inexpensive organisations to set-up. In its most basic form, what they required was a telephone, a few people, and some advertising. From the organisational point of view, in addition to rumour control staff to operate the telephones, there were communication hookups with police and fire departments as well as other city agencies to get the most recent updates regarding the incidents [37, 5]. From an institutional point of view, RCCs were adopting a similar institutional form by embodying within the government [172]. RCCs gradually disappeared in the early 70s due to funding and establishing legitimacy (in the eyes of the public) issues [37, 171]. One of the closest replacements to RCCs in the web era is the rumour control section in the Federal Emergency Management Agency (FEMA) website which provides reliable information about the running rumours during the disasters. This initiative confirms or denies major rumours that circulate during every disaster.

EDUCATION AND MEDIA LITERACY

Education is one of the most promising state policies against rumours. Open Society Institute calls education “the best all-round solution” and mentions “high-quality education and having more and more educated people is a prerequisite for tackling the negative effects of fake news and post-truth” [173]. This policy aims to improve media literacy and critical thinking among citizens. It would help them to reflect on the information they receive before believing or sharing them. Besides, it is a confrontation method with long-lasting effects. However, educational policies could be tricky, as too much emphasis on rumour might backfire and come with unintended consequences of undervaluing the real news outlets [17].

PUBLIC INOCULATION

Inoculation is a metaphor borrowed from biology and refers to the process of injecting the weakened doses of the virus which trigger the immune system to build up resistance against future infection. Inoculation in the rumour confrontation context also works in the same way, but instead of a virus, it confers resistance against influence and persuasion. This approach works like a rumour vaccine and aims to inoculate enough individuals, so the rumour does not have a chance to spread [174, 175]. Within the past couple of years, countering rumours using inoculation is increasingly drawing scholars atten-

tions. The primary reason for this is that debunking efforts and correcting rumours are not sufficient to stem the flow of online rumours [176].

This approach is composed of two primary steps. The first one is precisely similar to biological metaphor and comes with exposing the people to the weakened virus, which is in this case, information that challenges their existing beliefs or behaviours. It is worth noting that, as in the vaccination process, the weakened virus should not be so strong as to overwhelm the body's immune system [174]. Then, in the second step, one or more of the presented examples are directly refuted in a process called "refutational pre-emption" or "prebunking" [174, 177]. To improve the effectiveness of this approach, the public should also be vaccinated against the sources of misinformation, by drawing more explicit attention to exactly who is behind those information [12]. Although public inoculation studies have been limited to particular domains such as health, political campaigning, and climate change, the hypothesis is umbrella protection against the rumours regardless of the context [176].

Although public inoculation seems a promising strategy to preempt rumour campaigns, the implementation of it is shrouded in mystery. One of the most common implementation approaches is the collaboration of academics with reporters to echo the inoculation messages by their media. This approach can also be reinforced if elites and thought-leaders play an active role in the dissemination of inoculation messages [12]. Recently it is shown that serious gaming is a promising means to inoculate the public against rumours [176]. Additionally, in another research, critical thinking was introduced as an effective approach for public inoculation [177]. Currently, public inoculation is practised in experimental settings in academia. Expansion of this strategy beyond the academia would show its effectiveness in other domains and larger scale.

4.3. EVALUATION OF STRATEGIES

This section first presents a framework for the evaluation of the counter rumour strategies. Then, in the second part, the strategies that have already been introduced are assessed.

4.3.1. EVALUATION FRAMEWORK

The spread of rumours bears many similarities to the evolution and transmission of contagious diseases [178]. Almost half a century ago, Goffman and Newill [179] directed attention to the analogy between the spread of infectious diseases and the dissemination of information [141]. They argued that transmission of ideas do not need to be restricted to infectious diseases but is a more general process that might be applied to many contexts. For example, the development of the psychoanalytic movement in the early twentieth century was no less an epidemic than the outbreak of influenza in 1917 and 1918 [179]. This similarity between biological and intellectual epidemics is even caused the same modelling paradigm to be adopted in order to explain the dynamic of propagation [141, 180]. In the epidemiology, there is a control framework that has been successfully practised for reining in the epidemics [181]. It is composed of three mechanisms of (i) education, (ii) immunisation, and (iii) screening and quarantine. The first two are prevention measures which aim to minimise exposure to the disease and give a complete

protection to a person against infection, respectively while the third one has a more interventional nature with the purpose of reducing the transmission rate.

Education is one of the simplest and cheapest ways to control epidemics by training the public about simple techniques such as wearing masks, washing hands, social distancing, and gargling to reduce the likelihood of exposure to the disease. It is mostly about raising awareness about dos and don'ts regarding a particular disease. For example, in the case of AIDS epidemics, the educational campaigns in February 1987 tried to discourage risk-prone behaviours such as unprotected sex or needle exchange for drug users. The campaign was successful by reducing the spread of the virus in countries where educational campaigns were organised by the state or other organisations [181, 182, 183].

Immunisation through vaccination is one of the most effective and cost-effective strategies to control epidemics [184, 181, 185]. It is referred to as “one of the great public health triumphs of all time” due to achieving landmark gains over a relatively short period. For example, in the case of smallpox, a worldwide vaccination campaign succeeded in eradicating the disease. For instance, when the global immunisation program against diphtheria, pertussis, tetanus, poliomyelitis, measles and tuberculosis was initiated in 1974, only 5% of the world's children were fully immunised. However, in less than 20 years, more than 90% of the world's children had received BCG vaccine, and 75%–85% had received immunisation against diphtheria, tetanus, pertussis, poliomyelitis and measles [186].

The third approach to control the epidemic is screening and quarantine. This is an interventional approach as it is exercised when an epidemic has already started. It is a core public health approach as it can reduce and delay the spread of the disease somewhat at the earliest stage. During the epidemics, the susceptible individuals are screened, and the ones who are thought to pose a risk will be quarantined. “Many countries do not attempt these measures because of logistics, and cost-benefit considerations” [183, 181, 185].

Due to the strong similarity between the propagation of diseases and the information dissemination from one hand, and a comprehensive framework in disease eradication, this study proposes to adopt the same framework for the rumour confrontation. To this end, the past counter-strategies are set into the epidemic framework in order to understand, which phase of the epidemic control is less emphasised in rumour confrontation.

4.3.2. STRATEGIES EFFECTIVENESS

This section tends to assess the counter rumour strategies that are introduced in the previous section using the epidemic framework. As it is mentioned earlier, the epidemic framework presents three approaches to control the spread of disease: (i) exposure minimisation, (ii) immunisation or vaccination, and (iii) reducing the transmission rate. In this section, the same framework is used to assess the goal of counter rumour strategies. As Table 4.1 displays, rumour counter strategies pursue at least one of the epidemic control approaches.

The legislation mitigates the transmission rate and prevents further spread of rumours. It takes punitive measures against those who participate in rumour dissemination. This would make people more careful and cautious about sharing informa-

tion with their peers. In the political solutions, the goal is to cut the rumour from the source through negotiation by the main sponsors behind the rumours, so it belongs to the strategies with exposure minimisation view. The AI techniques for the filtering and downgrading tend to reduce the visibility of rumours in the online social network by taking down rumour related contents and accounts; thus they are considered as strategies with exposure minimisation approach. If they are used to just flag the misleading contents, then they also fall into the third category (reducing the transmission rate) as well. The collaboration between online social media platforms and fact-checking institutes provides information about the truthfulness of the posts circulating in the platforms. Fact-checking the posts may dissuade people from sharing the news with their network. The collaboration can also improve the accuracy of filtering and downgrading algorithms. The other type of platforms' collaboration is with a media organisation. It would empower professional journalism in the digital era and reduce the likelihood of rumour emergence in the news channels. The collaboration between platforms and the scientific community would contribute to the improvement of filtering and downgrading techniques which would eventually minimise the rumour exposure. Within the design-based strategies, the service retirement and the service modifications aim to reduce the likelihood of the rumour appearance and transmission. The new service development also can serve the same function; however, among the services that have already been developed, they are mostly for reducing the rumour transmission.

The purpose of ad-hoc fact-checking is to prevent mistakes and false information before making public; therefore, it falls into the exposure minimisation category. Conversely, post-hoc fact-checking corrects false information after they go public and decreases the likelihood of rumour transmission. The rules of thumb to deal with targeted rumours are basically a set of principles that rumour audience could use to reduce the likelihood of rumour transmission. The rumour clinics and public inoculation both try to create immunisation by teaching people how rumours deceive the mind. The educational approach tends to raise awareness and educate people. It works like the mind vaccination as it aims to train the brain not to be trapped by rumours. Control centres reduce the rumour transmission rate by filling the news channels gap by providing information about the floating rumours.

The analysis shows that the existing strategies against rumours unevenly cover all aspects of the epidemic framework. Although the vaccination is recognised as the most effective approach in the control of the epidemics, there is more emphasis on the counter-rumour strategies with other approaches, namely, exposure minimisation and reducing transmission rate.

4.4. CONCLUSION

This chapter investigated the counter rumour strategies. Based on their impact on different components of a communication process, they are classified into three groups of sender-, channel-, and receiver-related strategies. After introducing the strategies, they are assessed based on the epidemic framework. In this framework, strategies and measures are evaluated based on three criteria of exposure dissemination, giving complete protection, and reducing the transmission rate.

The introduced counter-measures are exercised intermittently over the past century.

Table 4.1: Analysis of the quelling strategies against epidemic control framework.

		Exposure minimisation	Giving complete protection (vaccination)	Reducing the transmission rate
Rumour counter strategies	Legislation			✓
	Political solution	✓		
	Filtering	✓		✓
	Downgrading	✓		✓
	Collaboration with fact-checking	✓		✓
	Collaboration for media literacy and digital journalism	✓		
	Collaboration in scientific project	✓		
	Service retirement	✓		✓
	Service modification	✓		✓
	New service development			✓
	Ad-hoc fact-checking	✓		
	Post-hoc fact-checking			✓
	Rule of thumbs to deal with targeted rumours			✓
	Rumour clinics		✓	
	Public inoculation		✓	
	Educational policy		✓	
	Control centres			✓

Despite tremendous efforts and developing all those strategies, diffusion of rumours not only has not shrunk but also escalated. The ephemeral reactions to sudden rise of rumours in different periods as well as the absence of a comprehensive plan are amongst the reasons for the failure of curbing the rumours. Besides, focusing on the rumour exposure minimisation and reducing the rumour transmission rate while neglecting the more effective approach of immunisation is another reason for the failure of the current set of strategies against rumour dissemination.

To tackle this problem, what is essentially required is a comprehensive plan that incorporates all aspects of the epidemic framework. It needs to especially focus on (i) immunisation approach due to its proven effectiveness and (ii) AI-based techniques due to the scale, scope, and speed of rumour spreading in online social media platforms. The following chapters focus on these two points.

5

AN ASSESSMENT OF ACADEMIC EFFORTS REGARDING RUMOUR CONFRONTATION

*The source of a fountain may be stopped with a bodkin
But, when it is full, it cannot be crossed on an elephant*

Sa'di

The previous chapter discussed the significance of countering rumours due to their potential to inflict damages at different levels. It was also pointed out that preventive measures are essential elements in a comprehensive plan to curb and control rumour spreading effectively. In the same vein, this chapter poses the following research question:

- *How ready is the academia regarding the spread of rumours?*

One of the key parties in the development of preventive strategies is academia. Despite a great deal of research in this arena, the amount of progress by academia is not clear yet. This may lead to misjudgements about the performance of this field of research, which can ultimately result in wrong science policies regarding academic efforts for quelling rumours. In this research, we address this issue by assessing the academic readiness in the topic of rumour spreading. To this end, we adopt the emergence framework and measure its dimensions (novelty, growth, coherence, and impact) over more than 21000 articles, published by academia about the rumour and its conceptual relatives. The results show the current body of research had an organic growth so far, which is not promising enough for confronting the large-scale problem of rumour diffusion. To tackle this problem, we suggest an external push strategy which reinforces the emergence dimensions and leads to a higher level in every dimension¹.

¹This chapter is based on the following publication: Fard, A. E., & Cunningham, S. (2019). Assessing the Readiness of Academia to Confront Rumours: A Systematic Literature Review.

5.1. INTRODUCTION

As it is discussed in Chapter 4, creating immunity among individuals to not being trapped by rumours would reduce the circulation of rumours drastically. The direct impact of this approach is to protect individuals' mind and to make them rumour-proof. When people are not carriers of a disease, automatically they will not be able to infect others. This is the indirect impact of vaccination approach, which automatically reduce the rumour transmission rate.

One of the most effective ways of creating immunity against rumours is through education [187]. The ultimate goal of different educational approaches is to create vigilance and awareness as well as training the mind to not being deceived and manipulated by rumours. In this vein, the role of academia in developing effective strategies is crucial. Since 1940, academia is in the front-line of confronting rumours.

Despite all the efforts in academia, we still do not know how much progress has been made in aggregate, and what is the readiness of academia. This can be problematic because when we do not know our readiness in a particular subject, we either overestimate or underestimate our ability in that subject. Both of these misjudgements are incorrect and lead to decisions irrelevant to the existing circumstance. So, in order to get an accurate picture of academia's readiness in the topic of rumour propagation, it is essential to evaluate this topic of interest. To this end, we deploy technology emergence framework and measure the academic readiness in the topic of rumour circulation. In this framework, we study four dimensions of emergence (novelty, growth, coherence and impact) over more than 21,000 scientific articles, to see the level of readiness in each dimension. This helps us to provide accurate recommendations for the improvement of those dimensions with a low level of readiness.

This chapter is organised as follows. Section 5.2 explains the theoretical framework of this study. Section 5.3 explains the methodology, including the data collection and operationalisation of emergence framework. In Section 5.4, we report the results of our analysis over each dimension of emergence framework (novelty, growth, coherence, and impact). In Section 5.5 we discuss the results, and finally Section 6.6 summarises and concludes this chapter.

5.2. SCIENTIFIC EMERGENCE

Emergence is a broad concept pointing to the process of coming into being, or of becoming important and prominent [188]. In the scientific context, emergence mostly refers to the emergence of new technology, scientific field or a topic of interest. There are competing theories that explain the emergence of scientific disciplines. For example, Thomas Kuhn [189] justifies emergence of scientific fields by his paradigm shift theory, or by Donald Stokes theory [190], the emergence of science is user-oriented and is triggered by societal needs.

Measuring the emergence has drawn much attention in recent years, owing to notable achievements in the operationalising of this concept. One of the primary works

ness of Academia in the Topic of False and Unverified Information. *ACM Journal of Data and Information Quality (JDIQ)*, 11(4), 1-27.

To access the codes used for the analysis of the bibliometric data, please refer to <https://gitlab.com/amiref/acm-jdiq-paper>

Table 5.1: Technology emergence dimensions [188]

Attributes	Definition
Novelty	Fulfilling a given function by using a different basic principle as compared to what was used before to achieve a similar purpose
Growth	Increase over the number of involved actors, public and private funding, produced knowledge, prototypes, products and services
Coherence	Reaching to a certain identity and momentum from those technologies/topic of interest still in a state of flux
Impact	Pervasiveness of the impact that emerging technologies may exert by crosscutting multiple levels of the socio-economics system
Uncertainty and ambiguity	Uncertainty refers to the possible outcomes and uses which might be unintended and undesirable. Also, ambiguity refers to the possibility that different groups associated with a given technology

in this domain is the FUSE² program, which suggested four dimensions for measuring emergent technologies: (i) novelty, (ii) persistence, (iii) community, and (iv) growth [191, 192]. “What is an emergent technology?” is the title of seminal work by Rotolo et al. [188]. In this work, they propose five dimensions as the denominator of every emergent technology (novelty, growth, coherence, impact, uncertainty and ambiguity). Table 5.1 introduces and explains the dimensions. Although the framework is developed to identify emerging technologies, it is also applied to other emergent phenomenon such as emergent topics of interest [188, 193].

In addition to the conceptualisation of the notion of emergence, one of the unique aspects of Rotolo’s work is identifying the dimensions of emergent technologies and demonstrating the dynamic of each dimension [188]. As Figure 5.1 displays, there are two dynamics for the components of the emergence framework: negative (decreasing) s-curve for radical novelty and uncertainty & ambiguity, and positive (increasing) s-curve for relatively fast growth, coherence and impact. Based on the level of each dimension, three phases are suggested: pre-emergence, emergence and post-emergence. Performance of technology in all dimensions shows the overall status of that technology in emergence framework. Nonetheless, trying to pin down the absolute values for emergence dimensions is rather meaningless. In fact, the dimensions of this framework, provide an indication of emergence when they are considered in the domain in which the given technology is arising. Therefore, an emergent technology may only be compared with the other technologies if they are also in the same domain [188]. Despite the limitations, this unique feature of Rotolo’s framework allows us to do a retrospective analysis and compare a topic of interest with itself through the time.

5.3. METHOD

This section comprises of two major parts: (i) data collection, and (ii) operationalisation. First, we need to collect empirical evidence which indicates the scientific efforts and achievements in this topic of interest so far. In the second part, we must operationalise dimensions of the theoretical framework we adopted in Section 5.2 to be able to measure the emergence dimensions. Finally, we measure the level of readiness by applying the chosen criteria to the collected data. Figure 5.2 summarises the steps that need to be

²“Foresight and Understanding from Scientific Exposition” (FUSE) research program funded by the US Intelligence Advanced Research Projects Activities (IARPA) in 2011

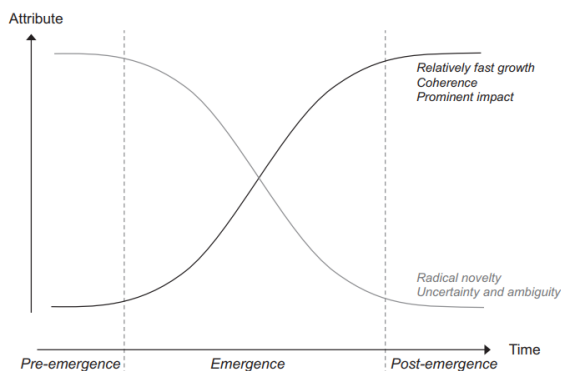


Figure 5.1: Different phases and attributes of emergence [188].

taken for the assessment. The dark part of the diagram indicates data collection steps, and the light part shows operationalisation procedure. Analysing the obtained metrics from operationalisation with collected data, gives us the results of the assessment. In the Section 5.3.1, and 5.3.2 we elaborate data collection and operationalisations procedures respectively.

5.3.1. DATA COLLECTION

To improve the reliability of our results, we need to apply the emergence framework on high-quality data. To this end, we follow three steps: (i) developing a comprehensive search strategy via finding relevant descriptive terms, (ii) choosing a reliable bibliometrics database and, (iii) defining a query, refining it and collecting the data. In this section, we elaborate every step of this procedure in details.

SEARCH STRATEGY

As the first step, we explain the search strategy. In this topic of interest, there is no consensus on the terms which are representing rumours. Besides, having more descriptive terms, allow us to perceive the phenomenon from different perspectives which would give us a more comprehensive picture of the topic of interest and would make the results more reliable. Therefore, we try to find as many relevant describing terms as possible.

To this end, a concept crawler for Wikipedia³ is developed to extract relevant terms to the field of rumour studies. Algorithm 1 shows how this crawler works. It is initialised with a list of seed terms and a depth argument. It goes to the Wikipedia page of each term, extracts the recommended concepts and adds them to the list of concepts. If an extracted term is already in the list, the crawler ignores it and just increases the frequency of that concept in the list. Because of the enormous size of Wikipedia's network, and to keep the collected concepts relevant to the scope of this study, a depth argument (d) is given to the crawler. It shows after how many hops the crawler stops.

³Our initial plan was to implement this crawler for Encyclopaedia Britannica, but Wikipedia covers more concepts and topics than Britannica. Hence the chance of finding more relevant terms in Wikipedia is higher than Britannica

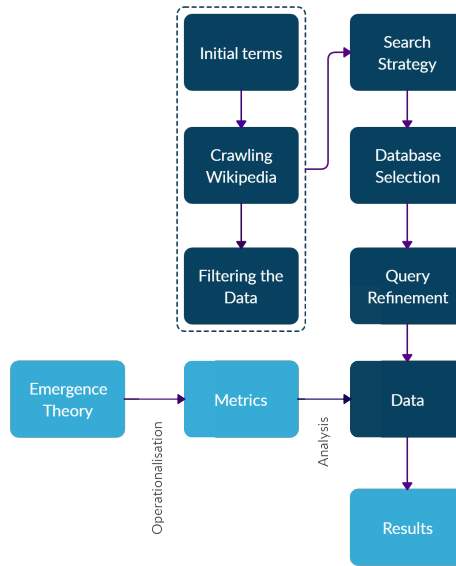


Figure 5.2: The method of assessing the readiness of the academia in the field of rumour studies.

Algorithm 1 Concept Crawler

Input: List (C) of initial concepts C_i , $m = 1, 2, \dots, M$ and depth of crawl d

while $depth < d$ **do**

$c \leftarrow C.dequeue()$

$new_concepts = fetch_recommended_concepts(c)$

$C.enqueue(new_concepts)$

$depth = depth + 1$

end while

Output Final list of concepts and the frequency of recommendation

We initialise the Wikipedia concept crawler with a list of 12 variations of rumours which are highly popular in the literature. We set the depth argument to 500 ($k = 500$)⁴ and then run the algorithm. As the network of the results is displayed in Figure 5.3 (each node shows a concept (which has a Wikipedia page), and each edge from node A to node B means concept B is recommended in the Wikipedia page of concept A), 3299 concepts are recommended by Wikipedia, but many of those are irrelevant and have to be eliminated. Thus, we conduct three-step filtering. First, based on node in-degree value, then based on page-rank centrality value, and finally, we manually filtered irrelevant concepts. After the first and second level of filtering, the number of concepts are reduced to 684 (almost 20% of the original set of concepts) and 102 (less than 5% of the original set of concepts) respectively. In the last step of filtering, after careful consideration, we

⁴it means after 500 hops, the script stops

reached to 19 terms in addition to nine initial terms, which altogether provides us with a list of 28 terms as the final list of search terms⁵.

BIBLIOMETRIC ENGINES

There are three major databases for indexing bibliometric data: Web of Science (WoS), Scopus and Google Scholar. None of those databases has absolute superiority over the others, and each one has its strengths and weaknesses (Table 5.2); however, we choose Web of Science for data collection because it includes high impact journals and has a long period of coverage. Here, we use the full cleaned dataset of WoS English publication records, including articles, reviews, editorials, proceedings, conference papers, and book chapters.

Table 5.2: Comparison between three major databases of indexing bibliometrics data [194].

Features	Scopus	Web of Knowledge	Google Scholar
Number of Journals	19809	12331	Unknown
Export records	Yes - <i>en masse</i>	Yes - <i>en masse</i>	Yes - <i>en masse</i> if you mark records which saves to My Library - then export from within My Library
Period covered	1966 -	1900 -	Unknown
Update	Daily	Weekly	Unknown

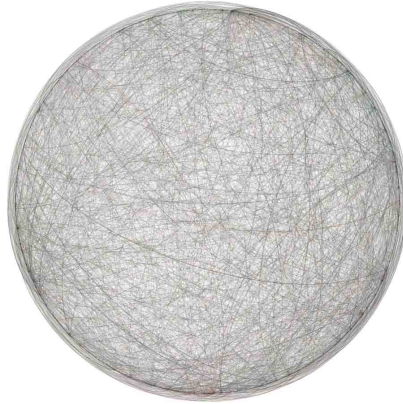
QUERY DEVELOPMENT AND REFINEMENT

Table A.1 displays the different queries we used to collect data⁶. We reached to our final dataset after four rounds of query refinement. In the first query, which was entirely based on the results of the concept crawlers, we observed many irrelevant papers about computer networks. They appeared in the dataset because of the “gossip” term in the search query, which returns a lot of papers about gossip protocols. Such papers are irrelevant to this study and must be eliminated. Hence, in the second query, we exclude them by adding the negation of those terms to the search query. After getting the second version of the dataset, we noticed a surprisingly high number of papers from oncology. After checking some of those papers, we discovered Web of Science OCR⁷ cannot distinguish “rumour” and “tumour”. As a result, all the papers about tumour which were indexed in WoS appeared in the dataset. For the third query, we added the negation of “tumour” to our search query to exclude the irrelevant oncology papers. This time the returned results were satisfactory; however, we performed manual filtering and removed a few of unrelated papers. Finally, we reached 21571 papers. We also evaluated the approximate level of the noise after data collection and refinement. For that purpose, we randomly selected 100 papers before and after data refinement, and observed the noise level has dropped from 19% to 8%.

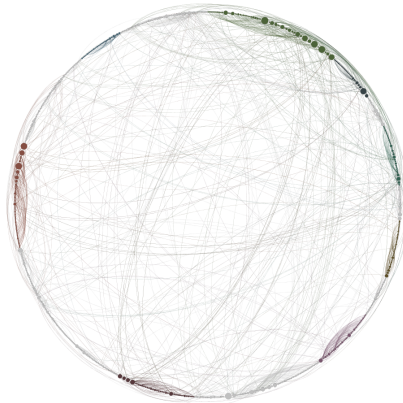
⁵Please refer to the Section A1 in the appendix for the final list of search terms

⁶For the detailed queries, please refer to the Appendix

⁷Optical character recognition



(a) The network of Wikipedia's recommended terms for initial set of concepts



(b) The network of Wikipedia's recommended terms after first level of filtering



(c) The network of Wikipedia's recommended terms after second level of filtering

Figure 5.3: Data collection and data filtering steps

Table 5.3: Queries for data collection from Web of Science

	Query	Returned results	Description
1	Q1	25769	Some of the returned manuscripts are about computer networks which is irrelevant to the topic of false and unverified information
2	Q2	24021	Surprisingly, many of the returned manuscripts are in the field of oncology which seems a little bit strange. After careful examination, we discover Web of Science OCR algorithm cannot distinguish "rumour" from "tumour" and significant part of the returned papers are about cancer
3	Q3	21675	Although, a considerable noise is removed from the dataset, there are still some irrelevant papers. For example there is a person called "Rumer", but Web of Science OCR algorithm recognizes it as "rumor", or there is Russian physicist called "Rumour" which is apparently confused by the concept of "rumour", or there is a geological site in Australia called "Rumour" which is confused by the rumour concept again
4	Q4	21571	This is the final version of the dataset on false and unverified information

5.3.2. EMERGENCE OPERATIONALIZATION

The operationalisation plays a central role in this research and works as a bridge between the theory and data. Without operationalisation, our research would freeze at a theoretical level. More precisely, operationalisation enables us to quantify emergence dimensions, which provides us with an opportunity for the comparison. This comparison can be between the different topic of interest or just a single topic of interest over the period. In summary, operationalisation makes our empirical approach meaningful and allows us to interpret our empirical evidence based on the theoretical framework.

Here, we measure all dimensions of the emergence framework [188] except the last one, no metric can genuinely operationalise uncertainty and ambiguity [188]. For each selected dimension, one or more indicator have been suggested which are either extracted from literature or developed by the authors. The indicators corresponding to each dimension are explained in Table 5.4.

5.4. RESULTS

In this section, we report results of measuring different dimensions of emergence namely, level of novelty, growth, coherence and impact in the topic of rumour spreading using the criteria introduced in the previous section.

5.4.1. NOVELTY

To measure novelty, several approaches such as co-word analysis [195] and citation analysis [196] have been proposed, but none of them has a meaningful interpretation in this topic of interest. That is the reason we develop a new method to measure novelty. At the heart of our method, topic modelling lied. Topic modelling is an exploratory statistical method to extract hidden thematic structure from a collection of documents. In this method, the algorithm receives two inputs: (i) a set of text documents, and (ii) the number of topics (p). Then it returns clusters of words where each cluster represents a topic. In this study, we adopt Latent Dirichlet Allocation (LDA) as our method for topic modelling [197]. To implement our method and measure novelty in the field of rumour studies, the following procedure needs to be followed:

In the following, the steps are elaborated. First, we perform topic modelling with

Table 5.4: Operationalisation of emergence framework. The new criteria are marked with †.

Dimensions	Operationalization criteria	Description
Novelty	<ul style="list-style-type: none"> • Maximum novelty† • Average novelty† 	We measure the similarity between topics in every two successive periods, then we investigate how it changes through the time
Growth	<ul style="list-style-type: none"> • Trend analysis over the number of publications • Trend analysis over the incumbents and newcomers authors† • Newcomers contributions† 	We study the growth in publications and authors
Coherence	<ul style="list-style-type: none"> • Dynamic of co-occurrence of major disciplines over underpinning discipline network† • Conferences and journals • Evolution of densification over co-authorship network 	We measure theme coherence (the first criterion) and community coherence (the second and third criteria).
Impact	<ul style="list-style-type: none"> • Trend analysis over the number of underpinning disciplines • Histogram over the discipline score • Trend analysis over number of fundings acknowledgements and variety of funding agencies • Histogram over the number of fundings acknowledgements 	We study funding agencies and underpinning disciplines to measure to what degree this topic of interest raise expectation

Procedure of measuring novelty

1. Applying LDA on four periods of data (Due to the lack of considerable academic activities in the field of rumour studies in the first 80 years of 20th century, we take that period (from 1900 until 1980) as the first one and call it the first period, then for every decade after 1980 we take it as an extra period)
2. Measuring the similarity between topics in different periods and filling the comparison table
3. Measuring degree of novelty (with *AverageNovelty* and *MaximumNovelty*) for each topic in every period
4. Getting the average of *AverageNovelty* and *MaximumNovelty* for all the topics in each period to have a big picture of change in the novelty level in this topic of interest

five topics. In topic modelling, making a decision about the number of topics is not a straightforward task. It is a matter of experience, and there is no gold standard for that. Here, after multiple times topic modelling with different topic numbers, we reach to this number. If $p > 5$, the excess topics are extremely similar to the others. And, if $p < 5$, some of the important topics that we see them separately before (when $p = 5$) are merged to other topics. After applying LDA on four periods of the dataset, we get five clusters of words for each period.

In the second step for every two topics in different years, we measure their similarity using cosine similarity measure and insert it into the comparison table⁸. The Figure 5.4 demonstrates a schema of the comparison table. Every cell in the comparison table illustrates the similarity between the two topics in two years. For instance, in Figure 5.4, the yellow cell indicates the similarity between topic 0 from fifth period (2010 ~ 2018) and topic 1 from fourth period (2000 ~ 2010).

⁸We take each cluster of words as a weighted vector

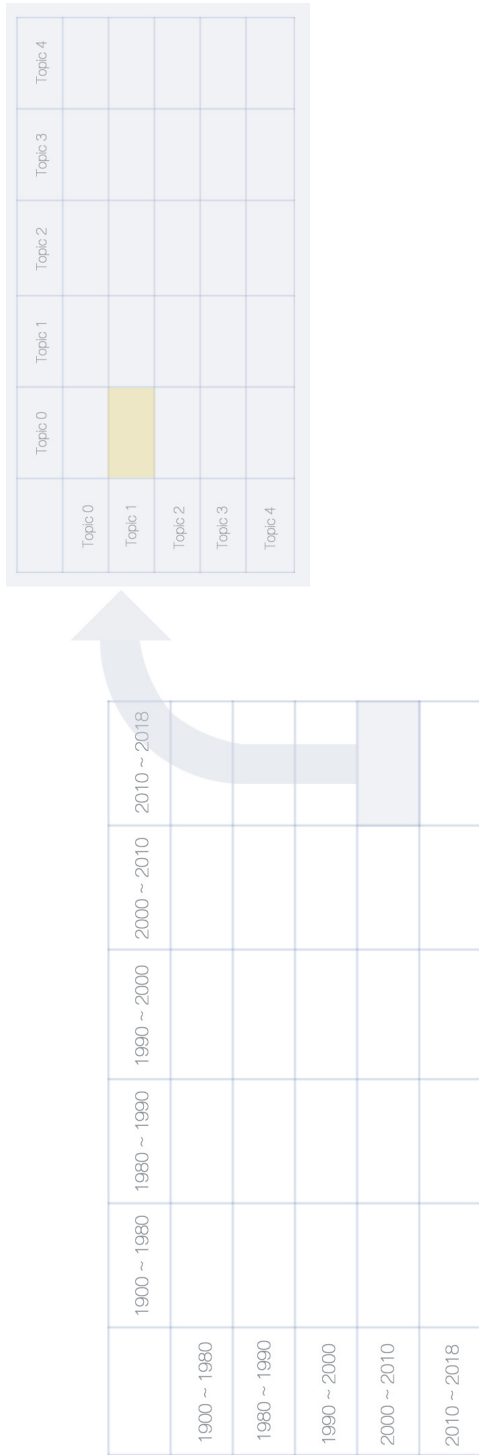


Figure 5.4: Schematic view of comparison table

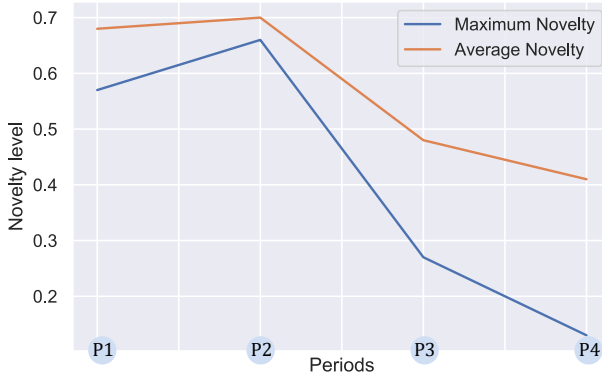


Figure 5.5: Change in novelty level in the field of rumour studies

5

After measuring the similarities and filling out the table, in the third step, we measure the degree of novelty for each topic in every period. Degree of novelty, measures the similarity between topics in every two consecutive years (Equation 5.1 and 5.2). It has two variations; one works based on maximum and the other based on average function. In the average approach, we get the average of similarity between topic i in period j with all the topics in period $j-1$. In the maximum approach, after getting the similarities, instead of getting the average, we pick the maximum value. It is worth mentioning; we subtract the result of average or maximum from one because the similarity between the topics, shows to what degree, same topics repeat in different years, while in this method the goal is to know, how much of the topics are new. In other words, without subtraction, the measurements show to what degree the topics are similar to each other.

$$MN_t^y = 1 - \max_{t' \in \text{topics}} (S(\text{topic}_t^y, \text{topic}_{t'}^{y-1})), \quad (5.1)$$

where MN is the maximum novelty and S is the similarity.

$$AN_t^y = 1 - \text{average}_{t' \in \text{topics}} (S(\text{topic}_t^y, \text{topic}_{t'}^{y-1})), \quad (5.2)$$

where AN is the average novelty and S is the similarity.

Finally, in the fourth step, we take the average of two key variables of *AverageNovelty* and *MaximumNovelty* for all the topics in each period. This gives us a big picture of change in the novelty level in different years (Figure 5.5). Both *AverageNovelty* and *MaximumNovelty* show the same pattern: first a modest increase from the first to the second period and then a significant drop to the third and then fourth period. This pattern demonstrates that the novelty is declining, and the research topics are becoming more and more similar to each other. However, we can also interpret the drop in the novelty as a sign of coherence in the field.

5.4.2. GROWTH

For growth dimension, we perform three levels of analysis which is summarised in the following:

Procedure of measuring growth
<ol style="list-style-type: none"> 1. Counting the number of publications per year 2. Counting the yearly publications of newcomer and incumbent authors 3. Measuring <i>NewcomersContribution</i> per year

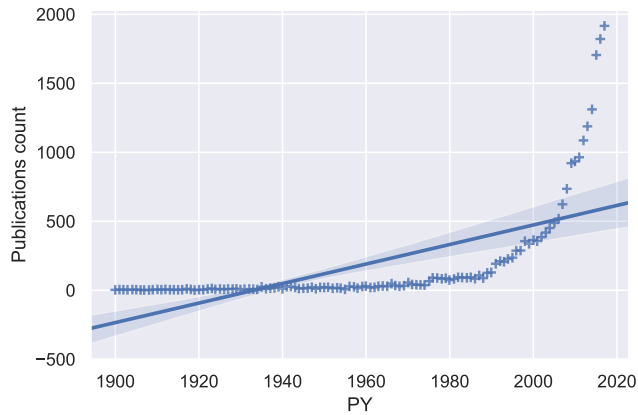
We start from trend analysis over the number of publications⁹. As the Figure 5.6a shows, for almost 80 years, this topic of interest was like a silent volcano which occasionally launched ash and smoke. But since almost 30 years ago, it has experienced dramatic growth in the number of publications; from almost 80 publications in the 80s to almost 2000 publications in 2017. But such a sudden dramatic growth is not necessarily a good sign. In fact, it may be just a hype, not a sign for the emergence of a new topic of interest with an independent identity.

In the second step, to explore this issue, and get a better picture of the growth in the field of rumour studies, we want to see how much of this growth is obtained by newcomers and incumbents authors. Newcomers are those authors who do not have any publication in this topic, while incumbents are those who already have at least one article in this area. As the Figure 5.6b illustrates, 87.7% of the publications in this area is produced by newcomer authors, and only 12.3% of the publications are produced by incumbent authors. In other words, among all the authors in this topic of interest, the fraction of newcomers is almost seven times more than incumbents. This provides us with a static picture of authors contribution in the field of rumour studies. Additionally, we study the dynamic of authors contribution by focusing on the speed of publication production for the newcomers and incumbents.

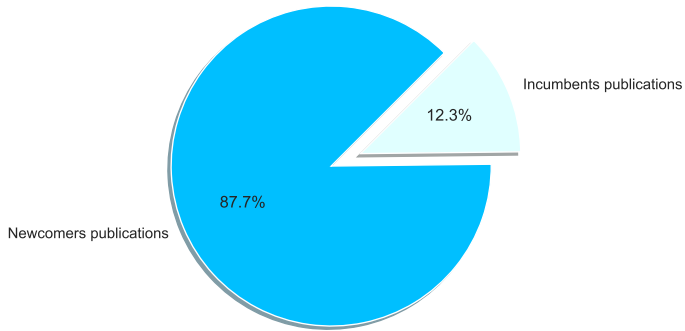
Figure 5.7a and 5.7b show the growth of newcomers and incumbents respectively. In both diagrams, the number of authors is growing but at a very different scale. We juxtapose both diagrams in Figure 5.7c to demonstrate how the number of incumbents and newcomers grow compared to each other. Due to this diagram, the rate of growth in newcomer authors compared to incumbent authors is significant. Newcomers grow exponentially while incumbents grow like a straight line with a small slope.

Finally, in the third step, to understand the difference between growth rate better, we define a new criterion called newcomers contribution and calculate the yearly contribution of newcomers to the field of rumour studies (Equation 5.3). We measure the level of contribution for each year by dividing the number of incumbent authors in that year by a cumulative number of newcomer authors minus a cumulative number of incumbents authors from the beginning until that year. The numerator demonstrates the number of authors who stayed in this topic of interest, and the denominator shows those authors

⁹This indicator is adopted from [188].

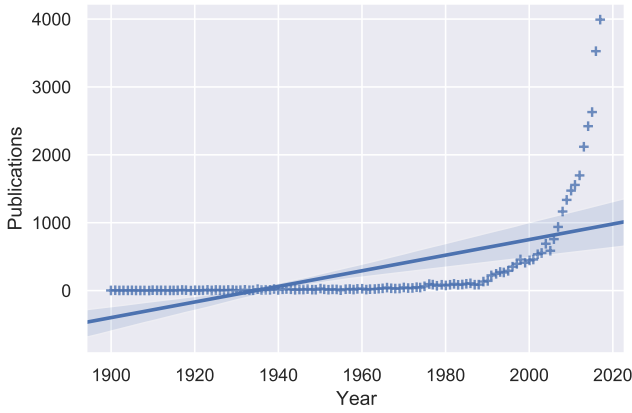


(a) Publication growth in the field of rumour studies

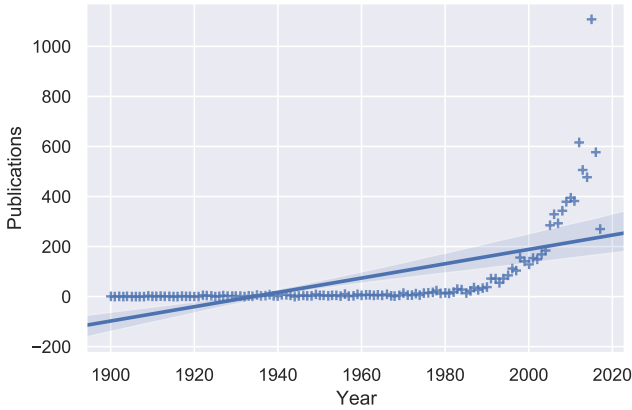


(b) The percentage of incumbents and newcomer authors in the field of rumour studies

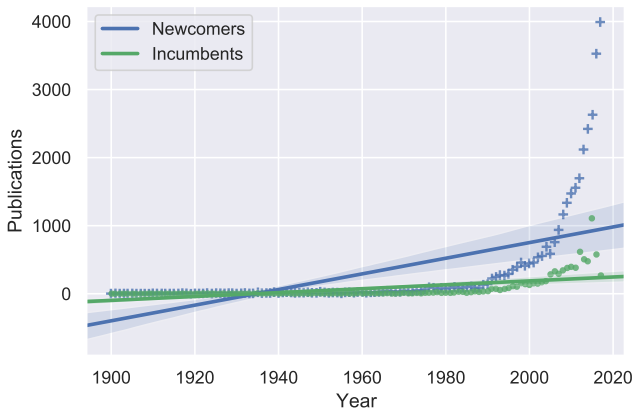
Figure 5.6: Growth and its composition in the field of rumour studies



(a) The growth of newcomer authors in the field of rumour studies



(b) The growth of incumbent authors in the field of rumour studies



(c) The comparison of growth between incumbent and newcomer authors

Figure 5.7: The composition of different communities in the underpinning disciplines network within 1900 to 2018

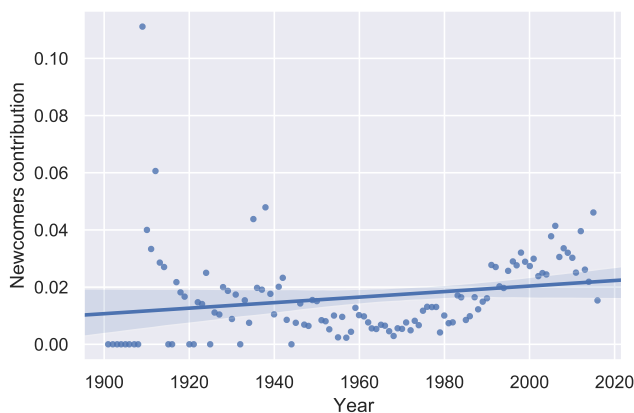


Figure 5.8: The yearly contribution of newcomers to the field of rumour studies

5

who could remain in this topic of interest, but they did not. In other words, the numerator shows the number of incumbent authors in year t and the denominator indicates the number of remaining newcomers from the beginning (year 1900) till a year ago (year $t - 1$).

$$NC_t = \frac{I_t}{\sum_{1900}^{t-1} N_t - \sum_{1900}^{t-1} I_t}, \quad (5.3)$$

where NC is newcomers contribution, I_t is the number of incumbents in year t , and N_t is the number of newcomers in year t .

The newcomers contribution shows the rate of newcomer to incumbent conversion. In other words, the numerator shows the number of incumbent authors in year t and the denominator indicates the number of remaining newcomers from the beginning (year 1900) till a year ago (year $t-1$). Dividing these two numbers, shows what rate of newcomers converted to incumbent this year.

This criterion shows, every year, what percentage of those who are still newcomers, join incumbents. As it is clear from Figure 5.8, except a few outliers, in general, less than 4% of newcomers join to the incumbents every year which is a very small fraction. This means most of the scholars who already have a publication do not come back to this topic.

All in all, by juxtaposing the results we can assert that the growth in this topic of interest is radical and imbalanced. Radical, because the number of publications started to grow exponentially since last 30 years, and imbalanced because the majority of publications are produced by newcomer authors, and incumbent authors do not show sustainable interest to publish more. The reason for the sudden rise in the number of publications by the newcomers might be the alarming trend of rumour circulation and its severe consequences in the past few years. Many of the scholars from a wide range of disciplines decided to respond this potential danger by their knowledge and expertise.

Besides, the reason for not publishing anymore by those who published once might be studying or conducting research in the fields other than rumour studies.

5.4.3. COHERENCE

In the third dimension of emergence framework, we study coherence in the field of rumour studies. We assess the coherence from two perspectives: *research theme* and *research community*. The criteria we use here, are the mixture of criteria extracted from [188] and the ones developed by the authors. In the following, the procedure of measuring coherence is explained:

Procedure of measuring coherence

- | |
|--|
| <ul style="list-style-type: none"> • Research theme coherence <ol style="list-style-type: none"> 1. Building the temporal network of underpinning disciplines for five periods 2. Applying community detection algorithm on the networks 3. Classifying the subject categories to GIPP research areas 4. Making the co-occurrence matrix of disciplines • Research community coherence <ol style="list-style-type: none"> 1. Counting the number of relevant dedicated journals 2. Counting the number of relevant dedicated conferences 3. Counting the number of relevant special issues 4. Counting the number of relevant conference sessions and tracks 5. Counting the number of relevant seminars and forums |
|--|

In the next two subsections, we elaborate each of these steps.

RESEARCH THEME COHERENCE

First of all, we need to build a network of underpinning disciplines for five different periods. This network is based on subject category labels given to each article and book in Web of Science database. We choose such a network structure to represent the thematic structure. In this network, each node represents one of the Web of Science subject categories, and every edge between two nodes shows there is a paper in both subject categories¹⁰. Because the Web of Science subject categorisation shows the theme of the paper¹¹, we can infer that this network represents the theme of the research activities in the field of rumour studies.

¹⁰The weight of the edge shows the number of papers between both subject categories

¹¹Subject categorisation in Web of Science is based on the theme of the journals

In the second step, to know which subject categories come together more often, we take the community detection approach. There are quite a few community detection algorithms [198], and we choose Lovain algorithm [199]. The main reasons for this choice are the speed, scalability, and simplicity of this method. It is one of the first scalable community detection techniques which has been successfully tested on networks of different types, and for sizes up to 100 million nodes and billions of links. Additionally, it is among the fastest community detection algorithms which works well with large graphs. The Louvain algorithm allows zooming within communities to discover sub-communities, and sub-sub-communities [200, 201]. It is also one of the most popular community detection algorithms and has been implemented in many of the network analysis tools and programming packages such as Gephi, or NetworkX.

Applying this algorithm on the underpinning discipline networks, specifies communities of the closest¹² subject categories in each period. But, because there are too many subject categories in the underpinning networks, it would be hard to make any comparison or draw any conclusion.

In the third step, we tackle this issue by grouping the Web of Science subject categories into six major research areas suggested by GIPP schema. In other words, we replace each subject category introduced in Web of Science with its parent discipline in GIPP schema. GIPP is a very broad categorisation comprising six disciplines (Arts & Humanities, Clinical, Pre-Clinical & Health, Engineering & Technology, Life Sciences, Physical Sciences, and Social Sciences) which covers all fields of scholarly research. GIPP is initially developed as a part of the Thomson Reuters Institutional Profiles project, and also is used in the Times Higher Education World University Rankings.

By applying GIPP classification on the identified communities, every subject category is replaced by its corresponding research area in GIPP schema. This turns communities with highly diverse composition into communities with a maximum of six disciplines. Such a small change makes the communities much easier to interpret and allows us to compare them within a single period and between different periods. Figure 5.9 illustrates communities we obtained in the third step and their compositions in each period. Every bar denotes one community and different colours in each bar represent contributing research areas in that community.

By comparing the five periods, we can see a significant drop in the number of communities from 50 in the first period to only 11 in the last period (Figure 5.9). It is mostly because of the singleton communities¹³ which their number decreases from 37 in the first period to 6 in the last one. Such a significant drop in the number of communities means new links appear in the network over time. That happens mostly because this topic of interest is getting more and more interdisciplinary. But, this does not necessarily indicate the lack of coherence around particular domains. In fact, in the rest of this section, we focus on this point to understand if there is any thematic research area that the field of rumour studies has shaped around it.

To this end, in the fourth step, we investigate the composition of the communities in different years. As Figure 5.9 shows, communities are the mixture of several disciplines. This can make the interpretation of coherence a bit hard. To tackle this issue, we make

¹²The closeness definition depends on the community detection algorithm.

¹³We define singleton communities as communities with only one member

a co-occurrence matrix. This matrix allows us to see the thematic composition of this topic of interest in an organised way in each period. Having such a matrix also enables us to compare different periods of this research topic.

The co-occurrence matrix is created in three steps. First of all, for each period, we pick the major communities. Major communities in each period are those communities which are larger than 30% of the biggest community in that period. After that, for each major community, we determine major research areas. We define major research areas in a major community as the largest research areas that altogether consist of at least 80% of that community. Finally, for each period, we count the number of major research areas in major communities and illustrate it in matrix (Figure 5.10).

In the co-occurrence matrix M , let D denotes the set of research areas and P be the set of time periods. In this matrix, the sample cell M_{ij} with the value of k ($M_{ij} = k$) shows that, in period i , there is k major communities with the major research area of j .

To measure the significance of each research area in the co-occurrence matrix, three criteria should be taken into account: (i) presence, (ii) continuity, and (iii) recency. The presence means, how many times a research area appears in major communities in different periods. The continuity explains to what extent this presence persists in consecutive years, and the recency indicates how recent is the appearance of research areas. We combine all these three criteria into one score called *Thematic Significance*(θ) (Equation 5.5). This score incorporates recency by associating periods with their index (Equation 5.4). Thus, the more recent a period becomes, the higher its associated index gets. Besides, this score involves presence and continuity by rolling windows with different sizes. The window with size one, counts in how many periods a research area appears¹⁴, and the bigger windows check whether those appearances are persistent or not. For instance, the window size 4 checks whether a research area appears for four consecutive periods or not.

$$\mathbb{I} = \{(i, j) \mid i, j \in P, i \leq j\}, \quad (5.4)$$

where $P \in \{1, 2, 3, 4, 5\}$.

$$\theta_d = \sum_{(i,j) \in \mathbb{I}} \left(\sum_{t=i}^j t \times \prod_{t=i}^j M_{rt} \right), \quad d \in D \quad (5.5)$$

where M is Co-occurrence matrix and D is Research areas existing in the first column of M .

As an example, to calculate *Thematic Significance* for research areas *Art & Humanities - Social Sciences* (the last column of Figure 5.10), we start moving the windows with different sizes. For window size 1, the score is (1+2+3+4+5), for the window size 2, the score is ((1+2)+(2+3)+(3+4)+(4+5)), for the window size 3, the score is ((1+2+3)+(2+3+4)+(3+4+5)), for the window size 4, the score is ((1+2+3+4)+(2+3+4+5)) and for the window size 5 the score is (1+2+3+4+5). If we add all the scores for different window sizes, thematic significance for *Art & Humanities - Social Sciences* would be 105.

After measuring *Theme Significance* for all the research areas, the results are shown in Figure 5.11. As the figure shows, two research areas have higher *Theme Significance*:

¹⁴The window with size one checks the criterion of *presence*

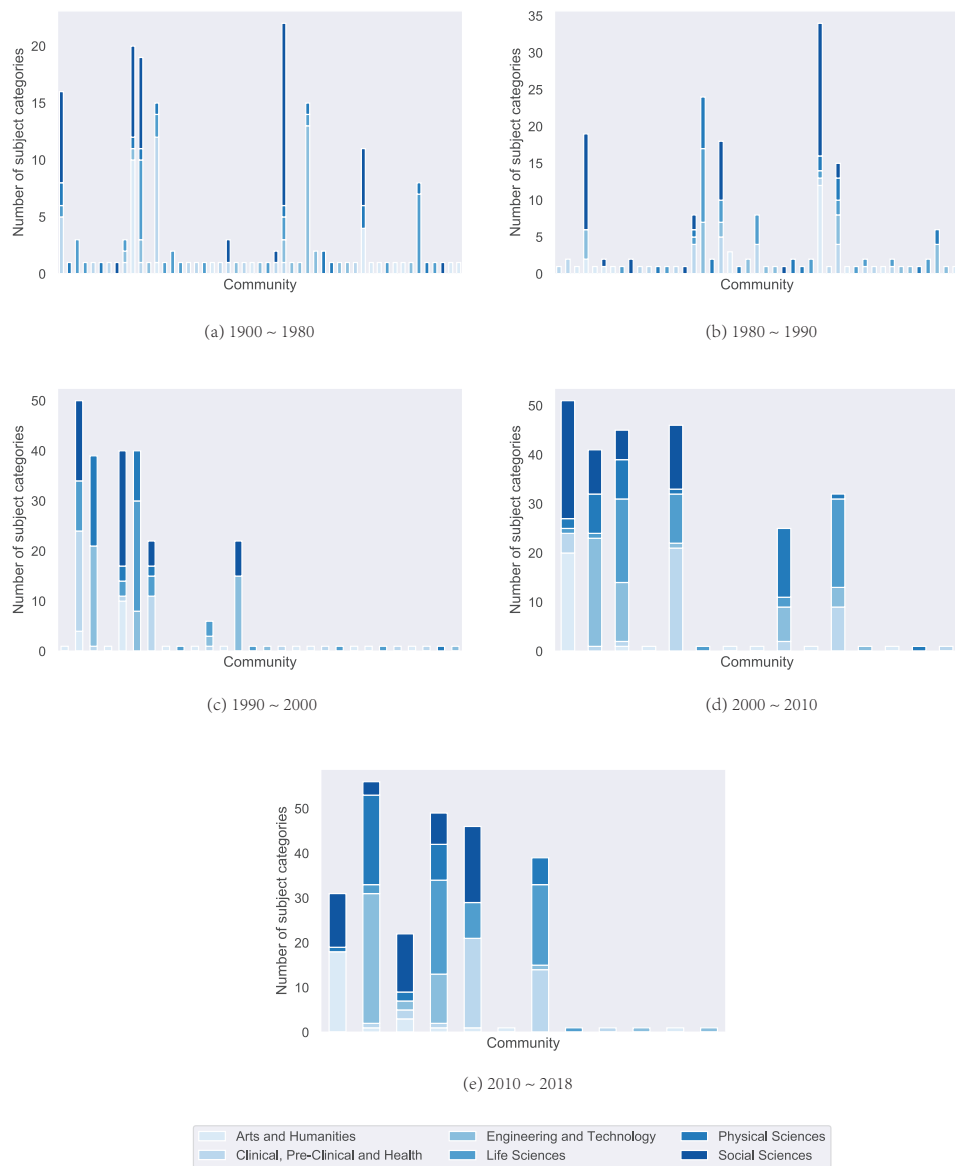


Figure 5.9: The composition of research communities in five periods of 1900 ~ 1980, 1980 ~ 1990, 1990 ~ 2000, 2000 ~ 2010, and 2010 ~ 2018. Every bar denotes one community and different colours in each bar represent contributing research areas in the corresponding community. The length of each bar displays the number of subject categories in its community.

(i) Art & humanities and social sciences and, (ii) Engineering and technology, and physical sciences. In the first group, both disciplines dominate one community and co-occur in all five periods, and in the second group, both groups co-occur and dominate one community in the last four periods.

By juxtaposing the results, we can infer that, the number of communities in the field of rumour studies is reducing significantly. The main reason for that is singleton communities which are joining the major ones. There are still a few singleton communities left, but compared to the early periods that the number of them is too low, even if they join the major communities, it is not a significant change in the structure of communities. The other finding is the relative stability and coherence in two research areas of (i) Art & humanities and social sciences, and (ii) Engineering and technology, and physical sciences. The coherence and stability of these two areas show that, since early days of research in the field of rumour studies, the scholars were interested in qualitative and quantitative approaches. In recent years, we can still observe affinity of the researchers to both perspectives; however, due to sudden increase in the scale, scope, and speed of rumour dissemination and democratisation & prevalence of computational techniques, the amount of research with computational approach is drawing more attention.

RESEARCH COMMUNITY COHERENCE

Another aspect of coherence is the community which we want to investigate it by looking at the conference sessions, special issues, specialised journals and conferences dedicated to the field of rumour studies [188]. For journals, we checked the list of journals covered by Web of Science and Scopus, but there is still no specialised journal on this topic of interest; however, due to the importance of this topic, some journals decided to publish special issues and explore different aspects of rumour based on their perspective. Table 5.5 indicates special issues on different aspects of this topic of interest. Currently, between the special issues, two themes are more prevalent: first, the study of fake-news and disinformation due to the importance of rumour spreading on political events and second, the study of gossip from a psychological perspective, most probably because of the important role of gossip in organisations and social communication.

Like journals, we have more or less the same situation for conferences as well (Table 5.6). Except for a conference on misinformation in eating disorders, there is no dedicated conference to the field of rumour studies. Still, due to the importance of the problem, some of the major conferences decided to hold a special track or workshop in conjunction with the main conference. There are also several forums and summits which their main purpose is to bring awareness about the problem.

Although zero dedicated journal or conference can be the sign of immaturity and lack of community in this topic of interest, the special issues, forums, workshops and conference tracks indicate the importance and urgency of this topic and interest of scholars from different disciplines to explore and study this area more deeply.

To investigate this point more accurately, we study the densification¹⁵ of co-authorship network to see how the collaboration per person is changed through the time. As the Figure 5.12 shows, the level of densification has increased gradually in co-authorship

¹⁵Densification in a network $G=(V,E)$, is defined as the fraction of edge number over node number ($Densification = \frac{|E|}{|V|}$)

2010 ~ 2018	2000 ~ 2010	1990 ~ 2000	1980 ~ 1990	1900 ~ 1980	
					Life Sciences
					Engineering and Technology
					Art and Humanities Clinical, Pre-Clinical and Health Social Sciences
					Clinical, Pre-Clinical and Health Engineering and Technology Life Sciences Physical Sciences
					Engineering and Technology Physical Sciences Social Sciences
					Life Sciences Physical Sciences
					Engineering and Technology Life Sciences
					Clinical, Pre-Clinical and Health Physical Sciences Social Sciences
					Engineering and Technology Life Sciences Social Sciences
					Clinical, Pre-Clinical and Health Life Sciences Social Sciences
					Clinical, Pre-Clinical and Health Social Sciences
					Engineering and Technology Social Sciences
					Engineering and Technology Life Sciences Physical Sciences
					Clinical, Pre-Clinical and Health Life Sciences
					Engineering and Technology Physical Sciences
					Art and Humanities Social Sciences

Figure 5.10: Co-occurrence of disciplines in the field of rumour studies

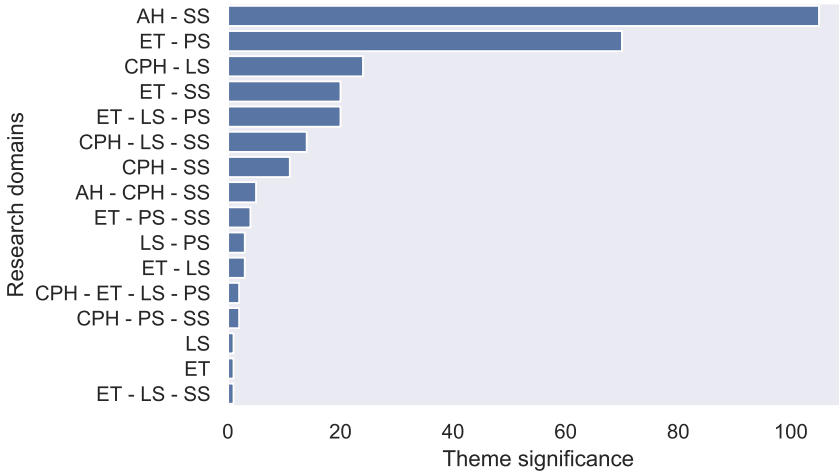


Figure 5.11: Theme significance of different research areas. To save the space, the following abbreviations are used: LS = Life Sciences, AH = Art and Humanities, CPH = Clinical, Pre-Clinical and Health, SS = Social Sciences, ET = Engineering and Technology, PS = Physical Sciences.

Table 5.5: Special issues in the field of rumour studies between 2000 and 2018.

Journal	Call	Time
Behaviour & Information Technology	Special Issue on Social Media in Conflicts and Crises	2018
Political Communication	Beyond fake news: The politics of disinformation	2018
ACM Journal of Data and Information Quality (ACM JDIQ)	Special issue on Combating Digital Misinformation and Disinformation	2018
Versus, journal of Semiotics and Philosophy of Language	Special Issue on "Fake news, misinformation/disinformation, post-truth"	2018
Frontiers in Psychology	Special Issue on Gossip	2018
Journal of Computational Science	Special Issue on Information Diffusion in Online Social Networks (IDOSN)	2017
Policy and Internet	Special Issue on Reframing 'Fake News': Architectures, Influence, and Automation	2017
Liinc em Revista	Disinformation, misinformation and hyper-information in contemporary digital networks	2017
Journal of Communication	Special issue on misinformation	2015
ACM Transactions on Information Systems (TOIS)	Call for Special Issue on Trust and Veracity of Information in Social Media	2015
Group & Organization Management	Special Issue on Gossip in/around Organizations	2008
Review of General Psychology	Special Issue on Gossip	2004

Table 5.6: The conferences in the field of rumour studies between 2000 and 2018.

Conference	Type	Time	Location	Other description
International Workshop on Rumours and Deception in Social Media	Workshop	Oct 22, 2018	Turin, Italy	
Digital Disinformation Forum	Forum	Jun 26-27, 2018	Stanford	
CyberSafety 2018 : The Third Workshop on Computational Methods in CyberSafety, Online Harassment and Misinformation	Workshop	Apr 24, 2018	Lyon, France	It was a workshop in the 2018 edition of The Web Conference (27th edition of the former WWW conference)
MIS2: Misinformation and Misbehavior Mining on the Web	Workshop	Feb 9, 2018	Los Angeles, California, USA	Workshop held in conjunction with WSDM 2018
Democracy and Disinformation	Forum	Feb 12-13, 2018	Ateneo de Manila, Philippines	
MisinfoCon: A Summit on Misinformation	Summit	Feb 24-26, 2017	Harvard and MIT Media Lab in Cambridge, MA	It explored exploring the psychology of misinformation and strategies to strengthen the trustworthiness of information across the entire news ecosystem
Journalism, Misinformation and Fact Checking	Conference Track	Apr 23-27, 2018	Lyon, France	It was a track in the 2018 edition of The Web Conference (27th edition of the former WWW conference)
ICWSM 2017 Workshop on Digital Misinformation	Workshop	May 15, 2017	Montreal, Canada	It was in conjunction with the 2017 International Conference on Web and Social Media (ICWSM)
Science journalism in a post-truth world	Session	Jun 29, 2017	Copenhagen, Denmark	It was one of the session of the 4th European Conference for Science Journalists (ECSJ2017)
Challenges in strategic communication and fighting propaganda in Eastern Europe. Solutions for a future common project	Workshop	Apr 25-27, 2018	Moldava, Czech Republic	It was held in Moldova with a Warsaw Institute's participation
Combating Fake News An Agenda for Research and Action	A working meeting	Feb 17, 2017	Harvard Law School	
Reporting Facts and the Future of Journalism	Forum	Aug 17-18, 2017	Singapore	
Defending Democracy: Civil and Military Responses to Weaponized Information	Forum	Apr 7, 2017	Princeton University	
Nordic Conference - Information & Misinformation	Conference	Sep 21-23, 2016	Helsinki, Finland	It was about misinformation and misconception about eating disorders
Rumors and Deception in Social Media: Detection, Tracking, and Visualization	Workshop	May 19, 2015	Florence, Italy	

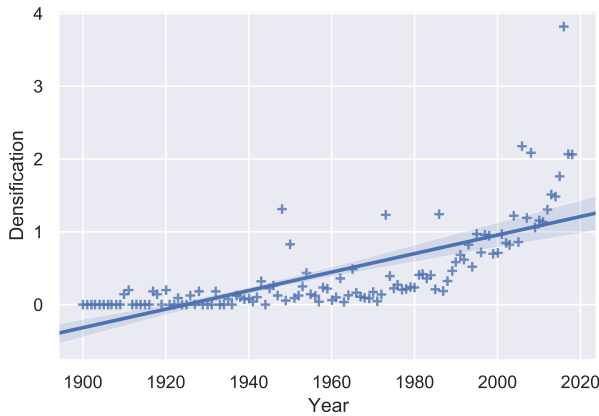


Figure 5.12: The growth of densification in author level

network in last 120 years, which confirms what we said previously that there is a growing interest among the scholars to study the field of rumour studies; however, this interest is in its infancy level.

5.4.4. IMPACT

For impact, as [188] explains we need to investigate how much expectation this topic of interest can build up. For this purpose, we study the underpinning disciplines and funding agencies to see how much expectation the study of rumours could create for the academic world. In the following box, we go through the procedure of measuring impact.

Procedure of measuring impact

- Underpinning disciplines
 1. Counting the number of underpinning disciplines in each year
 2. Measuring *DisciplineScore* for each discipline
- Funding
 1. Counting the number of fundings acknowledgements and number of engaged funding agencies per year
 2. Making the histogram for number of funding acknowledgements in different funding agencies

Firstly, we count the number of underpinning disciplines per year. As Figure 5.13a shows, the number of underpinning disciplines has an increasing trend in such a way that currently out of 252 Web of Science subject categories, most of them has contributed to this topic of interest. The number of underpinning discipline demonstrates a high level of engagement in the topic of rumour spreading from academia. However, this number does not tell anything about the quality of engagement. For example, some subject categories may contribute more than 100 articles and some others less than 10, and this criterion cannot distinguish them. So, relying only on the number of underpinning disciplines might be misleading.

To tackle this issue, in the second step, we define a criterion called discipline score. Discipline score measures the number of publications in a particular discipline. For instance, if in the topic X , subject categories of Y_1 and Y_2 contribute with 10 and 50 articles, respectively, then

$$DS(X, Y_1) = 10, \quad (5.6)$$

$$DS(X, Y_2) = 50, \quad (5.7)$$

Where DS is discipline score.

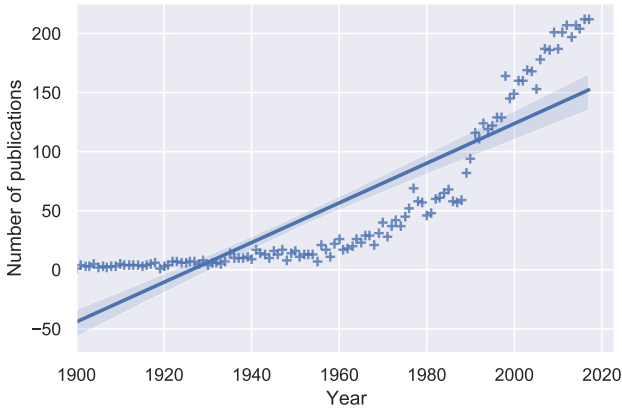
We measure discipline score for all the subject categories which contribute to this topic. Figure 5.13b, shows the histogram of discipline score. As this figure displays, except a handful of subject categories, most of them contributed to the topic of interest in a fairly small scale. In fact, more than 80% of the subject categories have less than 100 publications. This means, although this topic of interest has made a huge expectation in academia by attracting almost every subject category, most of them made a contribution by very few publications. In fact, a substantial part of academic contribution comes from a few subject categories which mean only a few disciplines attracted and contributed to this topic of interest genuinely, and for the rest, it has just made a temporary expectation.

The other part of the expectation analysis would be the analysis of funding agencies¹⁶. We first investigate the number of funding acknowledgements and a variety of funding agencies which provide those fundings acknowledgements. As Figure 5.14a illustrates, the number of fundings acknowledgements increased dramatically since last 10 years, but at the same time the variety of the engaged agencies also increased significantly (Figure 5.14b). This growing interest does not belong to a handful of funding agencies. In fact, there are plenty of agencies that are interested in this topic; however, they adopt prudent approach by providing funding acknowledgements to not many projects. It is also worth mentioning, those agencies are often among the giant funding organisations such as NSF, ESRC, and EU.

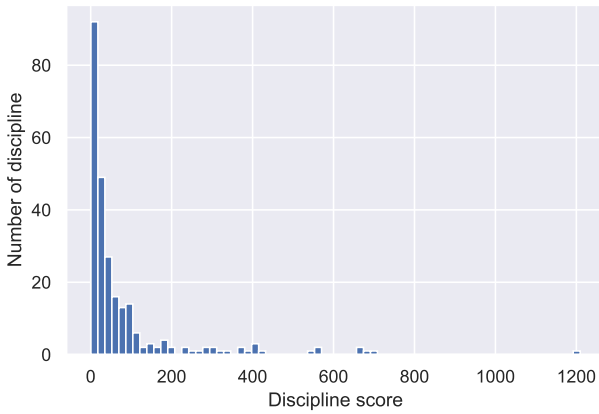
In the second step, we further clarify this point by making the histogram for number of funding acknowledgements in different funding agencies (Figure 5.14c)¹⁷. As it is evi-

¹⁶The funding data is available for records entered into Web of Science Core Collection (which covers a large collection of journals, books, and conference proceedings in 254 disciplines) since August 2008. However, in November 2016, funding data in Web of Science Core Collection was supplemented with funding information from Medline and researchfish. Medline started capturing funding data from 1981 on. Any Web of Science Core Collection record that did not have funding data was updated with grant information from Medline or researchfish if available.

¹⁷It is a log-log diagram

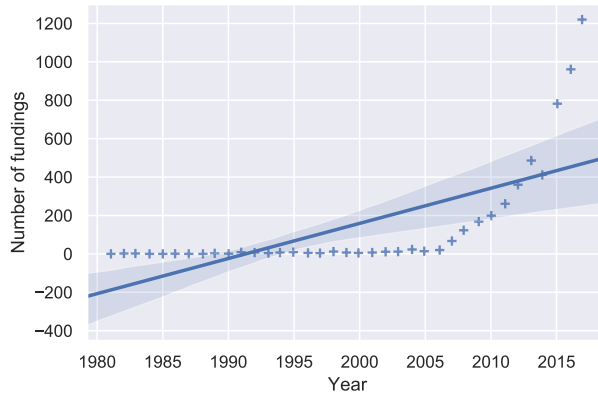


(a) The growth of contributing disciplines in the field of rumour studies

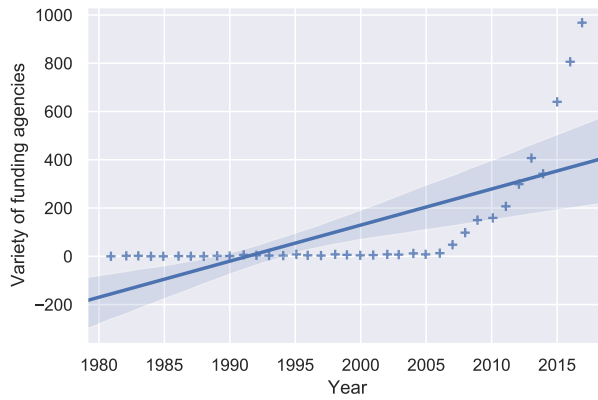


(b) Number of disciplines with different discipline score

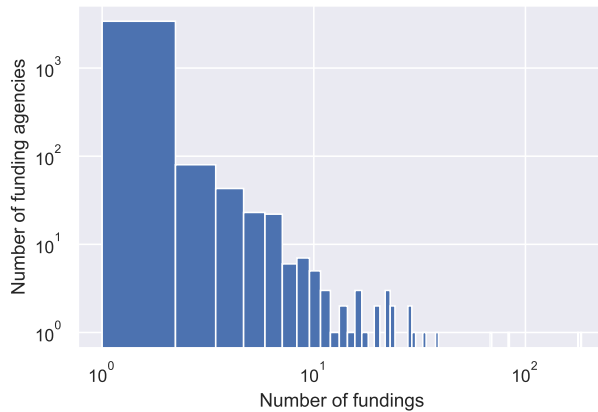
Figure 5.13: Assessment of impact using academic disciplines contribution to the field of rumour studies



(a) Growth of funding acknowledgements



(b) Growth of funding agencies



(c) Number of funding agencies with different fundings

Figure 5.14: Expectation analysis for funding acknowledgements and funding agencies in the field of rumour studies

dent, many agencies are engaged in this topic of interest by providing only one funding acknowledgement while very few agencies provide a relatively high number of funding acknowledgements. Nevertheless and despite the discreet approach taken by funding agencies, this topic has created a huge expectation by attracting too many funding acknowledgements; however, the temporary presence of the majority of them shows that the genuine expectation is only realised for a few of them.

5.5. DISCUSSION

This section tends to discuss the results of the analysis. We measured the performance of each dimension over time, to compare the performance of a research topic with itself. In the dimension of novelty, we observed a considerable decline in the novelty level through time. In the growth dimension, we saw a gradual increase. In the coherence dimension, we saw a broad consensus over the research theme and the low level of coherence in the research community, and in impact dimension, we observed that it pursues a growing pattern. By putting all those dimensions together, we will have a schema of the topic of interest (Table 5.7) which show the dynamic of change in the field of rumour studies from the emergence dimensions perspective.

Table 5.7: The schematic dynamic of emergence dimensions in the field of rumour studies

	Novelty	Growth	Coherence	Impact
Change direction	↓	↑	↑	↑

Although it is not possible to compare dimensions with each other, we can say our analysis shows an increasing trend for growth, coherence and impact and decreasing trend for novelty. But these trends need to be discussed. For growth, we see a considerable part of publications come from newcomers, while incumbent authors have a very small share in the growth of the topic of interest. For coherence, we see a broad theme coherence around (i) social science and humanity, and (ii) physical science and engineering & technology. In community coherence, we see a weak community around the topic of interest with a couple of special issues, conference tracks and forums. In impact, we see despite the significant contribution of funding agencies and academic disciplines, only a few of them actively contribute in the topic of interest.

This topic of interest had organic growth so far, which is not promising enough for confronting the problem of rumour dissemination. This problem can be tackled by a bit external push toward the formation of a stable community and realisation of an active field for rumour studies. Having an active field followed by a community of scholars cause to have dedicated researchers in this field which settles the problem of disparity between incumbents and newcomers via increasing the number of incumbents. Moreover, an active field entails having publication venues for sharing the most recent achievements with the fellow scholars. Besides, when there is an active community which produces a substantial amount of knowledge about a very severe challenge in a reasonable amount of time, funding agencies take that field into account more seriously and consider it for more funding opportunities. Also, having an active community of researchers in funded projects with the opportunity of publication in dedicated journals

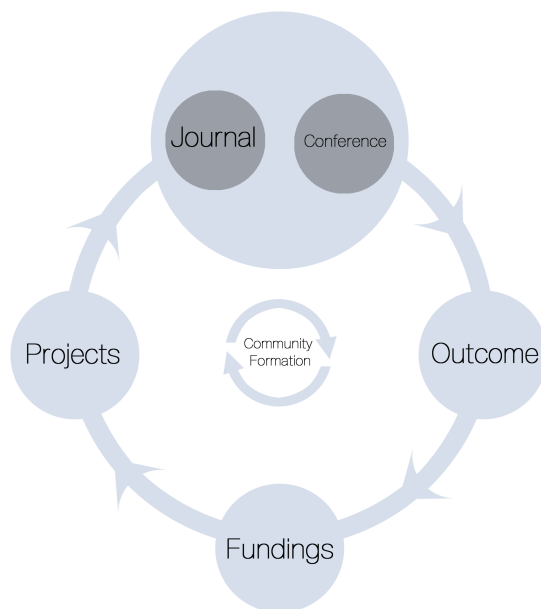


Figure 5.15: The community formation in a research field.

and conferences, cause researchers to explore undiscovered areas in this field and try to answer unanswered questions, which automatically improves the novelty level in the field.

As it is explained before, to make this topic of interest more active, we need an external push. The proposed solution is illustrated in Figure 5.15. The first step in our approach is arranging dedicated publication venues such as journals and conferences. They provide the opportunity of knowledge sharing and networking among the researchers. This can strengthen the ties between the scholars of the field and give them new ideas and inspirations for their research. Having regular conferences and journals improves the growth of the field due to the flow of the research outcomes coming to the conferences/journals. Stable growth in a topic of interest is the sign of productivity which is a positive signal for funding agencies and can convince them to dedicate more resources for doing research in this topic of interest. More fundings means more projects which require more researchers in this topic of interest which ultimately lead them to publish their works in conferences and journals. By repetition of this cycle over time, the community of the researchers around this topic of interest gets bigger, more structured and more established. Eventually, compared to the early stage of the field, such a setting leads to a scientific field with a higher level of novelty, growth, coherence and impact.

The other subject of discussion is the absence of baseline analysis. As it is explained earlier, the dimensions of the Rotolo's framework provide an indication of emergence when they are considered in the domain in which the given technology is arising, and, therefore, based on this framework, an emergent technology may only be compared with

the other technologies if they are in the same domain. To the best of our knowledge, there is no study from emergence perspective close to the research domain of this chapter; thus what we could do is to continue the retrospective approach in the analysis of this field of research every few years to keep the readiness of this field in check.

5.6. CONCLUSION

In order to counter rumours effectively, three approaches are required to be taken into account: exposure minimisation, immunisation, and reducing the transmission rate. Among those immunisation is one of the most effective approaches. Academia is on the front-line of developing immunisation techniques. Despite all the scholars' efforts, there is still no clear picture of the academic readiness. This might lead to the overestimation or underestimation of the potentials and abilities in creating societal resilience against rumour circulation. The lack of clear picture of the existing situation would deflect decision making and might result in the wrong policies to confront rumour dissemination.

To address this issue, this research aimed to study the scientific efforts of thousands of researchers all around the world in the field of rumour studies from the emergence perspective. Although the first signs of this research topic appeared almost a century ago, in the recent years, it has reached to a certain momentum (regarding knowledge production) which allows us to study different dimensions of emergence. To measure the status of this research topic, this study adopts the emergence framework suggested by Rotolo et al [188]. It is one of the first and most comprehensive models which identifies the components of the emergent phenomenon and studies their dynamics. Our results show while growth, coherence, and impact in this field are increasing, novelty is declining. Although there is no baseline that could be used to compare our results, one approach to make sure of the outcomes reliability is to use alternative operationalisation and see whether they indicate similar trends or not. This would also show the dependency of our analysis to a particular operationalisation.

Although the results of the analyses show slow organic growth in this field of research, it could be an alarming trend because the speed, scope, and scale of rumour spreading is increasing while the development of counter strategies is advancing at a much slower pace. This problem can be tackled by an external push strategy toward the formation of an active research community.

6

COMPUTATIONAL RUMOUR DETECTION USING ONE-CLASS CLASSIFICATION

I propose to consider the question, 'Can machines think?'

Alan Turing

As we discussed in Chapter 4, in addition to creating immunity against the rumours (Chapter 5), exposure minimisation is an essential piece of integral plan to confront this notorious phenomenon. To address this issue, the following question is put forward:

- *How could we identify rumours in social networks automatically, consistently and in a timely manner?*

One of the most effective ways to minimise the appearance and circulation of rumours within social media platforms is the computational approach that can tackle the large-scale spread of rumours in a timely fashion. The dominant computational technique for this goal is the binary classification which uses rumour and non-rumour for the training. Such an approach leads to unreliable classifiers which cannot distinguish rumours from non-rumours consistently. In this research, we tackle this problem via a novel classification paradigm, called one-class classification (OCC). In this approach, the classifier is trained with only rumours, which means we no longer need the non-rumour data-points. The experimental setup consists of feature extraction from two primary Twitter datasets and then running seven one-class classifiers from three different learning paradigms. Our results show that OCC can outperform binary classifiers with a high level of F1-score¹.

¹This chapter is based on the following publications:

6.1. INTRODUCTION

Rumours used to propagate by means of word of mouth, newspapers, radio, and television. However, in recent years, the emergence and rapid growth of online social networks turned this problem into a major socio-political challenge due to the easy, fast, and wide propagation of information in online social networks. Because of volume, velocity, and variety of rumours in social networks, researchers use large-scale data and computational techniques to control this phenomenon [202, 203, 204, 205]. One of the important steps in a computational rumour control system is rumour detection which has been a topic of interest for the community of computational social science in a past couple of years. The main goal of rumour detection is to identify rumours in online social networks automatically, accurately, and in a timely manner.

Binary classification is the dominant computational approach for rumour detection [206]. In this approach, a model is built by training the classifier with a dataset comprising samples of rumours and non-rumours. We then evaluate the discrimination power of the model by subjecting the trained classifier to a mixed set of rumours and non-rumours. The model with higher performance in separating rumours from non-rumours is considered as a better classifier. Although this is a well-established approach in the literature and several scholars used it for the computational rumour detection, it suffers from a serious flaw, namely, the non-rumour pitfall.

Non-rumour is a fuzzy concept which is widely used by scholars for computational rumour detection. Unlike the rumour which has a long-standing definition owing to its solid background in social psychology, non-rumour has an ambiguous meaning which is not supported by either epistemological or psychological studies. Ambiguity in this concept leads to different ways of data collection/annotation. In other words, the lack of a clear definition allows scholars to come up with their own readings of non-rumour [202, 207, 208]. For instance, Kwon et al. [202] consider any kind of factual information and news as non-rumour, while Zubiaga et al. [209] annotate tweets which are not rumour as non-rumour. This creates a confusing situation, because from one hand, there are some tweets that are not rumour, and on the other hand, they may or may not take the non-rumour label. Therefore, the model may receive data-points which do not belong to its pre-defined classes. Those data-points will randomly be classified by a binary classifier as rumour or non-rumour.

The implication of such discrepancy between various non-rumour definitions is that the binary classification is not the right approach for computational rumour detection.

- Fard, A. E., Mohammadi, M., Chen, Y., & Van de Walle, B. (2019). Computational Rumor Detection Without Non-Rumor: A One-Class Classification Approach. *IEEE Transactions on Computational Social Systems*, 6(5), 830-846.
- Fard, A. E., Mohammadi, M., & van de Walle, B. (2020, June). Detecting Rumours in Disasters: An Imbalanced Learning Approach. In *International Conference on Computational Science* (pp. 639-652). Springer, Cham.
- Fard, A. E., Mohammadi, M., Cunningham, S., & Van de Walle, B. (2019, June). Rumour As an Anomaly: Rumour Detection with One-Class Classification. In *2019 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)* (pp. 1-9). IEEE.

To access the codes used for the analysis of the tweets, please refer to <https://gitlab.com/amiref/iee-tcss-paper> and <https://gitlab.com/amiref/iee-tcss-paper>

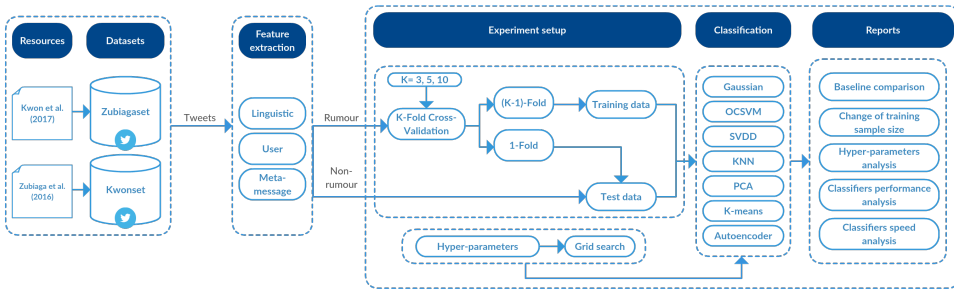


Figure 6.1: The research flow of rumour detection with one-class classification approach.

In other words, every scholar can define non-rumour according to which they collect data from social networks. This raises two important issues regarding rumour detection systems in real situations. First, the results of rumour classifiers become inconsistent, and second, the results of different classifiers cannot be compared. Therefore, the question that arises here is, *how can rumours be identified in social networks regardless of the non-rumour definition?*

We address this question from the one-class classification (OCC) perspective [210, 211, 212]. One-class classification is a supervised algorithm with the ability to identify class X between multiple classes of data when the classifier is trained by the class X only. For the computational rumour detection, this means that the classifier is trained with rumours only, and the trained classifier can be used to detect rumours from other kinds of tweets. In contrast to non-rumour, which is not well-defined and has a controversial conceptualisation, there is a consensus in the literature for rumour, and scholars often follow definitions with similar elements.

In this research, we benefit from two available datasets, one from Zubiaga et al. [204] with 6,425 tweets and the other from Kwon et al. [202] with 140,910 tweets. We extract 86 features from each tweet. After feature extraction, we build models with seven algorithms that implement one-class classification approach. Each algorithm has two hyper-parameters which impact its performance. We use grid search over the hyper-parameters to discover the best performance of the models. Figure 6.1 summarises the research workflow in three stages: (i) Dataset preparation, (ii) Feature extraction, and (iii) One-class classification.

The remainder of this chapter is organised as follows. In Section 6.2, we briefly review the literature of computational rumour detection and discuss the data, the feature, and the algorithm as three main pillars of this subject domain. Section 6.3 elaborates on the data by explaining the procedure of building datasets as well as introducing the publicly available datasets. Section 6.4 explains the set of features used in this work and their categorisation in detail. In Section 6.5, we scrutinise the concept of non-rumour and the fallacies of the binary classification in the computational rumour detection, then we discuss the general idea behind a one-class classification approach and explain seven renowned algorithms with this approach. In the last part of this section we present the experiments and report the results. Finally, we conclude this chapter by summarising the main finding and discussing the results in Section 6.6.

6.2. COMPUTATIONAL RUMOUR DETECTION

The computational solutions are among the most recent approaches that academia has adopted to tackle the problem of rumour spreading in social networks. This family of solutions is usually part of a framework called rumour resolution system which consists of four modules: (i) rumour detection, which specifies whether a piece of information is relevant to a rumour or not; (ii) rumour tracking, which collects the posts discussing the rumour; (iii) stance classification, which determines how each post orients to the rumour's veracity; and (iv) veracity classification, which determines the truth behind the rumour [206]. The four modules are shown in Figure 6.2.

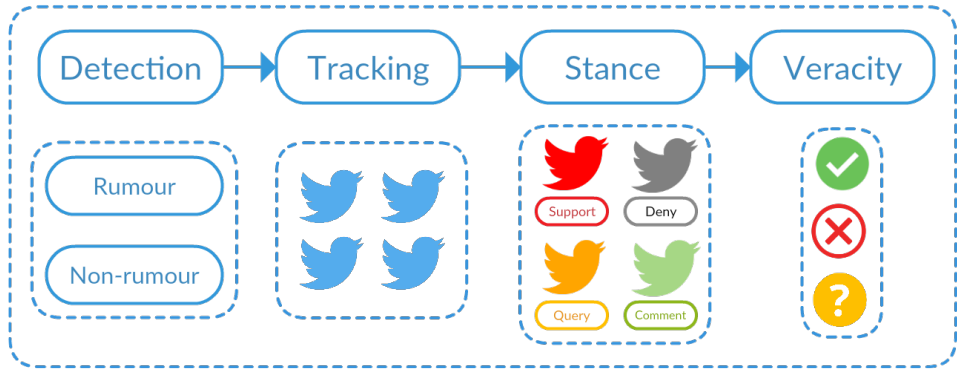


Figure 6.2: Rumour resolution system has four modules: (i) rumour detection for identifying rumour related information; (ii) rumour tracking for collecting the posts discussing the rumour; (iii) stance classification for determines posts' orientations toward rumours' veracity; and (iv) veracity classification for verification of truth behind the rumour [206].

In this chapter, we focus on the detection module in Twitter which aims to flag tweets automatically as rumour or non-rumour. This is a crucial module in rumour resolution system as it allows to identify newly emerged rumours in Twitter. In the rumour detection problem, the main focus is on providing an accurate representation of rumour for the computer in such a way that it can recognise rumours with minimum error. Such a representation is a function f of three elements, i.e., data, features and learning algorithm;

$$\text{Quality of Rumour Classifier} \propto f(\text{data}, \text{features}, \text{learning algorithm}).$$

If rumours are modelled via sufficient data, illustrative features, and a powerful algorithm, the rumour detection system is expected to identify rumours accurately, but any flaw in those elements can affect the quality of the rumour detection. In the study of digital rumours, obtaining sufficient data is an arduous task due to the difficulties and barriers of knowing the latest rumours and restrictions of social networks over the data collection. Therefore, collecting and annotating data in different topics and from various social networks is considered as an important contribution to this field. For instance, Zubiaga et al. [208] created a dataset of tweets, containing 6,425 rumours and

non-rumours, posted during five breaking news events². Sicilia et al. [207] also made a contribution by building a dataset of Twitter rumours and non-rumours containing 709 samples regarding Zika virus. In other work, Yang et al. [205] provided the first dataset of Chinese microblogs from the biggest microblogging service in China, namely, Sina Weibo. In the same vein, Turenne [16] created the first French dataset in this field by collecting 1612 rumour related texts. Additionally, in one of the biggest datasets in this field, Kwon et al. [213] provided a dataset comprising more than 140,000 rumours and non-rumours in a wide range of topics from politics to entertainment³.

The second important element of rumour detection is feature extraction. Rumours in their raw formats are incomprehensible for the computer. Hence, it is essential to find a way to make this concept machine-readable. The feature extraction aims to respond to the same need by selecting d quantifiable properties and representing each rumour with those properties to the computer. More formally, in this step, we map every data-point (rumour) to a d -dimensional space, where each dimension represents one property of rumours. The better the features can reflect different dimensions of rumours to the computer, the more realistic the understanding of the computer from this phenomenon will be. Many of the computational rumour scholars studied new features and their impact on the rumour detection. For instance, Castillo et al. [214] introduced features related to the account holders and their social networks. In the other work, Kwon et al. [213] introduced temporal features and studied the importance of temporal patterns in rumour detection. In another work, Kwon et al. [202] discovered the effectiveness of features in detection of rumour varies in the course of time. They concluded, some features (user and linguistic) are more effective for rumour detection in early stages of diffusion while some others (structural and temporal) are more effective for rumour detection in longer time windows.

The third important element in computational rumour detection is the learning algorithm. In this step, a classifier defines its decision boundary to understand what rumour and non-rumour are. As the decision boundary gets more accurate, the classifier performance improves. Computational rumour researchers apply new algorithms, statistical analysis, or methods to this domain in order to detect rumours more accurately. To be considered as a contribution, these algorithms should not be necessarily newly-designed. They might be popular algorithms in other contexts; however, they have not been used in rumour detection before. For instance, Ma et al. [215] applied recurrent neural networks (RNNs) to the rumour context for the first time. They evaluated their model with three widely used recurrent units, tanh, LSTM and GRU, which could perform significantly better than state of the art. In the other work, Zubiaga et al. [204] modelled the rumour tweets as sequences by extracting their features in the course of time and applied the conditional random field (CRF). This allowed them to obtain the context from a set of tweet and identify rumours with higher performance.

²This dataset is further explained in Section 6.3.2

³This dataset is further explained in Section 6.3.2

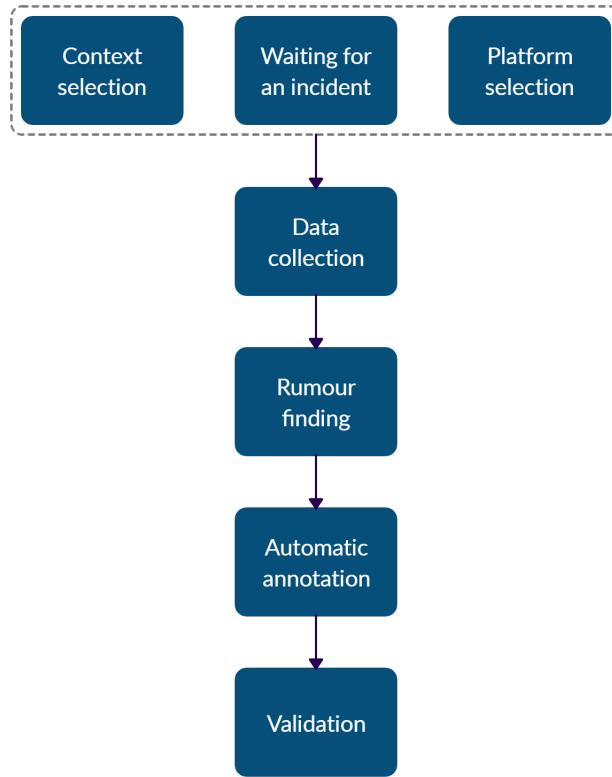


Figure 6.3: The methodology of building a dataset for computational rumour detection.

6.3. DATA

In general, there are two ways of access to data in machine learning problems: (i) building dataset, and (ii) using publicly available datasets. This section tends to elaborate on each approach in the context of computational rumour detection.

6.3.1. BUILDING THE DATASETS

This section explain the methodology of building datasets for the computational rumour detection. It is inspired by the methodology proposed by Kwon et al [202] and composed of four main steps: (i) data collection, (ii) rumour finding, (iii) automatic annotation, and (iv) validation. However, before starting the data collection, the system has to be set up. To this end, the context, the rumours, and the platform(s) should be specified. Figure 6.3 displays the required steps to build a dataset for the computational rumour detection. In the following each component of this methodology is explained.

Context selection In this step the broad context of data collection is specified. Most of the traditional studies on the rumour orient toward disasters and crises. Koenig [3]

introduces crisis, conflict, and catastrophe (or what he calls as three C's) as the three main contexts for the emergence of rumours; however, rumours are not limited to those contexts whatsoever. All subject domains that could satisfy prerequisite conditions for the emergence of rumours are good candidates for data collection, whether it is politics, health, or business [3, 7].

Waiting for an incident In this step the possible incidents for the rumour appearance are selected. Some of the incidents are predictable and the time of their occurrence could be anticipated within a particular time window. For instance, meteorological centres often forecast floods and hurricanes a couple of days before they begin. The time of political incidents such as elections is also fixed months before. Nevertheless, other incidents such as earthquakes, blasts, or terrorist attacks are unpredictable. For the predictable incidents, we can set up a data collection system a few days, weeks, or months (depending on how early we know about the incident). However, for the unpredictable incidents it is of utmost importance to constantly monitor the relevant news outlets. By relevant, it means if the topic of interest is within a particular geographical domain or has a specific subject domain, we should devote most of our attention to niche outlets in those domains.

Platform selection With the increasing availability of social media data over the past decade, scholars from multiple disciplines have been able to study human behaviour on micro-scales to not only test old social theories but also propose new ones. Figure 6.4 illustrates a growing trend of research conducted using social media data across disciplines. As this figure displays, Twitter is by far the most popular social media platform and has the highest rate of growth between different platforms. As this figure displays, Twitter is by far the most popular social media platform and has the highest rate of growth between different platforms. Perhaps having a relatively smooth API as well as keeping it open for public⁴ even after 2016 U.S. election incidents where in other major platform namely Facebook closed most of its APIs, are why researchers tend to use Twitter.

Data collection After the initial setting about the context, the incident, and the platform, the data collection starts. Assuming that the chosen platform is Twitter, the API has to be set up and the query should be defined. For the query it is recommended to be as inclusive as possible to include as many relevant tweet as possible. In this step, it is very important to collect data in conversation level which means when a query returns a tweet, the whole conversation of that tweet including replies, retweets, mentions, and likes should also be collected. Nevertheless this might not be possible due to the limitations of Twitter Standard API in collection of replies or mentions.

There is an important point which needs to be discussed. After tweet collection, it is highly likely that quite a few number of retweets appear in the dataset. Some of the scholars recommend to leave those tweets out of the dataset as they are repeating the exact same message [208]. Although the retweets have the same content, the other aspects

⁴Twitter API has three access levels of standard, premium, and enterprise. The Twitter API that we refer to it in this chapter is the standard API which has limitations, but it is free.

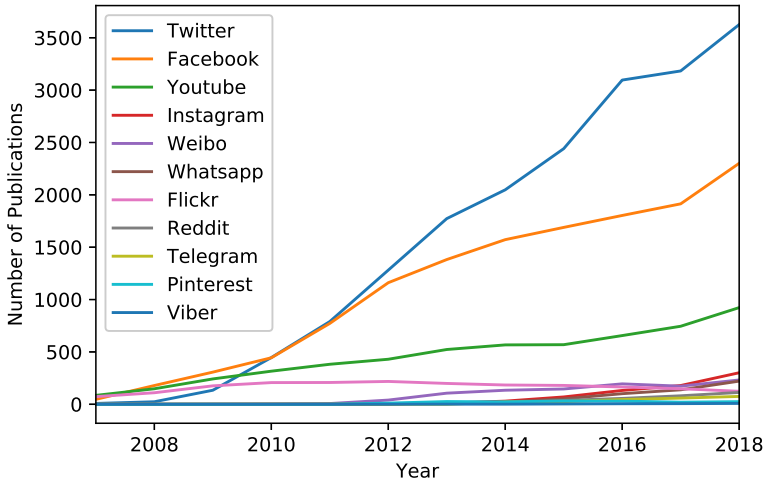


Figure 6.4: The number of publications regarding 11 popular social media platforms from 2008 to 2018 based on Scopus data. This figure illustrates the growing trend of using social media data by researchers. As the figure shows, scholars tend to work with Twitter data more than other platforms.

are totally different. They are produced by different people with different user profile in different locations, hence retweets are not identical whatsoever. Besides, elimination of the retweets would affect the phenomenon of rumour. In the non-digital version of rumour spreading, the original message changes and loses its original shape while it is transmitting between the people [4]. But nowadays, there is a possibility for transmitting the exact same message (retweet) as well as the reshaped ones (retweet with comments), thanks to online social media. Thus, we can infer that retweeting is part of digital rumour transmission as message reshaping is part of rumour spreading in a non-digital environment. On this account, elimination of retweets from a body of rumourous tweets, impacts the whole rumour spreading phenomenon and changes it to something else.

Finding rumours After data collection from a social media platform regarding a particular incident, it is the time to select rumour related tweets which are related to that incident. To this end, we should find a list of running rumours during the incident. To make that list, first, a list of credible local and international news and fact-checking outlets is required. It is notoriously difficult to measure the credibility of a news or fact-checking outlet. However, for the news-outlets, the ones with long and strong editorial histories are considered as credible. Besides, for the fact-checking outlets, being a member of IFCN (International Fact-Checking Network) is a signal of credibility. After creating the list of the outlets, by checking all the posts in those outlets concerning the incident, we would be able to create a comprehensive list of rumours regarding the particular incident.

Annotation Data annotation is an expensive, labour intensive, and time-consuming task; however, the growth of crowdsourcing systems provided us with the opportunity to tackle those issues by hiring low-cost annotators and get the dataset annotated in a reasonable amount of time. Despite those benefits, crowdsourcing approaches are also suffering serious flaws such as non-educated annotators, and lack of gold standard [216]. One way to tackle this issue is using expert annotators like Zubiaga et al. collaboration with journalists [209], which may not be easy, affordable, and accessible for all researchers.

The other approach is the automatic annotation which is proposed by Kwon et al. [202]. In this strategy, a tweet receives the rumour label if it contains explicit keywords (signal words) associated with a particular rumour. Like other annotation strategies, this method has some potentials and limitations. On the bright side, it is a fast, cheap, and scalable approach; however, there are two pitfalls regarding this method: (i) exclusiveness, and (ii) inclusiveness. The exclusiveness refers to choosing a very limited set of words which results in missing rumours without signal words. The exclusive annotation strategy has a tendency to high precision but low recall due to high false negatives and low false positives. On the other hand, inclusiveness means to cast our net unusually wide in order to identify more rumours, but it comes with the risk of assigning the label of "rumour" to non-rumours. Because of high false positive and low false negative the inclusive annotation strategy has a tendency to low precision and high recall. In order to tackle those pitfalls in automatic annotation, we should have a comprehensive knowledge about the incident, its context, and its associated rumours. We should also be aware of different wordings of the same rumours which help to not miss them in the dataset. Besides, this is an iterative task and after a few rounds of trial an error in building the signal set, the number of false negatives and false positives drop significantly.

One of the controversial points here is about debunking tweets. The debunking tweets tend to ascertain the truth behind the rumours by providing concrete evidence. However, such tweets often comprise the original rumour which means having the signal words. As it is discussed in Chapter 5, one of the main principles in rumour quelling strategies is to avoid repeating them [170]. Any strategy that results in rumour repetition will reduce the chance of countering this phenomenon. This principle implies that regardless of message content, it should not be repeated whether it is pure rumour message or debunking message. This would justify of being agnostic regarding tweet orientation in annotation process.

Validation The validation step allows us to measure the performance of the automatic annotation by counting the number of true positives, true negative, false positives and false negatives. In this step, a fraction of automatically annotated tweets is selected, then they are manually labelled by multiple annotators. In the last step, the automatic and manual labels are compared.

6.3.2. AVAILABLE DATASETS

In addition to data collection and building datasets by the scholars, they may also use available datasets for multiple reasons. First, the preparation of annotated datasets is a demanding task. It might also become costly due to hiring annotators. Besides, available

6

datasets in a field allows researchers to compare the performance of their models. There are multiple datasets in the field of rumour studies; however, not all of them are publicly available. Some are available upon request which means a request has to be issued to the owners of the datasets and then they decide to share their data or not. This is an unreliable situation since there is no guarantee that submitting a request would provide access to the datasets. For instance, based on the authors experience, in some cases we immediately received the datasets by asking the owners [16, 217]; however, sometimes they refuse to share their data with excuses such as lack of time for the preparation of data [207], or retirement of the project [218]. In addition to available upon request datasets, there are limited available datasets which are publicly available but not for arbitrary uses. One of the most famous dataset with such a condition is the massive dataset used in the seminal work of Vosoughi et al. [13]. In that project, the dataset is entirely available upon signing an access agreement stating that "(i) you shall only use the data set for the purpose of validating the results of the MIT study and for no other purpose; (ii) you shall not attempt to identify, reidentify, or otherwise deanonymize the data set; and (iii) you shall not further share, distribute, publish, or otherwise disseminate the data set". The first condition in this agreement deprives scholars of any experiment further than the ones in the original study. This would make the dataset not beneficial for research purposes. The other type of datasets in this field is publicly available datasets which are open to everyone to download and use the data. The two renowned publicly available datasets in computational rumour studies are introduced by Zubiaga et al. [204], and Kwon et al. [202]. We refer to these datasets as the Zubiagaset and Kwonset, respectively. They are quite renowned in this field and have been appeared in several research studies [219, 220, 221, 222]. They cover a wide variety of topics including disaster, health, and politics to name but a few.

In the first dataset, Zubiaga et al. [204] used Twitter Streaming API to collect tweets in two different situations: (i) breaking news that is likely to spark multiple rumours; and (ii) specific rumours that are identified a priori. Tweets are collected from five cases of breaking news events (Charlie Hebdo shooting, Ferguson unrest, Germanwings crash, Ottawa shooting, and Sydney siege)⁵. Given the large volume of tweets in the early stages of the data collection, they only sampled the tweets that provoked a high number of retweets. Then, they manually annotated the tweets as either rumour or non-rumour. In total, they collected 6,425 tweets including 4,023 non-rumour and 2,402 rumour tweets.

In the second dataset, Kwon et al. [202] made a list of popular rumours by searching fact-checking websites. They also made another list for non-rumours by searching for notable events from news media outlets. The entire list of rumours and non-rumours cover wide variety of topics from health and business to politics and technology⁶. After preparing the list of rumours and non-rumours, they crawled one of the largest and near-complete repositories of Twitter to collect tweets relevant to the list of rumours and non-rumours. In total, they identified 140,910 tweets for 111 events (44,394 tweets from 60 rumours and 96,516 tweets from 51 non-rumours). To the best of our knowledge, Kwon dataset is the biggest publicly available dataset in this field to date.

⁵For further information about the dataset please visit <https://bit.ly/39YOOPI>

⁶For further information about the dataset please visit <https://bit.ly/3cXsEz7>

Table 6.1: The statistical information regarding Zubiaga [204] and Kwon [202] datasets.

	Zubiagaset	Kwonset
Number of tweets	6,425	140,910
Number of rumours	2,402	44,394
Number of non-rumour	4,023	96,516
Description	Based on five breaking news events (Charlie Hebdo shooting, Ferguson unrest, Germanwings crash, Ottawa shooting, and Sydney siege)	Based on 111 notable events and popular rumours (in wide range of topics such as politics, entertainment, health, etc.).

6.4. FEATURE EXTRACTION

This section explains the second key step in computational rumour detection which is the feature extraction. Although plenty of features have been proposed by the rumour scholars [207, 202, 214], some of them are suffering from the lack of early availability and volatility. By the early availability, we refer to those features that are ready to extract at the moment a tweet is released, and by volatility, we mean the change in the value of a feature over the time. The early availability is highly important because the ultimate goal of rumour detection is to identify rumours and taking them down as early as possible. Otherwise the rumour exposure increases as the tweet lingers and circulates in Twitter. The feature volatility is also very important. If a tweet feature changes over the time then the model built upon such features would be unreliable. Thus it is essential to avoid collecting features which are not early available and change constantly over time.

Among the features that are introduced in this section, some of them are extracted from the existing literature of computational rumour detection, some are borrowed from the literature of similar subject domains such as detection of cyberbullying, inauthentic behaviour, or spam, and some are proposed based on the soft literature of rumour mostly in psychology and journalism.

Among the literature based features, many of them are redundant. They all try to capture the same aspects of the tweets but with different lenses. One of those redundancy cases is when one feature counts a specific element amongst a pool of elements, while the variation of the same feature measure the fraction of that specific elements in the pool of elements. For instance, two highly popular features are number of uppercase characters and fraction of the uppercase characters. Although they seem two different features, they are referring to the same concept. Another case of redundancy is when one feature counts something while the other variation measure whether that thing exists or not. For instance, Castilo et al. [214] introduced *Contains Hashtag* as a binary feature indicating whether a tweet contains hashtag or not; however, Bahuleyanand & Vechtomova adopted a non-binary version of this feature and measured the number of hashtags in a tweet. For such cases where there are multiple versions of a same feature, only one of them is kept. Another kind of redundant features, is when we refer to exact same features but with different names. For instance, both followers count and user influence, count the number of users who follow the account holder of a tweet. For such

features, we also only retain only one of them.

In addition to eliminate the redundant features, we had to also leave out the significant number of features which tried to capture the temporal and network behaviour of the tweets as they are not available during the initial rumour propagation phase. The remaining features are introduced in the rest of this section.

The features are often presented in different categories. There is no gold standard for the feature categorisation and different scholars develop their feature categorisation schema arbitrarily. For instance, features about tweet content has been labeled as *content features* [223, 224], *linguistic features* [218], *post-centric features* [225], *tweet-based features* [226]. Similar categories have been further distilled into two sub-categories of *message* and *topic* [214]. Here, we propose a two-level categorisation. The first level broadly classifies features in three groups of *linguistic & content*, *user*, and *meta-message*. The linguistic & content features capture synthetic and semantic aspects of the tweets, the features related to the account holders and their social network are in the user category, and all the features about tweets meta-data fall into the meta-message category. In the second level, we go one level deeper and cluster features based on the conceptual closeness. Figure 6.5 illustrates the feature categorisation schema and the features associated to each level of categorisation. The blue, yellow, and red layers represent *linguistic & content*, *user*, and *meta-message* categories in the first level, respectively. The second level composed of 12 broad categories is represented as the orange boxes. In the following, the categories and their associated features are explained in more details.

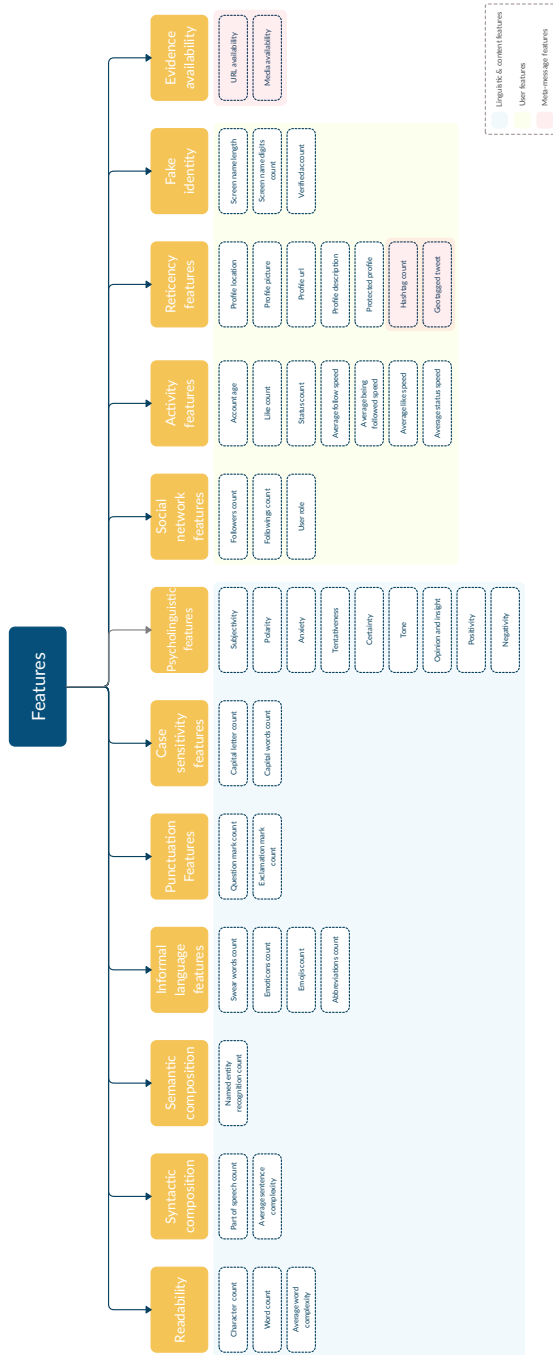


Figure 6.5: Categorisation of features for computational rumour detection.

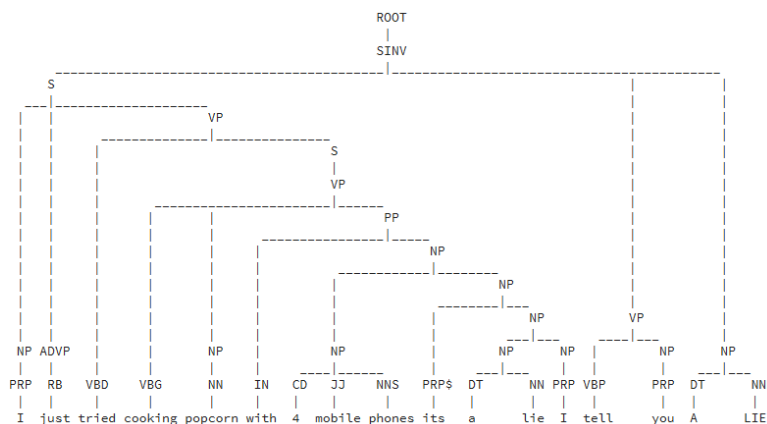


Figure 6.6: The dependency tree of "I just tried cooking popcorn with 4 mobile phones its a lie I tell you A LIE".

6.4.1. LINGUISTIC & CONTENT FEATURES

Linguistic & content analysis enables us to scrutinise semantic and syntactic aspects of the tweets. The features in this category are further classified into 7 categories of readability, syntactic composition, semantic composition, informal language, punctuation, case sensitivity, and psycholinguistics. Every category and the associated features are explained in the following.

READABILITY FEATURES

For this set of features, the readability of messages is measured. The three features for this purpose are *character count*, *word count* [214], and *average word complexity* [218]. The first two, measure the length of a tweet based on two different unit of analysis: (i) character and (ii) word. The *average word complexity* estimates the average length of words in a tweet. For example, the tweet *I just tried cooking popcorn with 4 mobile phones its a lie I tell you A LIE* has 17 words containing 1,4,5,7,7,4,1,6,6,3,1,3,1,4,3,1,3 characters respectively. The average word complexity is therefore 3.5.

SYNTACTIC COMPOSITION FEATURES

In this feature category, the tweet syntax is taken into account. One of the important set of features is the *frequency of Part-of-Speech (PoS) tags* [227]. PoS tagging is the process of assigning one of the pre-defined grammatical roles to every token in a sentence. Here we consider 19 types of PoS tags and count the frequency of every tag in each tweet. The PoS tags and their explanations are tabulated in Table 6.2. The other feature in this category is the *average sentence complexity* [218] which estimates the average depth of a tweet dependency parse tree. We use Stanford CoreNLP to generate dependency tree. The dependency tree of the previously mentioned tweet is shown in Figure 6.6.

SEMANTIC COMPOSITION FEATURES

In order to capture the semantic aspects of a tweet, we use Named-entity recognition (NER). NER intends to classify every token in a text into pre-defined conceptual cate-

Table 6.2: The PoS tags and their description.

POS	Description
ADJ	adjective
ADP	adposition
ADV	adverb
AUX	auxiliary
CONJ	conjunction
CCONJ	coordinating conjunction
DET	determiner
INTJ	interjection
NOUN	noun
NUM	numeral
PART	particle
PRON	pronoun
PROPN	proper noun
PUNCT	punctuation
SCONJ	subordinating conjunction
SYM	symbol
VERB	verb
X	other
SPACE	-

gories. Here we consider 17 types of NER tags and count the frequency of every tag in each tweet. Table 6.3 shows the tags and their description.

PUNCTUATION FEATURES

The features related to tweets punctuation fall into this category. Two of the most frequently used features in computational rumour detection are *question mark count* and *exclamation mark count* which belong to this category [214, 228].

CASE SENSITIVITY FEATURES

This category is associated with the features about case sensitivity. Similar to punctuation category, here, there are also two features that are often used in rumour detection solutions. Those features measure the number of uppercase letters and words with capital letters within a tweet [214, 228].

INFORMAL LANGUAGE FEATURES

This category captures features related to use of informal language in the tweets. In the *swear word count* feature, we count the number of vulgar words or expressions in the tweets. For this feature, we made a collection using online dictionaries⁷ including 1585 terms and checked each tweet against this collection⁸. The other feature in this category is the *abbreviations count* which count how many abbreviations are used in a tweet. To

⁷<https://www.noswearing.com/dictionary> and <https://www.cs.cmu.edu/~biglou/resources/>

⁸The list of vulgar terms is publicly available in this address: <http://bit.ly/2CtO4Rz>

Table 6.3: The NER tags and their description.

Type	Description
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	NonGPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including "%".
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	"first", "second", etc.
CARDINAL	Numerals that do not fall under another type.

measure this feature we first made a list including 2622 abbreviations using online dictionaries⁹ and Crystal's book on the language used on the internet [229], and then check tweets against this list¹⁰. The other features in this category are *emoticons count* and *emojis count*. Both these pictographs indicate different types of feelings via a tiny figure which sometimes need many words to be described. For counting emojis we use an available Python library called Emoji¹¹; however, for the counting the emoticons, we compiled a list of emoticons and made it publicly available¹²[218].

PSYCHOLINGUISTICS FEATURES

To study the semantic layer of a text, one of the most common approaches that scholars adopt is capturing the emotions, attitude or mood conveyed by a piece of text. For measuring such semantic proxies, in this research we borrow relevant features suggested in the literature such as, *tone* [230], *subjectivity* [231], *polarity* [230], *number of positive and negative words* [205]. There are also features inspired by the rumour literature such as *anxiety score* [16], *tentativeness score* [218], *opinion and insight score* [218], and *certainty score* as many of the rumours emerge at the time of uncertainty.

For the *anxiety*, *tentativeness*, *opinion and certainty* scores we used Linguistic Inquiry and Word Count (LIWC)¹³. LIWC is a tool for analysing the cognitive and psycho-

⁹<http://www.netlingo.com/category/acronyms.php>

¹⁰The list of abbreviations is publicly available in this address: <http://bit.ly/2Bxim4e>

¹¹<https://github.com/carpedm20/emoji/>

¹²<http://bit.ly/2EDDhG2>

¹³<http://liwc.wpengine.com/compare-dictionaries/>

logical theme of text documents [232]. LIWC is a commercial tool thus we are not sure about the set of signal words that it used to give a score to each of the above-mentioned psychological/cognitive dimensions. To measure the *subjectivity* and *polarity*, *positiveness* and *negativity* we use a Python library called Textblob¹⁴.

6.4.2. USER FEATURES

User analysis enables us to investigate the credibility of Twitter accounts. In this research, we carry out this analysis via extracting features related to the account holders and their social network. The features in this category are classified in four groups of social network, activity, reticency, and fake identity. Every category and the associated features are explained in the following.

SOCIAL NETWORK FEATURES

In this category, features regarding the social network of tweets' account holders are discussed. Three features in this category are *number of followings*¹⁵, *number of followers*¹⁶ [233, 230, 205, 202], and *user role* [218]. This feature measures the ratio of the followers count to the followees count for a user. A user with a high follower to followee ratio is a broadcaster. Conversely, a user with a low follower to followee ratio is considered as a receiver.

ACTIVITY FEATURES

In this category, features are about the account holders activities in Twitter. We use 7 features to capture different aspects of a user's activity in Twitter. The *account age* [214] calculates the age of the users based on their registration date in Twitter. The *number of likes and status* [205, 233, 214, 202] counts how many times a user has published or liked a post. In the remaining features, we assess the speed of engagement in four common Twitter activities namely, following, being followed, liking and posting. Since many of the trolls and social bots join social networks just for a short period, the speed of their activities is surprisingly high. Thus this set of features could distinguish them from the ordinary users with non-strategic motives. All the average speed features are calculated by dividing the value of feature by the account age.

RETICENCY FEATURES

This feature category tends to capture the reluctance of the account holders to reveal their identity. To this end, we use *profile picture* [234], *profile location* [205], *profile description* [214, 205, 230, 233], and *profile URL* [230] as a proxy to show how much a person is willing to reveal her personal information. In the same vein, the *protected profile* feature indicates whether somebody is fine if unknown users check her account.

FAKE IDENTITY FEATURES

This category is borrowed from the literature of social bot detection. Many of the bots in Twitter have screen names consisting of a random permutation of letters and numbers.

¹⁴<https://textblob.readthedocs.io/en/dev/>

¹⁵In some studies it is called the *number of friends*

¹⁶In some studies it is called the *influence score*

Using three features of *screen name length*, *screen name digits count* [234], and *verified account*, we try to highlight users with suspicious screen names.

6.4.3. META-MESSAGE FEATURES

As with the linguistic & content analysis, in the meta-message feature category, unit of analysis is tweet itself; however, here the focus is tweet meta-data, instead of its textual content. There are three categories of evidence availability, and reticency for the meta-message features. In the following each category and its associated features are explained.

EVIDENCE AVAILABILITY FEATURES

There two different kinds of evidence that might accompany a tweet: (i) a multimedia evidence such as picture or video, or (ii) a link to an external resource. This feature group uses two features of *URL availability* [214] and *multimedia availability* [233, 205, 230] to capture whether a tweet contains an evidence or not.

RETICENCY FEATURES

Similar to the user category, here also reticency features refer to the willingness of the account holders to reveal their information; however, here the features are slightly different than the one in the user category. These features are about the tweets not the account holders of those tweets. Thus, we check whether account holders leave any trace about themselves in their tweets. The first feature is *geotagged tweet* which shows whether a geographical location is assigned to a tweet or not. The other feature is *hashtag count* [223, 214]. Using a hashtag in a tweet is a clear signal of willingness to be part of a particular movement in Twitter. Twitter users can pinpoint those who are engaged in different movements by just following or searching the hashtags.

6.5. CLASSIFICATION

After data preparation and feature extraction, this section tends to explain OCC based solution for the computational rumour detection. We first argue why OCC is suitable approach for the rumour detection, then we elaborate on OCC idea and explain seven algorithms with this approach. Finally, the experiments are discussed and their respective results are reported.

6.5.1. PROBLEM STATEMENT

In Section 6.2, we explained the binary classification as the dominant approach for computational rumour detection in the literature. However, the binary classification suffers from a major drawback in detection of the rumours. In this section, we discuss the drawback and the reason of why binary classifiers are not suitable for identifying rumours. To this end, we first discuss the concept of non-rumour and provide some evidence regarding different ways of defining non-rumour in the literature. Then, in the second step, we argue how this controversial concept makes the binary classification unreliable and inconsistent for detecting rumours in real-world situations.

THE CONCEPT OF NON-RUMOUR IN COMPUTATIONAL RUMOUR DETECTION

As explained in Section 6.2, we train the classifier with features obtained from rumour and non-rumour tweets. But the question that remains here is “how do the data-points take rumour/non-rumour labels?”. It is an important question since the way of annotating data points defines each class in the model.

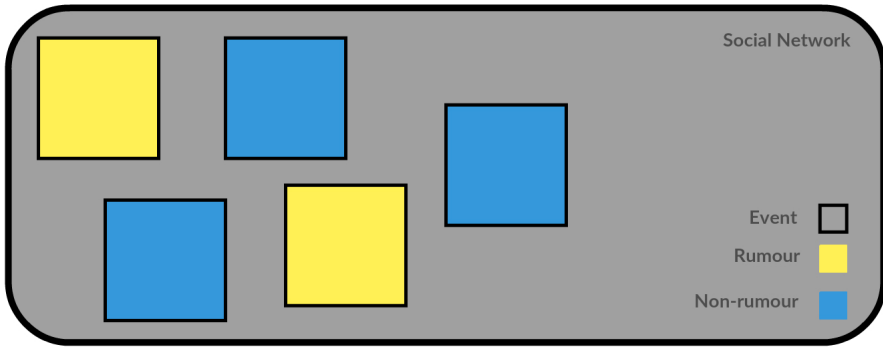
For the rumours, researchers most often refer to definitions with similar elements. That is due to the years of research by rumour scholars which ultimately led to the relative convergence on some aspects of this field such as rumour conceptualisation. This allows researchers to annotate rumour related tweets consistently. But what about non-rumour? What is exactly the non-rumour? Are rumour and non-rumour complementary concepts, in a way that by specifying one, the other will be automatically specified? Or, they are not complementary, and there are other sorts of tweets that fall into neither rumour nor non-rumour category? These are non-trivial questions which to the best of our knowledge have not been addressed in the literature yet.

Non-rumour is an ambiguous term that is coined mostly by computer scientists who used the binary classification for detecting rumours. From a historical point of view, this term has been mentioned two times in the non-computational part of the rumour literature [235, 236], but it has not been defined in any of them. Similarly, non-rumour has been used frequently in the literature of computational rumour detection; however, it has not always been portrayed in the same way. In other word, there is no consensus between the researchers about the conceptualisation of non-rumour, which would engender in different ways of annotating the non-rumour tweets.

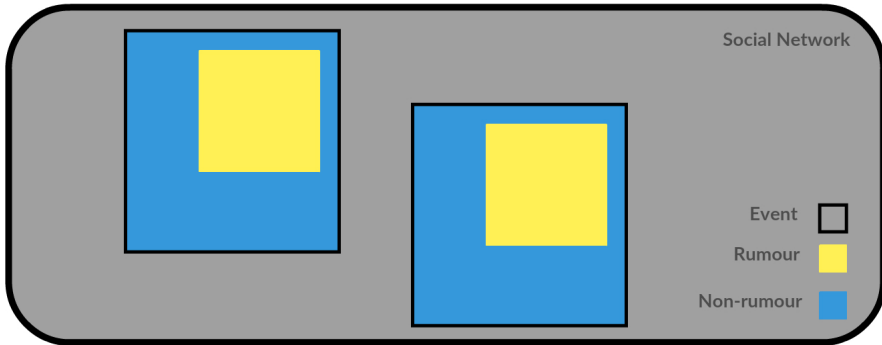
We investigated two primary and widely-known datasets of Kwonset [202] and Zubiagaset [208] in the literature of the computational rumour detection. They do not define non-rumour in the same way. Kwonset takes non-rumour as news items which are extracted from credible news sources, while Zubiagaset treats non-rumours as relevant tweets that cannot take the rumour label. Figure 6.7 represents the conceptualisation of rumour and non-rumour in these two datasets in a schematic way. In Figure 6.7a, which displays the idea of Kwonset about rumour, the big grey rectangular area demonstrates the social network space. Each square in this area illustrates a distinct event. The ones with yellow colour correspond to rumour worthy tweets, while the blue squares show tweets related to the reliable news. On the other hand, Figure 6.7b demonstrates the approach of Zubiagaset. Similar to the Kwonset, the big grey area is the social network space, and each square with black border shows the relevant tweets to a particular event. Unlike the approach of Kwonset in which an event corresponds to either rumour or non-rumour, here an event is a mixture of rumour and non-rumours tweets. The yellow part of the squares shows the rumour relevant tweets, and the remaining blue area demonstrates the non-rumour tweets.

BINARY CLASSIFICATION AND COMPUTATIONAL RUMOUR DETECTION

In this section we argue how various interpretations of non-rumour can impact the quality of the rumour detection systems. To answer this question, we need to inspect the multi-class classification and specifically binary classification more closely, to understand how they classify their input into pre-defined classes. To this end, we use the case of compute making factory, to show how unexpected input can violate the consistency of the classification system.



(a) Non-rumour as fact [202]. Five events are depicted in this diagram; two rumour worthy and three non-rumour worthy.



(b) Non-rumour as any piece of information that cannot take rumour label. [208]. Two events are depicted in this diagram. In each event, the yellow and blue area show the rumour and non-rumour worthy part in each event.

Figure 6.7: Schematic description of two primary perspectives toward non-rumour. In both diagrams, squares with border show different events. Also, yellow and blue are denote rumour and non-rumour area respectively. In this figure, size does not mean anything and cannot be a basis for comparison.

We assume a compote making factory produces five types of fruit compotes based on five different fruits (there is a one-to-one association between fruits and compotes). Because each type of compote is prepared in different parts of the factory, first of all, fruits must be categorised based on their type (at the beginning, the fruits are mixed). In fact, this machine is a multi-class classifier which is trained with samples from five types of fruits. We assume any fruit enters to this machine belong to one of the five pre-determined fruit types, but what would happen if this condition is violated? What would happen if once ten types, once five types, and once eight types of fruits enter to the machine? What would happen if new types of fruits (which are different from the ones that are used for the training) enter to the machine?

When a model is trained for k -classes and a new data-point without belonging to any of these k -classes is given to the classifier, the data-point will be definitely classified into one of the k -classes. This is a case of false positive since that data-point gets a wrong label. This is exactly similar to the case of the computational rumour detection using binary classification techniques. If we build a binary classification model for the

rumour detection (according to any of the existing definitions for non-rumour), there might be data-points coming to our system without belonging to the rumour or non-rumour classes (based on the definition we used). Hence, if we build a rumour detection system to identify rumours in the real world¹⁷, (depending on the definition of non-rumour we use to train the classifier) new data-points which do not belong to rumour or non-rumour classes may increase the number of false positives.

The lack of consensus on the meaning of non-rumour causes every researcher to come up with his/her definition. This diversity of definitions would create two major problems; first of all, the rumour detection systems become inconsistent and unreliable as we cannot be sure about their functions and what they separate. Second, we cannot compare the outcome of different models as they measure different things. Figure 6.8 summarises the chain of reasoning behind the problematic consequences of non-rumour in the binary classification.

Due to these difficulties, questions, and ambiguities of the binary approach to the rumour detection, we are thinking of the rumour detection with rumour data only. Such an approach can address both problems that we raised before. It makes the classifiers reliable and effective by flagging the rumour related information as rumour. From the data collection perspective, we can annotate rumour relevant data points easily, while the same task can be extremely controversial for the non-rumours. Therefore, in a dataset, the annotation of rumours is a feasible task while it is quite challenging for non-rumours. But how could we detect rumours without having and representing non-rumours? The answer is with an approach called one-class classification (OCC) [212].

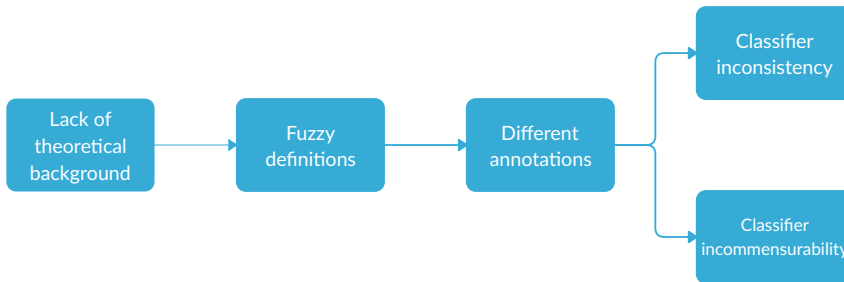


Figure 6.8: Chain of the reasoning behind the problematic consequences of non-rumour in binary classification. It starts with a lack of sufficient theoretical background for the concept of non-rumour. It leads to the emergence of ambiguous and contradictory definitions of non-rumour. Lack of clear definitions causes data annotation to be done arbitrarily, which makes the rumour classifier unreliable (it is not clear, what it separates) and incomparable (it is not possible to compare the results of different classifiers).

One-class classification is a novel machine learning approach which can address such a problem as it works based on the recognition of one class only and classifying it as a target class while annotating all the other data-points as an outlier class. Table 6.4

¹⁷It means, in a non-experimental setting

shows the differences between multi-class and one-class approaches. The second and fourth columns demonstrate different aspects of multi-class and one-class classification, respectively. The third column shows two situations we use one-class classification: when the dataset is imbalanced and when the number of classes is unknown. In the next section, we discuss the one-class classification in more details.

Table 6.4: Comparison between multi-class classification and one-class classification.

	Multi-class Classification	Transition	One-class Classification
Number of classes	$n \geq 2$	Lack of either of the following conditions suffices for transition from multi-class to one-class classification: <ul style="list-style-type: none"> • well-represented classes • well-identified classes 	2
Training dataset	A dataset comprising balanced samples of n classes		A dataset comprising only one class that we know
Test dataset	A dataset comprising balanced samples of n classes		A dataset comprising both classes
Training	Train the model with n classes		Train the model with one class
Test	Testing the model with n classes		Testing the model with two classes
Performance	Reporting the model performance via confusion matrix		Reporting the model performance via confusion matrix

6.5.2. ONE-CLASS CLASSIFICATION APPROACH

The binary classification is the dominant strategy in the realm of pattern recognition and machine learning. Binary classifiers try to learn a function which is able to discriminate the samples of two given classes. To compute such a function based on the given samples, a plethora of methods exist, each of which has an underlying idea and is predicated on some presumptions. The classification problem compounds for the cases where we have the samples of one of the classes only, i.e., one-class classification problem. Several techniques have been proposed to solve one-class classification problem. Tax [212] classifies one-class methods into three categories of (i) the density estimation, (ii) the boundary methods, and (iii) the reconstruction methods. Figure 6.9 provides an overview of the categories and their corresponding methods. In the following, several one-class classifiers for each of these categories are discussed.

DENSITY ESTIMATION

In this approach, the underlying idea is that target samples follow a distribution, which needs to be estimated in the training phase. The most prevalent distribution, which is used for the density estimation is the multivariate Gaussian distribution. In this model,

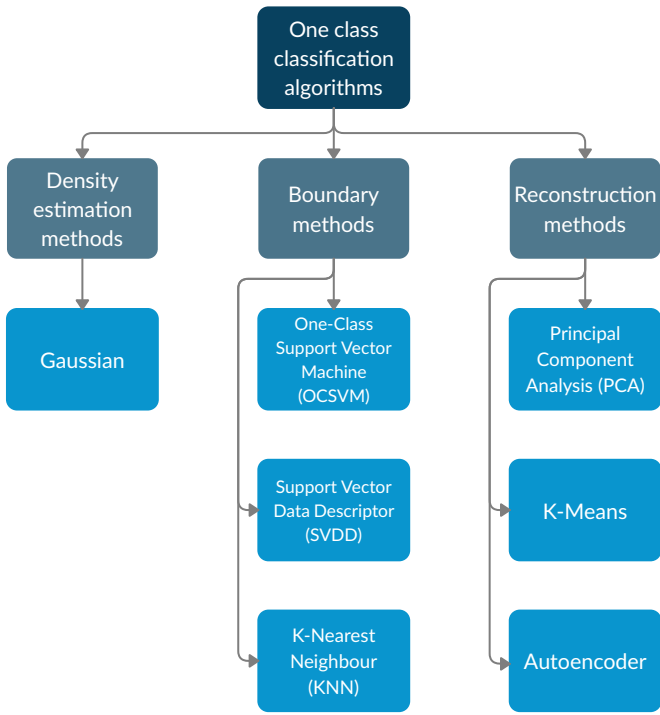


Figure 6.9: Categorisation of one-class classification algorithms.

it is assumed that each training data $x \in R^d$ is a sample of a multivariate Gaussian distribution, i.e.,

$$p_N(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right\}, \quad (6.1)$$

where $\mu \in R^d$ is the mean and $\Sigma \in R^{d \times d}$ is the covariance matrix. Thus, the training of this method entails the estimation of the mean and the covariance matrix. The maximum likelihood estimation (MLE) can swiftly estimate these parameters; thus, the one-class classification based on the density estimation is usually faster compared to other methods. Having estimated the distribution of the target samples, the probability that a test sample z belongs to the target class can be simply computed by the likelihood of z with respect to the estimated distribution. If the computed likelihood is less than a threshold, then the sample z is said to be an outlier; otherwise, it belongs to the target class.

BOUNDARY METHODS

Due to limited data availability in some cases, the density estimation does not provide a comprehensive overview of the data and the result of the consequent one-class classifier is thus not acceptable. To tackle this problem, boundary methods are proposed which try to optimise a closed boundary around the target class. In this chapter, we consider three algorithms with this perspective. The two widely-used one-class classifiers are the one-class SVM (OCSVM) [237] and support vector data description (SVDD) [238]. Given a set of training samples $\{x_i\}_{i=1}^n$, $x_i \in R^d$, the goal of these methods is to specify a region for the target samples. The other well-known classifier with a different perspective is based on k -nearest neighbours, which decides if a sample belongs to the target class based on its distance to k -nearest data-points in the training set, in contrast to SVDD and OCSVM that define a region for the target class. Thus, this classifier does not assume a fixed boundary for the target class, and the size of the boundary is flexible and reliant on the nearest neighbours of a test sample.

One-class Support Vector Machine The OCSVM aims to specify a function which takes a positive value for a small region that the training data live, and take -1 elsewhere. This is done by maximising the distance of the desired hyperplane to origin. The optimisation problem is

$$\begin{aligned} \min_{w, \psi, \rho} \quad & \frac{1}{2} \|w\|_2^2 + \frac{1}{\nu n} \sum_i \xi_i - \rho \\ \text{s.t.} \quad & w^T \phi(x_i) \geq \rho - \xi_i, \quad i = 1, \dots, n \\ & \xi_i \geq 0 \end{aligned} \quad (6.2)$$

where $\phi(\cdot)$ is the high-dimensional space to which data are mapped and ν is the prior probability for the fraction of outliers. The OCSVM decision function can be computed using the kernel trick as

$$f(x) = \text{sign}(w^T \phi(x) - \rho) = \text{sign}\left(\sum_i \alpha_i K(x, x_i) - \rho\right), \quad (6.3)$$

where x is a test sample, $K(\cdot, \cdot)$ is the kernel function used for the training, sign is the sign function, and α is the Lagrangian multiplier of problem (6.2).

Support Vector Data Description This method also seeks to estimate a region for the target class. In contrast to the OCSVM, the SVDD has a distinct approach to computing the desired region for the one-class classification. The SVDD describes the region of training samples as a hypersphere that is characterised by a centre a and a radius R in the feature space. The corresponding minimisation is

$$\begin{aligned} \min_{R, a, \xi} \quad & R^2 + C \sum_i \xi_i \\ \text{s.t.} \quad & \|\phi(x_i) - a\| \leq R^2 + \xi \\ & \xi \geq 0 \end{aligned} \quad (6.4)$$

where C is the regularisation parameter. Given a new test sample z , it is from the desired class if its distance from the center is less than R .

K-nearest Neighbours (KNN) One of the other techniques in the boundary-method class K-nearest neighbour (KNN). Unlike other methods in this class, KNN does not seek a region for the target class. Instead, it classifies a test sample based on its distance to its nearest neighbours. In other words, this method does not assume a fixed region for the target class. Rather, the region is flexible and reliant on the nearest neighbours of the test sample. This algorithm has no training algorithm, and the decision for the test sample z is directly obtained based on the training samples. The underlying notion behind one-class k-nearest neighbour is the ratio between the distance of the sample z from its k neighbours and the distance of its k -neighbours from their nearest neighbours. In order words, assume that $q_i, i = 1, \dots, k$ are the k -nearest neighbours of the test sample z , and $\hat{q}_i, i = 1, \dots, k$ are the nearest target sample to q_i from the training data. The decision function $\rho(\cdot)$ of this method for the test sample z is

$$\rho(z) = \frac{\sum_{i=1}^k \|z - q_i\|}{\sum_{i=1}^k \|q_i - \hat{q}_i\|}. \quad (6.5)$$

If the value of $\rho(z)$ is less than a predefined threshold, then z is deemed to belong to the target class. Although the method has no training, the computation of (6.5) is time- and memory-costly since we need to hold all the samples in memory, and also, we have to compute the distance of a test sample to other points to get the k -nearest neighbours.

RECONSTRUCTION METHODS

The underlying idea of this class of methods is that the target-related test samples must be properly reconstructed from the training samples at hand. The proper construction is typically measured by the *construction error*, which is the distance between a test sample z with its reconstructed point \hat{z} . The difference between various methods of this class is basically their difference in constructing test samples based on the training data. The decision is also made simple: If the construction error of a test sample is less than a threshold, which is determined beforehand, then it belongs to the target class, and it is otherwise outlier. In the following, three well-known reconstruction-based one-class classifiers are discussed.

Principal Component Analysis (PCA) PCA is a data-transformation technique which can be applied to cases where data lie on a linear subspace. The computation for the transformed data is based on the eigenvalue decomposition of the covariance matrix: The top \hat{d} eigenvectors pertaining to the top \hat{d} eigenvalues span a lower-dimensional space that best represents the overall data.

For one-class classification, we transformed the target samples into a lower-dimensional space by using PCA. Thus, the training $X \in R^{d \times n}$ is transformed into $\hat{X} \in R^{\hat{d} \times n}$ where $\hat{d} \ll d$. For a test sample z , we need to find its projection \hat{z} into the transformed space

\hat{X} and compute the reconstruction error by finding the distance between z and \hat{z} . The projection of z onto the subspace is simply as,

$$\hat{z} = \hat{X}(\hat{X}^T \hat{X})^{-1} \hat{X}^T z = \hat{X}^T \hat{X} z,$$

therefore, the reconstruction error for the test sample z is computed as

$$\epsilon(z) = \|z - \hat{z}\|^2 = \|z - \hat{X}^T \hat{X} z\|^2,$$

where $\epsilon(z)$ is the reconstruction error for the test sample z . If this error is less than a threshold, the sample belongs to the target class.

K-means K-means is one of the first techniques in unsupervised learning, which aims to separate the input data into a predefined number of clusters, i.e., k . Each cluster is represented by its center; therefore, the outcome of k-means is k centers of clusters. For one-class classification, we first cluster training data into k groups and represent the center of each cluster by $\mu_i, i = 1, 2, \dots, k$. Having k centers from k-means, the distance of a test sample z from each center is computed. If the distance of the sample z to *only one* of k centers is less than a threshold, then z is said to be a target sample. Otherwise, if the distances of the sample z to *all* cluster centers are larger than the given threshold, it is then an outlier.

6

Autoencoder An autoencoder is a neural network that is well-known to learn the data representation. The autoencoder aims to reproduce the pattern at the input layer in the output layer. Therefore, the objective function for training this neural network is the distance between the input and output layers. For the one-class classification, the training samples of the target class are subjected to the autoencoder so that the weights of the neural network are computed. Then, for a test sample z , the output of the trained network for the input z is deemed as its reconstructed point \hat{z} , i.e., $\hat{z} = f_{ae}(z)$, where $f_{ae}(z)$ is the output of the trained autoencoder for the input z . The reconstruction error for the test sample z is then simply computed as

$$\epsilon(z) = \|z - f_{ae}(z)\|^2.$$

Similar to other methods in this class, if the reconstruction error is less than a predefined threshold, then the sample belongs to the target class; otherwise, it is an outlier.

6.5.3. EXPERIMENTS

In this section, we first explain the details of the experimental setup, then we report the results of our experiments and interpret them from three different perspectives.

EXPERIMENTAL SETUP

As we discussed in Section 6.5.1, designing an experiment for the binary classification and one-class classification are quite similar. Both need feature extraction, training and test datasets, and performance score to evaluate them. For the training and test sets, in this chapter, we use Zubiagaset and Kwonset.

After preparing the datasets, the next step is feature extraction in which every tweet is represented by 86 features. Then, the one-class classifier must be trained. This is the only distinct step of one-class classification compared to the binary classification. Unlike the binary classification, which needs both classes for the training, we train the one-class classifier with one class only. In the rumour detection problem, this means that training a binary classifier needs both rumour and non-rumour while one-class classifier is trained only by rumour. To test the performance of one-class classifiers, we follow the same evaluation approach as the binary classification and test the classifier with both rumour and non-rumour (Table 6.5).

Table 6.5: Confusion matrix for one-class classification [212].

	Object from target class	Object from outlier class
Classified as a target object	True positive, T^+	False positive, F^+
Classified as an outlier object	False negative, F^-	True negative, T^-

To get more reliable results on the datasets, we use k -fold cross-validation. In this technique the rumour dataset is partitioned into k bins of equal size, then we perform k separate learning experiments by picking $k - 1$ folds of rumour dataset for the training and one remaining fold along with non-rumour class for the test in each experiment. In the end, the average performance of the k experiments is reported as the performance of the model. In this work, we repeat k -fold cross-validation for $k = 3, 5, 10$ to show the sensitivity of the models' performance to the training sample size.

For the classifier selection, we consider seven classifiers belonging to three one-class classification paradigms, including Gaussian classifier as a density estimation method, one-class support vector machine (OCSVM), support vector data descriptor (SVDD), and k -nearest neighbours (KNN) as boundary methods, and K-means, principal component analysis (PCA), and autoencoder as reconstruction methods. For parameters tuning and sensitivity analysis over the model hyper-parameters, we use grid search technique and measure the models' performance regarding the different combination of hyper-parameters. In kernel-based methods, namely SVDD and OCSVM, we used the radial basis function (RBF) as it was suggested and used in many of rumour detection works [206]. To apply the selected algorithms, we used MATLAB and Python. The OCSVM is implemented in scikit-learn (a python machine learning library) [239], and the rest of the algorithms come with MATLAB PRTools [240] package.

From the implementation point of view, in the above programming libraries and packages, the methods that are essentially using kernel or distance matrix are not well suited for relatively large datasets. We tried to tackle this issue by establishing a powerful computer system to perform the experiments; however, it could not manage to run SVDD over Kwonset. Therefore we decided to perform SVDD experiments by subsampling the Kwonset.

We report the model performance via precision, recall, and F1-score. Precision is the fraction of correctly retrieved instances that are relevant, while recall is the fraction of relevant documents that are retrieved. F1-score is the harmonic mean of precision and recall [241]. We also assess the models' efficiency by measuring the execution time of the experiments in both datasets.

ONE-CLASS CLASSIFICATION RESULTS AND DISCUSSION

In this section, we report and discuss the results of the experiments. To this end, we first make a baseline analysis by comparing the results of the experiments with the baseline of each dataset. Then we measure the impact of training sample size on the models' performance. After that we evaluate the impact of hyper-parameters in each model. Then we report the models' performance in different feature categories, and in the end, we assess the models' execution time in each dataset.

Baseline Analysis In this subsection, we study the performance of one-class classifiers in comparison with baselines in both operational datasets. In the first baseline, Zubiaga et al. [204] propose a rumour identification classifier using conditional random field (CRF) with 5-fold cross-validation, and in the second one, Kwon et al. [202] apply a random forest with 3-fold cross-validation. Both baselines use F1-score, precision, and recall to report classifiers performance. For the baseline analysis, we replicate the same experimental setup as the original studies, therefore for the Zubiagaset and Kwonset we perform the experiments in 5-fold and 3-fold cross-validation, respectively. We also report the classifiers performance with the same metrics as the original studies.

Table 6.6: Baseline analysis on the Zubiagaset and Kwonset [204, 202]. We could not apply SVDD on the whole Kwonset since the standard solver of SVDD does not suit the large-scale datasets. We tackled this problem by subsampling the training set and experiment with a subset of the original dataset.

	Zubiagaset			Kwonset		
	PR	RE	F1	PR	RE	F1
Autoencoder	48.61%	77.90%	59.86%	97.31%	90.14%	93.59%
Gaussian	38.33%	87.80%	53.36%	95.69%	90.03%	92.77%
K-means	53.82%	81.20%	64.73%	96.11%	90.08%	93.00%
KNN	69.20%	80.20%	74.30%	98.19%	90.11%	93.98%
SVDD	51.78%	82.20%	63.54%	95.20%	91.21%	93.16%
OCSVM	13.08%	51.30%	20.85%	95.99%	88.24%	91.95%
PCA	41.38%	90.20%	56.73%	96.99%	90.03%	93.38%
Baseline	66.7%	55.6%	60.7%	89%	90%	89.5%

Table 6.6 demonstrates the precision, recall, and F1-score of classifiers in both datasets along with the results of baselines. Based on this table, in the Zubiagaset, KNN, *k*-means, and SVDD outperform baseline with the F1-score of 74.30%, 64.73%, and 63.54% respectively. In the Kwonset, all seven one-class classifiers achieve better F1-score than the baseline and thus outperform it.

In terms of precision and recall, in the Zubiagaset baseline, precision is better than recall. This is another way around for the one-class classifiers on the same dataset, which means recall is higher than precision in one-class classifiers. On the other hand in the Kwonset, baseline precision and recall are almost equal, while precision is slightly higher than recall in one-class classifiers.

Compared to baseline, in the Zubiagaset, one-class classifiers can identify more rumours but with less precision. Therefore it is highly likely to identify the bigger fraction of rumours as well as mislabelling the bigger fraction of non-rumours as rumour when

we use one-class classifiers in the Zubiagaset. Repeating the same comparison in the Kwonset results in the same number of identified rumours by one-class classifiers but with higher precision. This means, it is highly likely to identify the same fraction of rumours but with less mistaken flags, when we use one-class classifiers in Kwonset.

We can infer one-class classifiers often outperform baselines' binary classifiers or achieve to their close proximity in spite of the fact that one-class classifiers are trained in the absence of non-rumour samples.

Training sample size impact In this section, we analyse the impact of training sample size on the performance of the classifiers. To this end we measure the performance of different one-class algorithms with k -fold cross-validation when $k \in \{3, 5, 10\}$. Since F1-score combines precision and recall, we use it to represent classifiers performance. Figure 6.10 displays how changing the training sample size affect classifiers performance. It consists of two subfigures which represent the performance changes in each dataset correspondingly.

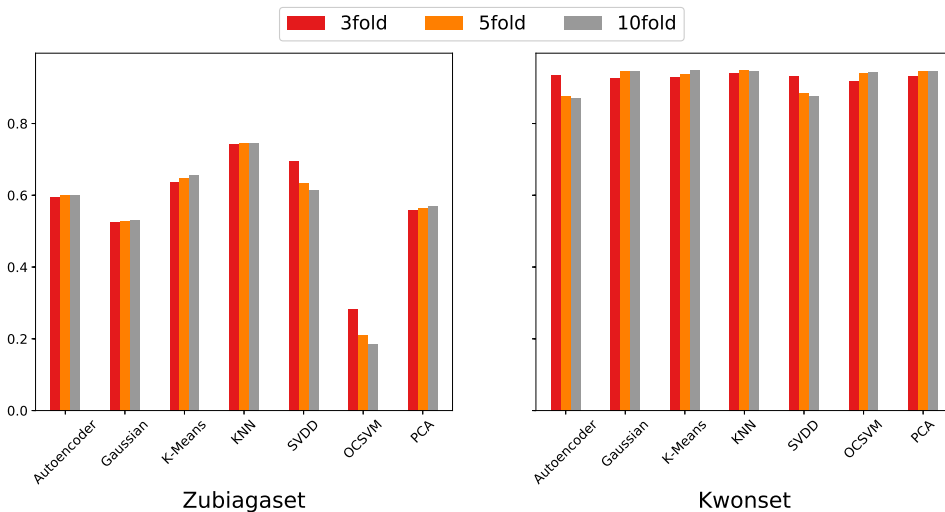


Figure 6.10: The impact of training sample size on the performance of classifiers in the Zubiagaset and Kwonset. The horizontal axis displays different classifiers, and the vertical axis shows their F1 score.

By comparing classifiers performance with different cross-validation, we can observe two different patterns: performance gain and performance loss by the growth of the training sample size. In the Zubiagaset, except SVDD and OCSVM, the other classifiers performance improve as the training sample size increases. In the Kwonset, all the classifiers but autoencoder and SVDD experience performance enhancement when the training sample size grows.

Despite the heterogeneous impact of training sample size on the classifiers performance, the difference between the highest and lowest performance in 10 out of 14 classifiers is less than 2%. This means our models show a high level of robustness against training sample size alteration.

Hyper-parameters Impact In this section, we discuss the sensitivity of the models to the hyper-parameters value. We look at the classifiers hyper-parameters in each of datasets and between them. Table 6.7 summarises the classifiers hyper-parameters and their valid range. Each classifier has two hyper-parameters which we change them within their valid range.

Table 6.7: The classifiers hyper-parameters and their valid range.

Classifiers	Hyper-parameters	Valid range
Autoencoder	FRACREJ: Fraction of target objects rejected N: Number of hidden units	$\text{FRACREJ} \in (0, 1)$ $\{i i \in N, i \leq \text{datapoints}\}$
Gaussian	FRACREJ: Error on the target class R: Regularization parameter	$\text{FRACREJ} \in (0, 1)$ $R \in (0, 1)$
K-means	FRACREJ: Error on the target class K: Number of clusters	$\text{FRACREJ} \in (0, 1)$ $\{i i \in N, i \leq \text{datapoints}\}$
KNN	FRACREJ: Error on the target class K: Number of neighbors	$\text{FRACREJ} \in (0, 1)$ $\{i i \in N, i \leq \text{datapoints}\}$
SVDD	FRACREJ: Error on the target class P: Inverted kernel parameter	$\text{FRACREJ} \in (0, 1)$
OCSVM	v : Regularization parameter γ : Kernel parameter	$v \in (0, 1)$
PCA	FRACREJ: Error on the target class N: Number of PCA components	$\text{FRACREJ} \in (0, 1)$ $\{i i \in N, i \leq \text{features}_n \text{ number}\}$

Figure 6.11 demonstrates the classifiers F1-score regarding different combination of hyper-parameters. For each classifier, two heatmaps are used to represent the performance space in both Zubiagaset and Kwonset. In the heatmaps, darker colours represent higher performance.

As Figure 6.11 illustrates, for each classifier (in both datasets) the best performance is achieved when the target-class error is in the vicinity of its lowest value. Some of the classifiers such as Gaussian and KNN are indifferent to the second hyper-parameter, which for instance in the case of KNN, it does not matter how many neighbours are selected; however, for some others, it is not the case and the best performance depends on both hyper-parameters. For example, in OCSVM, the best performance is achieved when the kernel parameter has a small value or autoencoder perform better when the number of hidden unit is higher. By inter-dataset analysis, we can see classifiers pursue similar hyper-parameters pattern in both datasets. This means the high- and low-level of performance for a classifier are achieved in similar areas in hyper-parameter space. We have also observed SVDD is almost indifferent to its hyper-parameters and perform similarly in different combinations of hyper-parameters.

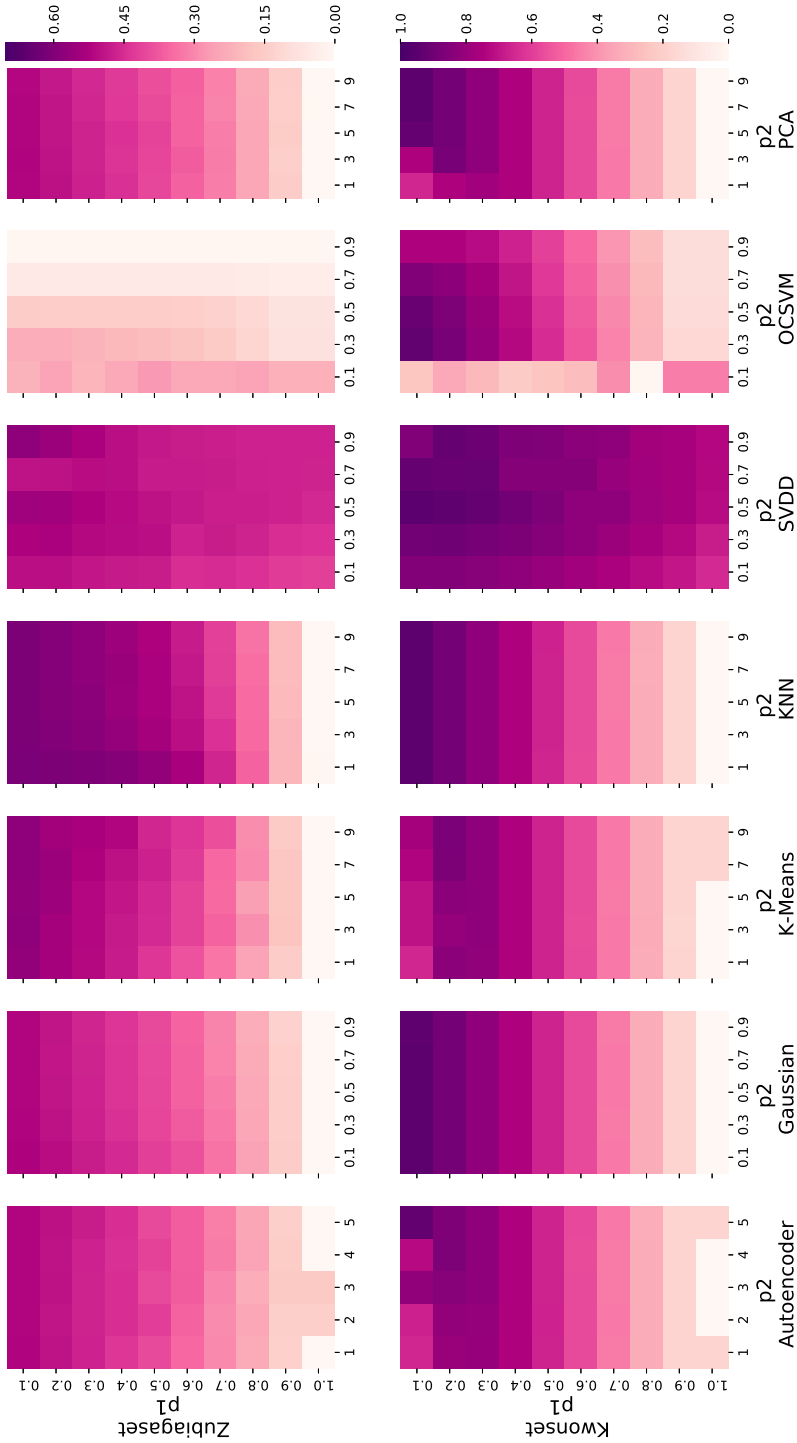


Figure 6.11: The impact of hyper-parameters on models performance in the Zubiagaset and Kwonset.

Classifiers Performance In this section, we report and discuss the results of the experiments in different feature categories. Figure 6.12 illustrates the performance metrics of the one-class classifiers on two datasets for different feature categories. This figure consists of two panels representing the performance scores in different datasets. Each panel is also composed of three parts which display the performance of different feature categories in terms of F1-score, precision, and recall. We first report classifiers performance in the Zubiagaset where all features are considered. Then we go through Kwonset and report one-class classifiers performance when all features are present. Finally, we investigate the synergy between feature categories in both datasets.

As the left panel of Figure 6.12 shows, among the one-class classifiers which are trained on the Zubiagaset with full features, KNN has the highest F1-score. It also has the highest precision and lowest recall among the same group of classifiers. SVDD has the second-highest F1-score and precision, while it outperforms other classifiers. Although SVDD has the best recall and second best precision, KNN high precision compensates its low score in recall and gives it the highest F1-score. After KNN and SVDD, the next one-class classifiers with highest F1-score are K-means, autoencoder, PCA, and Gaussian, respectively. The lowest performance belongs to OCSVM which delivers the poor F1-score of 28% by 81% recall and 17% precision.

The experiments on the Kwonset is also displayed in the right panel of Figure 6.12. In this set of experiments we could not apply SVDD on the whole Kwonset since the standard solver of SVDD does not suit the large-scale datasets. We tackled this problem by subsampling the training set and conduct the experiment on a subset of the original dataset. For the rest of the classifiers, despite the long execution time, the experiments were finished and produced expected outcomes. In the Kownset all full features trained classifiers achieve a high level of F1-score which is obtained by a high level of precision and recall. The performance of one-class classifiers is very close to each other, in such a way that the difference between maximum and minimum of F1-score, precision, and recall is less than 1%. The high level of both precision and recall mean that one-class classifiers could identify most of the rumours with high level of precision.

Last but not least, by considering Figure 6.12, one can simply realize that the synergy of all features from the three categories is positive based on F1-score. In particular, the F1-score of the autoencoder, Gaussian, K-means, KNN, SVDD, OCSVM and PCA based on all features is superior to the F1-score obtained by training with individual feature categories only. However this superiority is more significant in some classifiers, for instance, SVDD in the Zubiagaset or Gaussian in Kwonset. In some of the experiments, such as the KNN trained by the linguistic features on the Zubiagaset, recall is higher compared to the training with all features, but its precision is also lower which results in the overall lower F1-score.

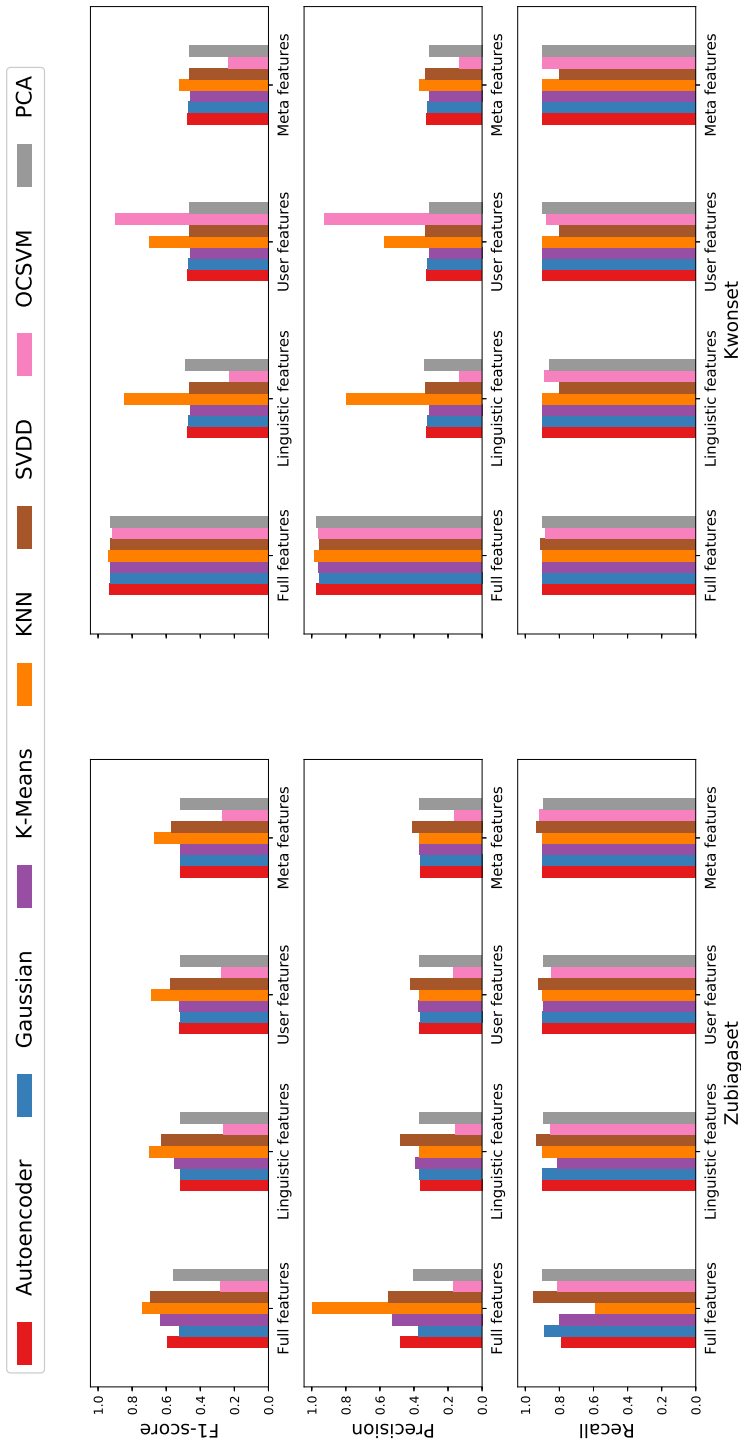


Figure 6.12: The classifiers performance in different feature categories in the Zubiagaset and Kwonset.

Classification Speed The other metric to assess and compare the classifiers is their speed. To measure the classification speed, we gauge the execution time of classification which means the average time of training and test for one iteration of k -fold cross-validation. Figure 6.13 displays the execution time of classifiers across the Kwonset and Zubiagaset. We use a log-log diagram, owing to the substantial difference between execution times.

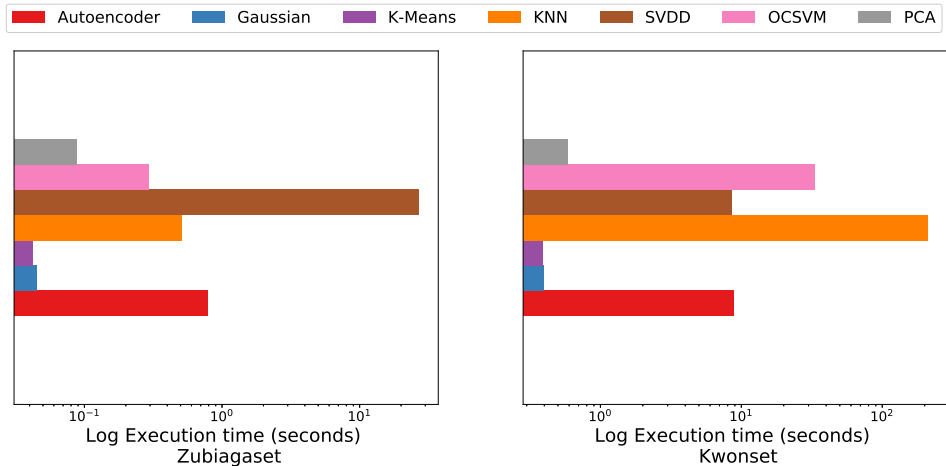


Figure 6.13: The execution time of classifiers in the Zubiagaset and Kwonset.

In both datasets, Gaussian, K-means, and PCA are the fastest classifiers. The execution time of the other classifiers varies in both datasets. In the Zubiagaset, SVDD is the slowest model, while in Kwonset, KNN has the longest execution time. It is worth mentioning that, due to the scalability problem in MATLAB SVDD package, we use a subset of Kwonset for SVDD. In the Zubiagaset, after SVDD, autoencoder, KNN, and OCSVM are the slowest classifiers respectively. In Kwonset, the next three classifiers with the longest execution time are OCSVM, autoencoder, and SVDD.

6.6. CONCLUSION

In this chapter, we studied the binary rumour classification pitfall by addressing the long-standing and unnoticed concept of "non-rumour". Non-rumour is a term coined by computer scientists when they formulated rumour detection as a binary classification problem. There is neither a clear definition nor a consensus among scholars for this pseudo-concept. Some studies imply that rumour and non-rumour are complementary concepts, while in some other work, non-rumour is considered as factual information. This ambiguity and lack of consensus about the definition of non-rumour prevents us from making a comparison between different classifiers as they do not annotate tweets consistently. It also makes the classifiers unreliable as there is no consensus regarding the definition of the non-rumour. We also do not know which of the existing definitions is the correct one. Based on the stated flaws, we reached to the conclusion that the bi-

nary classification with the current approach to the non-rumour might not be beneficial to the computational rumour detection.

To tackle the issue addressed above and avoid dealing with non-rumour, we adopt a novel classification approach called one-class classification. This approach goes very well with the special characteristics of our problem as the classifier is trained by one class only. Hence, it provides us with an opportunity to have a classifier for detecting rumour without touching the controversial area of non-rumour. For the feature extraction, we take two principles into account, first, to focus on early available features due to the importance of early detection of the rumours, and second, to avoid features with high degree of volatility.

To evaluate the quality of the proposed approach, we trained seven one-class classifiers on two major datasets and compared their performance. We observed that the one-class classification approach can recognise rumours with a high level of F1-score. In the Zubiagaset, this approach could achieve F1-score of 74%, and in Kwonset F1-score reaches to 93%. We extended the experiments to different feature categories and analysed the performance of each classifier on individual feature categories. We observed a positive synergy when individual feature categories are aggregated. We also studied the impact of training sample size and hyper-parameters on the classifiers performance. We reported the model performance in different settings using F1-score, precision, and recall. Additionally, to understand the efficiency of the one-class classifiers, we compared their speed by measuring the execution time of each classifier.

We can summarise our findings into a few lessons learnt. SVDD performs very well in terms of precision and recall; however, it is not time and memory efficient. Hence, it is an ideal one-class classifier for small sized problems. KNN performs very well in terms of precision, while its performance subjected to recall is poor. It is also time and memory inefficient but not as bad as SVDD. If the dataset is not too large and precision is the main concern, KNN can be the right choice. The other one-class classifier is autoencoder, which achieves good results with respect to both precision and recall. In contrast to SVDD and KNN, the autoencoder is memory-efficient and can be executed on a desktop computer even for relatively large datasets such as Kwonset. On the other hand, it is time-consuming, especially when it is set up with a high number of hidden units. Therefore, in the case of relatively large datasets, lack of adequate computational resources, and having a considerable amount of time autoencoder can be a proper one-class classifier. Gaussian, K-means, and PCA produce mediocre results with a relatively low level of precision and high level of recall in small datasets; however, they are very fast even in large datasets. Therefore, if the datasets are large, they are ideal choices. One of the other lesson learnt is the context insensitivity in one-class classification. Although Zubiagaset and Kwonset covered very different subject domains, one-class classifiers overall perform well.

7

MODELLING RUMOUR CAMPAIGNS: A PROACTIVE APPROACH

To know your enemy, you must become your enemy.

Sun Tzu

The current paradigm of rumour confrontation has a passive approach which is reactive and after-the-fact. Besides, it focuses on the impacts, rather than the functional mechanisms leading to the rumour spreading in social media. To tackle this issue, the following research question is posed:

- *How could we take preemptive measures regarding rumours in social media?*

To address this question, this chapter proposes a fine-grained model that can capture the underlying mechanisms of rumour campaigns. This model takes a proactive stance against rumours and provides us with an opportunity of developing preemptive measures. To implement this approach, an ontology model for a rumour campaign is developed in an iterative procedure. The model is evaluated by experts opinion and through exemplification on three notoriously famous social manipulation campaigns¹.

¹This chapter is based on the following under review manuscript: Fard, A. E., & Maathuis, C. (2020). Capturing the Underlying Offensive Mechanisms of Social Manipulation: A Data Model Approach. Applied Ontology (Under Review).

7.1. INTRODUCTION

Sometimes rumour spreading becomes a strategic tool for the nefarious self-serving motives [7]. In wartime, rumours are intentionally spread to demoralise the enemies troops and nations. During the elections, rumours are deliberately used to discredit the candidates. Business competitors also use rumours to drive customers away from rival products to their own. Deliberate rumour spreading might also be a coping strategy to manage psychological threats² or to boost self-esteem by derogating the outgroups [7, 6, 14].

The emergence and drastic growth of social media platforms over the past couple of years have scaled up self-gaining rumours [25]. Those platforms facilitate inception and amplification of messages due to their unique characteristics that we discussed in Chapter 3. Therefore it is of utmost importance to develop counter-strategies to confront such a malevolent phenomenon; otherwise, it would influence political, economic, and social well-being.

As we discussed in Chapter 4, plenty of approaches have been proposed so far to limit the spread of rumours; however, they mostly focus on the understanding, prevention or reduction of the impacts rather than the discovery of the underlying functional mechanisms leading to those impacts. Besides they react to the rumour campaigns passively and usually after it has made some progress [47, 220, 242, 206]. Even in strategies focusing on the educating people to not being deceived by false information, the forward-looking approach is absent, and the training is based on the historical cases [4].

To tackle the above-mentioned issues, this research aims at proposing an operational-level model that can capture the functional mechanisms of deliberate rumour spreading within social media. We call this model Deliberate Rumour Spreading Model, and we refer to it as DRSM to avoid confusion. Building this model would lay a concrete foundation for proactive rumour confrontation because it identifies the components of a rumour campaign and the relationships between them in an operational level; and thus it provides a framework that could be used for the development of ABM based simulation on top of it. Those simulations could generate future rumour campaigns which allows us to take preemptive measures and develop resistance in advance.

Figure 7.1 illustrates the flow of building DRSM. It is composed of three main steps of (i) model selection, (ii) model building, and (iii) model evaluation. In the first step, the foundation of the model is discussed. More specifically, in the first step, the models that constitute the backbone of DRSM are introduced. In the second step, the model's components are thoroughly explained. In this step, we analyse further related research in order to be able to capture the particular aspects of rumour spreading campaigns. In the third step, we evaluate the model through experts opinion and exemplification. We use the insights gained in this step to enhance model realism, accuracy, clarity, conciseness, and adaptability. The bidirectional arrow between the last two steps refers to several rounds of evaluation and incorporation of the feedback (from the evaluation step) into the model.

The remainder of this chapter is structured as follows. Section 7.2 introduces the context of this study and summarising the relevant models for this research. Section 7.3 describes the skeleton of the DRSM, the operational-level components, and the data

²Psychological threats arise at the time when whatever we cherish such as our identity, values, community, party, or ideology is ridiculed, criticised, derogated, or humbled.

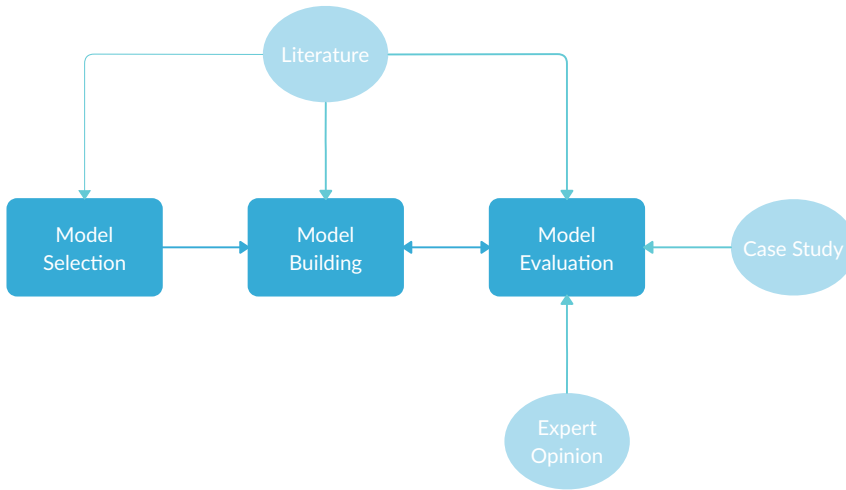


Figure 7.1: The flow of building DRSM. The bidirectional arrows show iteration between the phases.

model. Section 7.4 presents the evaluation process conducted for the proposed model: (i) expert-based evaluation, and (ii) evaluation through exemplification on three different Information Operations: Operation Infektion, Operation Heshmat Alavi and the IRA Copy Pasta Campaign. Finally, the last section concludes this research by discussing the limitations and possible extensions on this study.

7.2. RESEARCH BACKGROUND

This section provides some context about deliberate rumour spreading and then introduces and investigates the misinformation machine model and Maathuis Cyber Operation Model as the baseline for the proposed model in this research.

7.2.1. DELIBERATE RUMOUR SPREADING AS A MEANS OF INFORMATION OPERATION

The development and growth of information technology in the twentieth century extended the areas of confrontation significantly. There is no need for the physical confrontation anymore as it could take place in the information environment in a cheap, covert, and less fatal manner. This way of confrontation which is called information operation (IO) might arise between individuals, groups, organisations, or even countries. IO opens up the possibility for three kinds of confrontation [79]: (i) electronic confrontation, (ii) cyber confrontation, and (iii) deliberate rumour spreading. Electronic confrontation refers to the use of electromagnetic and directed energy to control the electromagnetic spectrum or to attack the enemy. The second kind of IO is called cy-

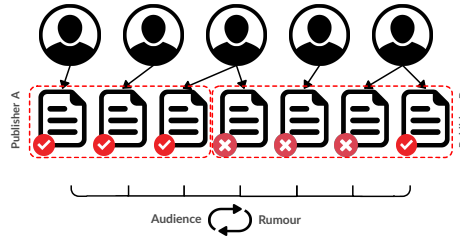


Figure 7.2: The Misinformation machine model.

ber confrontation, meaning the use of malicious computer programs to achieve hostile intents [243, 244]. The third kind of IO conducted in social media is deliberate rumour spreading which refers to the circulation of misleading information to produce harmful social, political, and economic outcomes [79].

The drastic growth of the Internet and social media in the absence of institutional maturity in societies, as well as lack of proper control measures in the infrastructure, opens a window of opportunity for adversarial behaviour by launching rumour campaign. This is an alarming trend that has to be controlled; otherwise, the repercussions might be inevitable [25].

7.2.2. MISINFORMATION MACHINE

This section presents one of the baseline models that this research is built upon it. The misinformation machine [81] is the model developed to explain the dynamic of misleading information in the media sphere. In this model, the process of producing false information is composed of five key elements: (i) publishers, (ii) authors, (iii) articles, (iv) audience, and (v) rumours (Figure 7.2). The publishers provide a broad range of distribution media from the ones with minimum codes of conduct and guideline such as entirely informal websites (e.g. blogs and content mills for clickbait) and social media platforms to the publications and news outlets with long and strong editorial histories and high journalistic standards. The authors are content providers for publishers. Sometimes they cross-post their articles in multiple media such as the original publisher, their blogs, and their social media pages. The articles are the messages produced by the authors and published in different mediums. The audience is those who primarily interact with the articles. They discuss the articles by circulating their impressions, interpretations, or reactions among their social networks. In this model, the discussion among the audience is broadly called rumourmongering.

There are two key moments in this dynamic leading to the spread of false information. The first one is called the article-audience interaction, and it happens when the malicious articles are planted into the media, and the audience are exposed to them. The second key moment is called the audience interaction, and it occurs when the audience discusses the message. Regardless of the kind of discussion that can support, deny, query, or comment the original content [206], it fulfils the main purpose of au-

thors, which is maximum visibility of their message [170, 3].

The misinformation machine model provides us with the first comprehensive model that can capture the dynamic of deliberate rumour spreading from emergence to implantation and circulation. However there are three missing pieces in this model which we highlight in this section. First, one of the key steps in every deliberate rumour is the evaluation [14]. In this step, the operation's success is measured, and the lessons for the future operations are learned. Despite the significance of the assessment phase, it is absent in the misinformation machine model. The second missing piece of misinformation machine model is the iteration. The iterative approach works like the constant dripping of water on a rock. A drip on a rock might not have any impact, but after a long period it would create a hole on the rock. Studies have shown that many of the deliberate rumours work in the same way, and they comprise multiple iterations [60]. Finally, the misinformation machine explains rumour spreading from a broad perspective and does not take the operational level mechanisms into account. Increasing the model resolution would allow us to capture the underlying mechanisms that contribute to the rumour circulation. It would, in fact, provide us with the opportunity of emulating the offensive side of the deliberate rumour spreading. Such a system allows to simulate a variety of offensive operations in different settings and study the impact of each element within the whole operation. This would also enable us to stay one step ahead of adversaries by being vigilant and developing preemptive measures regarding different rumour spreading scenarios.

7.2.3. CYBER OPERATIONS MODEL

This section presents Cyber Operations model developed by Maathhuis et al. [244]. To avoid confusion, we refer to this model as Maathuis Cyber Operations Model (MCOM). The MCOM is a knowledge/data model for Cyber Operations implemented as a computational ontology following a design science approach grounded on extensive technical-military research. This model incorporates the essential entities of Cyber Operations at an operational level and allows to simulate Cyber Operations by populating the model with actual data [244]. Figure 7.3 displays inter-connected components in this model. There are 11 classes in MCOM inherited from the superclass of owl:Thing. They are explained in Table 7.1.

7.3. MODEL DEVELOPMENT

This section explains the development of the proposed model or as we refer to it DRSM. In the previous section, we explained, misinformation machine model and MCOM. We discussed the strengths and shortcomings of each model. The misinformation machine is a coarse-grained model which does not have the assessment module and multi-iteration approach. On the other hand, MCOM is a fine-grained model for Cyber Operations which includes both the assessment module and multi-iteration approach.

The Cyber Operations and deliberate rumour spreading are two different yet very close variations of Information Operations. In both of them, a set of actors initiate an operation against a target group within the information sphere. They both benefit from computers and weaponised form of information to fulfil their plans. Additionally, they

Table 7.1: The explanations of MCOM classes [244].

Class	Description
Context	It comprises the following dimensions: Political, Military, Economic, Informational, Historical, Sociocultural, and other context.
Actor	It refers to distinct types of actors who are either responsible for planning, executing, or assessing Cyber Operations. The actors are the targeted ones or the ones unintentionally impacted by Cyber Operations.
Type	It refers to distinct types of Cyber Operations, specifically offensive, defensive, and intelligence.
MilitaryObjective	It is the military goal that actors want to achieve in Cyber Operations.
Phase	It refers to the phases of Cyber Operations from planning to assessment
Target	It is a military entity (person or object) legally targetable in Cyber Operations.
CyberWeapon	It refers to the means employed in Cyber Operations to achieve military objectives.
Asset	It is either humans or objects unintentionally impacted in Cyber Operations.
Geolocation	It is incorporated geolocation information about targets or assets.
Action	It is the actions and tasks involved or performed in Cyber Operations
Effect	It is the implications and consequences of Cyber Operations. It can break into intended and unintended effects. The intended effects support the achievement of military objectives (Military Advantage) by targets' engagement; and unintended effects do not contribute to the achievement of military objectives, but do still unintentionally impact other assets (for instance Collateral Damage).

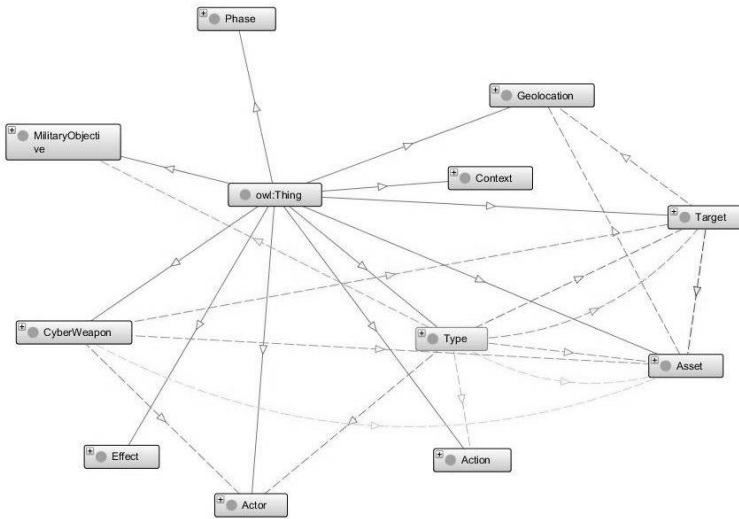


Figure 7.3: Maathuis Cyber Operation Model (MCOM) [244].

plant their seed (computer program in Cyber Operations/ misleading message in the deliberate rumour) into a network (of computers in Cyber Operations/ of people in deliberate rumour spreading).

Due to the high degree of similarity between the deliberate rumour spreading and Cyber Operations as well as the complementary aspects of the models mentioned earlier, we use them as the skeleton of DRSM. The misinformation machine describes the logic of rumour spreading in high level, thus we use it to form the conceptual components and the relationships between them in the DRSM. In this step, we also use MCOM to fill the missing aspects in the misinformation machine. To this end, we add an assessment component to evaluate the effect of the operation on the target. We also incorporate the multi-iteration approach to the model.

For the operationalisation of the model, we mainly use MCOM. Besides, due to the exclusive aspects of deliberate rumour spreading such as the major role of media, we also benefit from the instrumentally relevant studies in this subject domain to incorporate those particular aspects to the model. Additionally, to make this model more realistic, accurate, clear, concise, and adaptable, we do several rounds of evaluations through an interview with the field experts and exemplification of major intentional rumours. We incorporate the insights gained through the evaluation step into model development.

In the next few sections, the above-mentioned steps are explained with more details. In Section 7.3.1, the DRSM architecture and the components are discussed. Section 7.3.2 elaborates on each component and explain it in an operational level. In Section 7.3.3 the data model is presented.

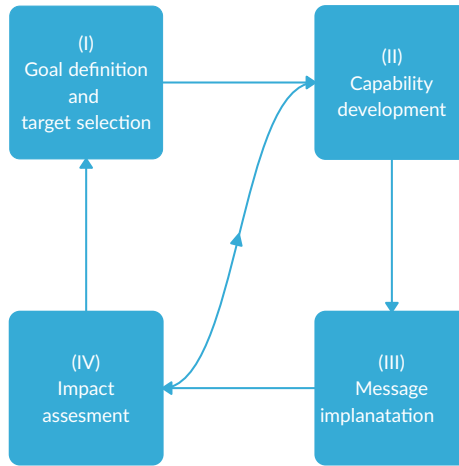


Figure 7.4: The block diagram of DRSM.

7.3.1. THE BLOCK DIAGRAM

In this section, we introduce the components of DRSM using block diagram. The components are inspired by misinformation machine and MCOM. Figure 7.4 displays the block diagram of the model. It starts by goal definition and target selection. In this phase, the preparations before commencing the operation are being made. This means the goals are defined, the involved authors are determined, and the audience is selected. The second step would be the development of the required capabilities to address the target audience and fulfil the intended goals. In this phase, the deceiving technique is chosen, and the message regarding the selected technique is developed. The implanatation of the message into media and the (organic³ or inorganic⁴) message circulation takes place in the third component. The fourth component is the assessment module that evaluates the impact if the deliberate rumour spreading on different actors. After the assessment, the gained insights and knowledge is transferred to the first module for a new operation or to the second module for another iteration of the same operation. In this diagram, the first and the fourth components are based on MCOM, and the misinformation machine inspires the second and third components.

7.3.2. THE OPERATIONALISATION OF THE MODEL

After introducing the components of the DRSM, in this section, we explain how each component is operationalised. As we explained before, due to the similarity between cy-

³By the organic message circulation, we mean the spread of the opinions, impressions, interpretations, or reactions by non-automated accounts and with non-strategic intents

⁴By the inorganic message circulation, we mean the spread of messages by inauthentic accounts (including social bots and trolls) and with strategic intents

ber and deliberate rumour spreading, we mainly use MCOM for the operationalisation. For the exclusive parts of rumour spreading, we benefit from instrumentally relevant literature. For every component, first, the entities are introduced, and then relationships between them are explained.

GOAL DEFINITION AND TARGET SELECTION

This component reflects the moment when a deliberate rumour starts. The three entities of Actor, Aim, and Plan make these components. The first two are equivalent to the Actor and MilitaryObjective in MCOM. They are slightly changed to adapt to the deliberate rumour spreading context. For the Actor⁵, it is an abstract entity which has three types of Offender, Victim⁶, and Unknown⁷. They are the main actors of a deliberate rumour. The Offenders are those who initiate the operation (and sometimes even they appear as an execution group). The Victims are the targets of the deliberate rumour. The Unknowns represent those who are not in the target group but the operation might influence them, so they are collateral actors. The Aim⁸ entity shows the ultimate goal of the Offender. The last element in this component is the Plan which shows how the Aim could be achieved. It is like a roadmap that should be followed throughout an operation. Figure 7.5 represents the entities and their relationships in this component.

CAPABILITY DEVELOPMENT

After defining the goal and selecting the target, it is a time for the preparation of the misleading message. The four entities of Execution Group, Execution Strategy, Information, and Techniques make this component. They are inspired by two entities of CyberWeapon and Action in the MCOM. The Execution Group represents the responsible actors commissioned by Offender to implement the Execution Strategy⁹. The Execution Group is often manifested as social media accounts [25]. They are either known or anonymous. Social media accounts in the latter group might be hacked or stolen, or impersonated. From the execution point of view, the accounts are driven by humans, bots, and cyborgs [55]. The Execution Strategy is the Plan at an operational level. Part of the Execution Strategy refers to a set of Techniques that are often used in deliberate rumour spreading. They are often Propaganda, Conspiracy theory, and Fake news¹⁰; however, there is always the chance of emergence of a new technique which is the reason for having Unknown techniques in this set. Besides, sometimes those techniques are used in combination. In the following, each of those techniques is explained.

Another important entity in the component of capability development is Information which represents the knowledge base behind the operation and comprises all the information that are instrumentally relevant to the elements of deliberate rumour spreading including the target group, context, and goal. Information is mainly used for the

⁵Actor is equivalent to Actor in MCOM.

⁶Victim is equivalent to Target in MCOM.

⁷Unknown is equivalent to Asset in MCOM.

⁸Aim is equivalent to MilitaryObjective in MCOM. Unlike the MilitaryObjective, Aim is not limited to military goals.

⁹The Execution Group is, in fact, the Offenders' proxy; however, sometimes Offenders decide to act directly and without using its execution wing

¹⁰It is worth noting that, other techniques such as disinformation are subsets of those three techniques [6, 7, 8].

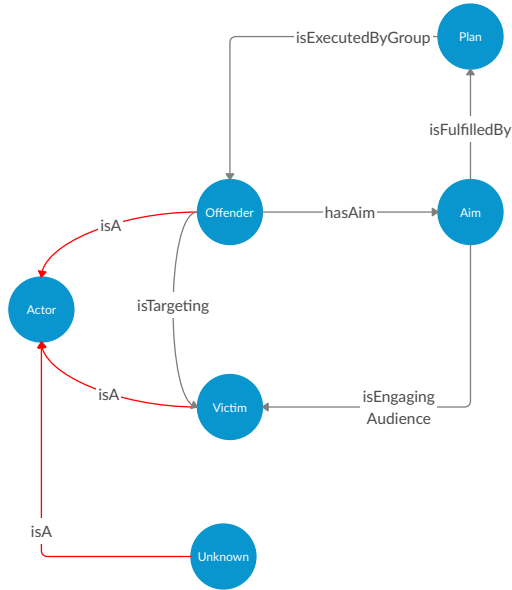


Figure 7.5: The component of goal definition and target selection.

7

development of narratives used in different Techniques. Figure 7.6 displays the entities and their relationships in the component of capability development. It also illustrates how the components of goal definition and target selection and capability development are related to each other.

MESSAGE IMPLANTATION

After choosing the right technique and development of the message, in this component, the implantation and circulation of the message is explained. This phase is one of the distinguishing factors between Cyber Operations and deliberate rumour spreading; hence there is no entity in MCOM regarding the message implantation into public, and all the entities in this component are inspired by other literature. The two entities in this component are Post and Other media. The Post refers to releasing the message among public using online social media. The Other media shows any kind of conventional media except online social media. For example, radio, television, newspaper, and web fall into this category. Sometimes an operation initiates in an online social media platform and then continues in parallel or transfers to conventional media. Sometimes this happens in the opposite, which means the operation begins in a conventional media and then continues or is followed in online social media platforms [245, 246, 247]. In this phase, the Offender or Execution Group plant the message among the public. Those who are exposed to the message might ignore it or engage in rumour process by reacting in different forms of sharing, commenting, liking, or reporting. The deliberate rumours need to reach their target audience (Victims); however, it is likely that other people are

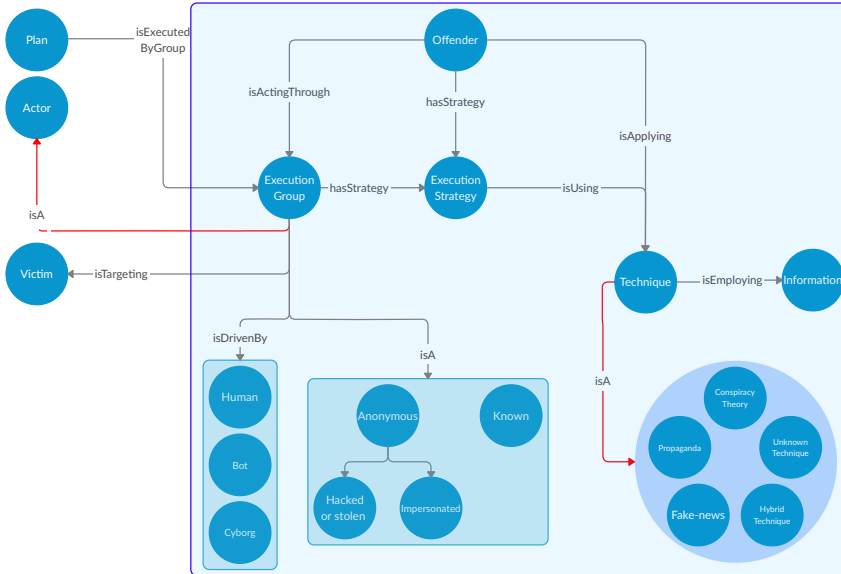


Figure 7.6: The sub-component of capability development (the highlighted part).

also exposed to their messages and manipulated by them (Unknowns). Figure 7.7 displays the entities and their relationships with other components.

IMPACT ASSESSMENT

This component reflects the implications and consequences of a deliberate rumour on different actors. It is entirely inspired by the MCOM and is composed of Effect and Quality/Aspect/Value elements. The Effect element captures the exact type of effect that an operation could have e.g. influence, damage, and disturbance on different Actor types such as the targeted actor – the Victim – as well as other types of actors that are Unknown, and possibly even by the actor responsible for conducting the operation – Offender. The exact aspect, quality, or value of an actor/system that is being impacted through this Effect is captured by the Quality/Aspect/Value element [248]. For instance, we could model an operation that might damage (as an Effect) the reputation (as a Quality/Aspect/Value) of an actor. Figure 7.8 displays the entities and their relationships in this component.

THE DRSM IN OPERATIONAL LEVEL

This section aggregates the elements of the DRSM and gives a one unified representation of the model. As Figure 7.9 displays, the blue circles are the elements, and the arrows show the relationships between them. The red arrows indicate the special relationship of inheritance. This figure gives us a bird-eye view on deliberate rumour spreading at an operational level. Based on this model, in a manipulation operation in the social

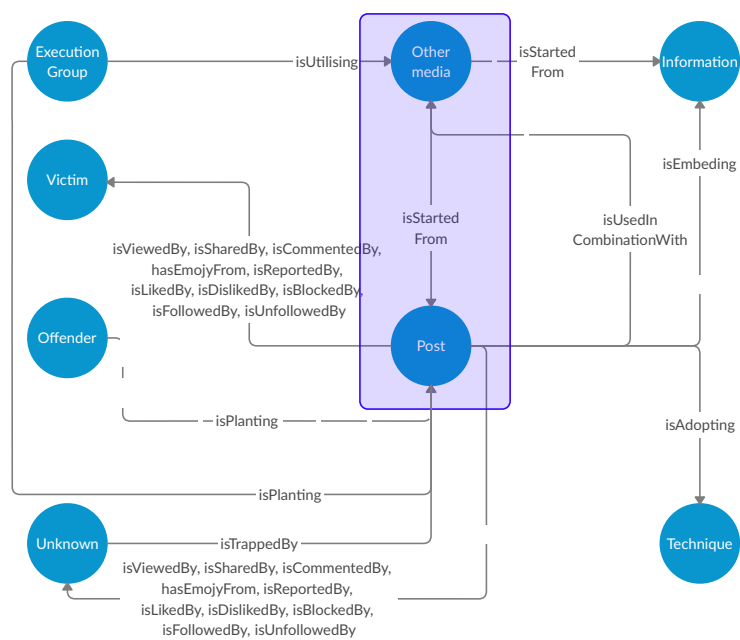


Figure 7.7: The sub-component of message implantation.

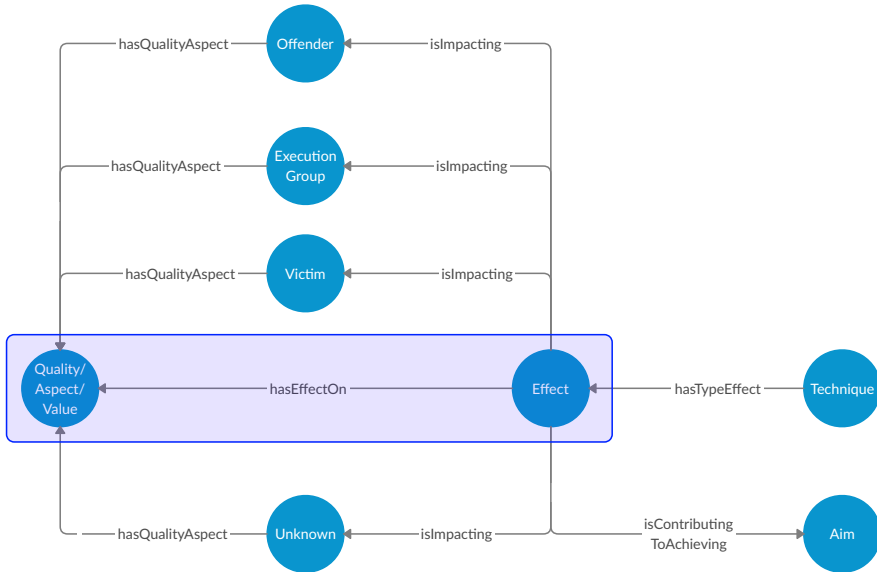


Figure 7.8: The sub-component of impact assessment.

media, an offender (which is the actor that conducts or executes the operation in social media) can either conduct an operation directly or through an execution group in order to achieve its aim (goal) by building a plan of action. To be able to do that, a strategy is executed by applying a specific technique (fake news, conspiracy theory, propaganda etc.) based on the information about the target group. For any of those techniques a message is required to be developed and planted into the social media. This series of actions can also be followed or practised in parallel in other media. The spread of the message has an effect or impact on qualities, values, or aspects of the target (victim), or other unknown actors or even the offender.

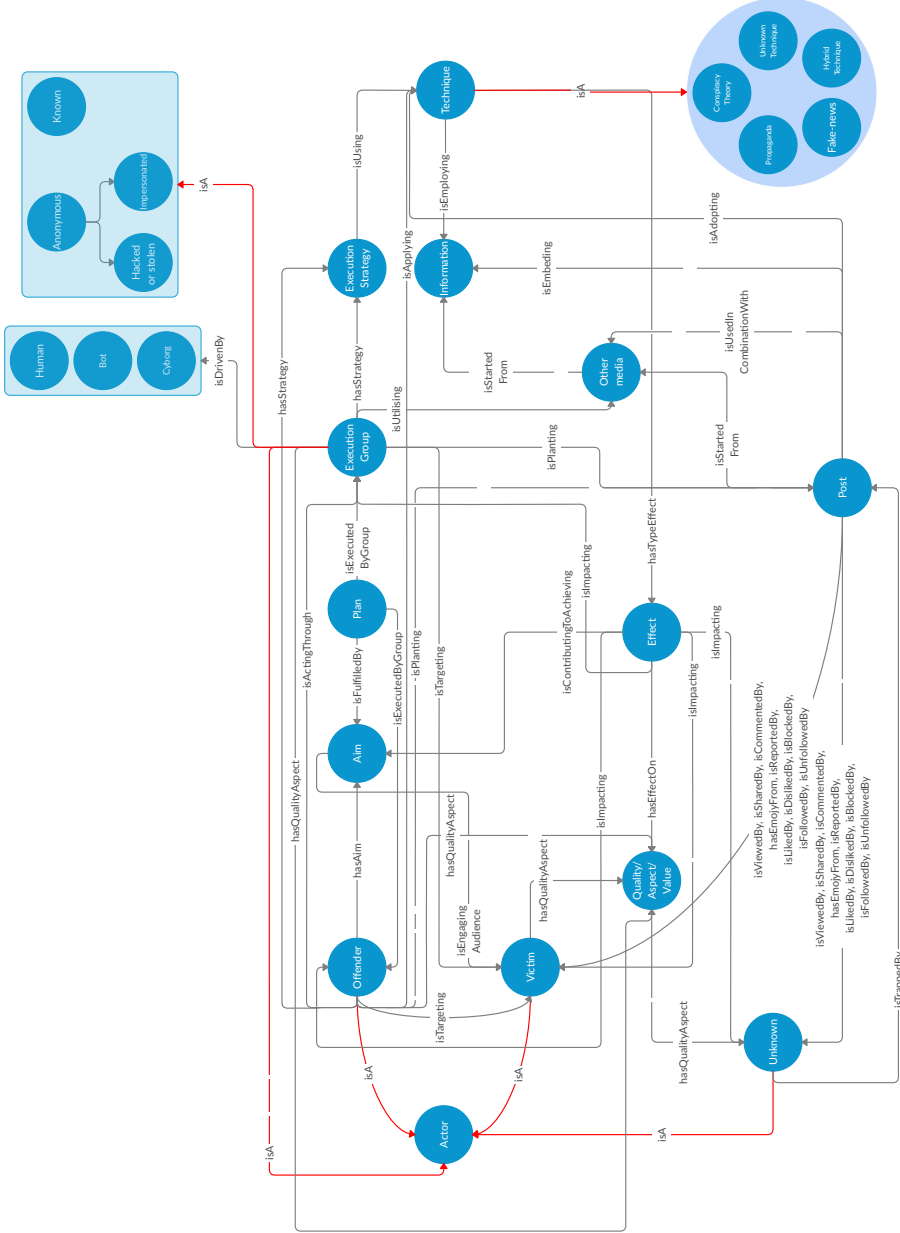


Figure 7.9: Model architecture design – logical flow of the deliberate rumour spreading in social media.

7.3.3. THE DATA MODEL

After explaining each component of the DRSM, this section introduces the corresponding data model which conceptualises and formalises the manipulation operations in social media using the OWL (Web Ontology Language) [249], and is defined as the following quadruple:

$$DRSM = \{E, D, R, I\} \quad (7.1)$$

where

E: the set of classes/entities/nodes of the model,

D: the set of data properties or attributes that characterise the classes,

R: the set of relations between instances/objects/data values of the classes;

I: the set of instances corresponding to classes.

The entities of the model (E), as well as its data properties (D) and relations (R) between the nodes, are further described. The individuals or instances of the model (I) will be explained in Section 7.4.2; I is further left to be filled in with data from more incidents as one of the extensions of this model. To build the data model, we need first to define the classes, data properties, and object properties, and then populate the model with data. In the following, the code snippets of each element are displayed.

The code snippet¹¹ 1 shows an example of defining the classes in the model. Here we can see that classes such as Actor, Technique, and Effect are child classes of the superclass Thing, while classes such as Offender and Propaganda are sub-classes (i.e. child classes) of classes Actor and Technique, respectively.

The code snippet 2 shows an example of defining the data properties of the model. Here we can see that classes Offender and ExecutionGroup (i.e. the offending actors that conduct such operation) are characterised by a property named StrategicNarrative which is of type String.

The code snippet 3 shows an example of defining the object properties of the model. Here we can see how individuals or objects of classes Offender and Aim are connected in the sense that a type of offending actor has an aim that has to be fulfilled in an operation. In this particular example, we can also see that this property is of type topObjectProperty which implies that it directly reflects that this relation applies to all the individuals from both considered classes.

The code snippet 4 shows an example for the operation that we will further discuss in 7.4.2, but we briefly tackle in this paragraph. At it can be seen, the individual with the name MojahedinEKhalq is the individual of the class Offender, i.e. that the offending actor of the operation used as exemplification (Alavi) is MojahedinEKhalq.

Moreover, a global view of the model that illustrates its complexity is depicted in Figure 7.10 which comprises all the entities, attributes, and relations. In this figure, the blue

¹¹To access a sample of the ontology, please refer to <https://bit.ly/36XyXPd>. For the complete code, please contact the authors.

Listing 1 Defining the classes (E) in OWL.

```

//Actor class:
<owl:Class rdf:ID="Actor"/>
//Technique class:
<owl:Class rdf:ID="Technique"/>
//Effect class:
<owl:Class rdf:ID="Effect"/>
...
//Offender class:
<owl:Class rdf:ID="Offender">
  <rdfs:subClassOf rdf:resource="#Actor"/>
</owl:Class>
//Propaganda class:
<owl:Class rdf:ID="Propaganda">
  <rdfs:subClassOf rdf:resource="#Technique"/>
</owl:Class>
...

```

7

Listing 2 Defining the properties (D) in OWL.

```

//StrategicNarrative data property:
<owl:DatatypeProperty rdf:about="...#StrategicNarrative">
  <rdfs:subPropertyOf rdf:resource="...#ExecutionGroup"/>
  <rdfs:domain rdf:resource="...#Offender"/>
  <rdfs:range rdf:resource="...#string"/>
</owl:DatatypeProperty>

```

circles represent the classes of the graph, the green rectangles represent the data properties of nodes with their yellow rectangles represent the types of attributes, and the blue rectangles represent relations between the classes (object properties), as depicted in Table 7.2. Moreover, detailed explanations regarding the classes, attributes, and relations that the models contain are provided in the Appendix.

Listing 3 Defining the relations (R) in OWL.

```
//hasAim object property:
<owl:ObjectProperty rdf:about="...#hasAim">
  <rdfs:subPropertyOf rdf:resource="...#topObjectProperty"/>
  <rdfs:domain rdf:resource="...#Offender"/>
  <rdfs:range rdf:resource="...#Aim"/>
</owl:ObjectProperty>
```

Listing 4 Defining the individuals (I) in OWL. Here we define an individual as an example. Further explanations about individuals will come in Section 7.4.2.

```
<owl:NamedIndividual rdf:about="...#MojahedinEKhalq">
  <rdf:type rdf:resource="...#Offender"/>
</owl:NamedIndividual>
```

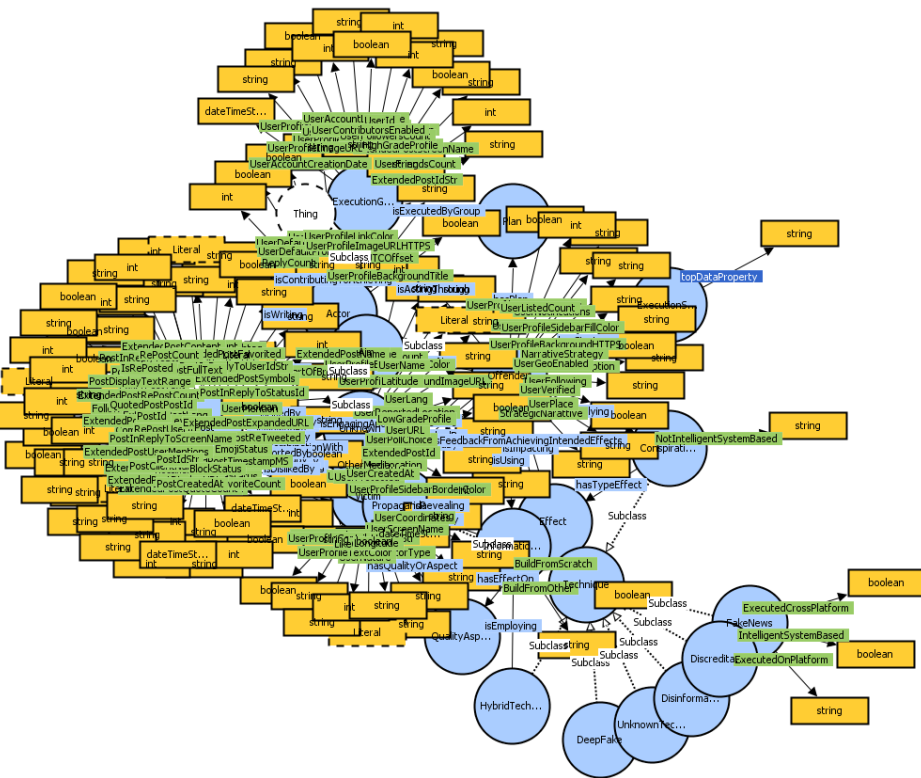









Figure 7.10: Global view of the model.

Table 7.2: Graphical representations of the model.

Graphical representation	Definition	Example
	The root node	Thing
	Entity	Actor
	Data property	HighGradeProfile
	Data property type	Boolean
	Object property	isTargeting
	Connection between two nodes as support for relations	For instance, arrow from node Offender to node Victim
	Connection between a parent node and child node	For instance, arrow from node parent Technique to node child Disinformation

7.4. MODEL EVALUATION

In this research, the model evaluation is conducted in two ways [250]. Firstly, through an interview with four experts with significant experience in the fields of Information Operations in social media. And secondly, through exemplification on real deliberate rumours. As a prerequisite, a technical evaluation using the HerMiT reasoner in order to guarantee the consistency and usability of the model was done, and no error occurred.

For the expert-based evaluation, five criteria of (i) realism, (ii) accuracy, (iii) clarity, (iv) conciseness, and (v) adaptability [251, 250] are taken into account. The results are discussed and compared in Section 7.4.1. The exemplification of the proposed model was done through a series of notoriously famous cases of Information Operation. Accordingly, in Section 7.4.2, each case study is introduced and further analysed.

7.4.1. EXPERT-BASED EVALUATION

We started the evaluation by building a database of potential domain experts. To this end, we searched LinkedIn (as the main database of organisations and individuals affiliation) by eight keywords: disinformation, social manipulation, misinformation, fake-news, propaganda, information operation, fact checking and conspiracy theory. We ended up with a list of nine relevant organisations. In the next step, we tried to prepare a contact list of experts who work in those organisations via LinkedIn, Twitter or in their website. After this step we could collect contact information of 19 domain experts. We contacted them all and six people responded us and four of them agreed to have an online conversation with us¹². Each interview took almost an hour and followed through a guideline that we prepared before. During the interview we provided the interviewees

¹²Disclaimer: Because social manipulation is a sensitive topic and due to security issues we cannot reveal any information about the organisations and the experts we contacted.

with a visualisation of the model (through screen sharing) and discuss them different aspects of the model from their point of view¹³.

The evaluation provided by the experts is further explained as follows. Firstly, the results of this evaluation are reported to see to what degree the proposed model complies with the dimensions of evaluation. Then in the second part, we go deeper into the interviews and discuss the experts remarks in more details.

We used five criteria of realism, accuracy, clarity, conciseness, and adaptability to evaluate the proposed model. Realism captures to what degree the proposed model is aligned with the real rumour campaigns. Accuracy measures how precise the proposed model describes its intended aim. Clarity shows to what extent the model is easy to understand and communicate. Conciseness evaluates the model from a redundancy point of view, and adaptability is responsible for measuring the generalisability of the model in a variety of deliberate rumour cases. We brought up a set of questions regarding measuring each dimension and discuss them with the interviewees.

For realism, all the experts stated the model reflects real deliberate rumour cases; however, they also said that rumour campaigns are constantly under development; thus, this model requires to receive constant updates. This leads us to one of the main limitations of this study, namely, the lack of publicly available datasets for rumour campaigns. Tackling this issue is an essential step towards developing effective and efficient models, tools, and strategies against deliberate rumour spreading. To overcome this issue, they suggested us to exemplify the model on older rumours launched during the Cold War such as Operation Infektion. Accordingly, this was done in the exemplification section of this chapter.

For accuracy, although the experts acknowledged the accuracy of the model, they provided us with a couple of suggestions. First, we need to take the collateral actors into account. To this end, a relation called `isTrappedBy` is added to the model. We should also take the fact into account that the content could be built from scratch (the expert that has suggested this called this process original behaviour) or be reused from another or former operation (the same expert that suggested this called this process advantageous behaviour). Correspondingly, two attributes were added: `BuildFromScratch` and `BuildFromOther`. Moreover, the final suggestion was to take the fact into account that the offender actors are always in advantage by developing unknown or hybrid techniques for achieving their aims. In this way, two new entities (`Unknown` and `Hybrid`) were added.

For the clarity, they expressed the model is understandable if the explanations accompany it. About the model conciseness, they said the model is concise, and there is no redundancy in it. For the adaptability, the experts pointed out that although different rumours have a similar skeleton, every culture, platform, and context has its nuances. Thus depending on the case, some details in the model would differ.

7.4.2. EVALUATION THROUGH EXEMPLIFICATION

In this section, we evaluate the proposed model by exemplification through three notoriously famous deliberate rumours: Operation Infektion, Operation Heshmet Alavi, and IRACopyPasta Campaign. In this section, each case study is introduced and furthermore

¹³The interview setting including the questions and consent form are in Section A.2.3 in appendix.

is mapped on the DRSM in order to reflect its realism as well as applicability to different contexts, actors, and time periods.

OPERATION INFEKTION

The Operation Infektion[252, 253, 254, 247] was an Information Operation conducted in the 80s by KGB Department A (Komitet Gosudarstvennoy Bezopasnosti – Committee for State Security in the Soviet Union) using disinformation as a major technique. This comes out as news published in July 1983 by an Indian magazine called Patriot under the name “AIDS may invade India: Mystery disease caused by U.S. experiments”. The news claimed that AIDS was a weapon aimed to kill gay people and African Americans, and was secretly created by U.S. scientists that were conducting biological warfare experiments in Fort Detrick Maryland. In September 1985 the news spread in Africa and was supported by two medical doctors. Two years later, in March 1987, the news was presented on a national American T.V. channel, from there rapidly spread all over the world and became one of the greatest cons with global impact. From further investigations and declarations provided by former KGB officers, the operations conducted by KGB aimed to change the perception of reality through different techniques applied to information such as implanting or altering it. Hence, Operation Infektion was a highly effective and successful Information Operation planned and executed in years to make sure its aim will be fulfilled: it produced mass manipulation through altering people’s perceptions together with distrust and reputation damage for U.S. Further, denial campaigns against this operation were conducted by U.S., and were responded by the Soviet Union with defensive denial campaigns and lead in the same year to apologies and stopping Operation Infektion.

7

OPERATION HESHMAT ALAVAI

“Heshmat Alavai” is a persona run by the political wing of Mojahedin-e-Khalq (MEK) -an Iranian opposition group which is advocating overthrowing the government- in order to broadcast propaganda against the Iranian government. Although this persona is highly active in Twitter, it used to appear in more conventional media such as Forbes, Hill, and AlArabiya by its article until an article in The Intercept revealed that it is a propaganda persona operated by MEK members in Albania [106]. Although Alavi introduces himself as “human rights & political activist”, one of his main strategies is to attack journalists, politicians, and long-standing news outlets which support diplomatic efforts regarding Iran issues. In the same vein, he constantly shows his support on non-diplomatic efforts such as imposing sanctions by promoting pro-war and pro-sanction people. After the Intercept report, Twitter appears to have suspended Alavi’s account; however, the account was reinstated shortly afterwards. Besides, many of Alavi’s articles are taken offline by the news outlet he used to publish¹⁴. Besides, it seems all those news outlets even the ones that did not remove Alavi’s pieces, stopped publishing its articles. However, Alavi’s Twitter account is still active and produces content against the Iranian government.

IRA COPYPASTA

The third deliberate rumour campaign that we exemplify is called IRA Copypasta. It is the name of a disinformation campaign on Instagram, which was discovered by Face-

¹⁴Please refer to the Appendix for more details regarding Alavi’s publication record in the news outlets

book in October 2019. They immediately took down 50 Instagram accounts belonging to this campaign. The posts were mostly about U.S. social and political issues and the 2020 presidential election. Facebook concluded that the operation “originated from Russia” and “showed some links to the Internet Research Agency (IRA)”. These accounts, all linked to the same operation, claimed to represent multiple politically active U.S. communities: black activist groups, advocates speaking out against police violence, police supporters, LGBTQ groups, Christian conservatives, Muslims, environmentalists, gun-rights activists, southern Confederates, and supporters of Senator Bernie Sanders and President Donald Trump. A minority of posts focused directly on the 2020 election.

Table 7.3 tabulates the exemplification of three above-mentioned deliberate rumour campaigns using the DRSM. In this table, different parts of each campaign is mapped to the components of the model.

Table 7.3: Model exemplification

Nodes	Operation Infektion [252, 253, 254, 247]	Operation Heshmat Alavi [106, 255, 256]	The IRACopyPasta Campaign [257]
Offender	Soviet Union (state /union)	Iranian opposition group Mojahedin-e-Khalq, which is known by the initials MEK	Russia
Victim	United States	Journalists and activists who support peaceful efforts regarding Iran issues and the Iranian government	American citizens
Unknown	The world	-	-
Aim	To tarnish U.S. brand by spreading the idea that AIDS was a biological weapon created by U.S. scientists which aimed to kill gay people and African Americans.	Influencing English-language audience and exerting pressure on political discourse against the Iranian government by mixing scathing denunciations of the Iranian government with not-so-subtle suggestions that it might be replaced by the MEK and its leader.	To reinforce division and hostility between different societal groups in U.S.

Table 7.3 continued from previous page

	<p>The planning is done considering the following rules:</p> <ol style="list-style-type: none"> 1. finding a crack in the society regarding economic, politic, or demographic issues. 2. making a big lie. 3. mixing a bits of truth into the lie. 4. concealing the hands (making it seem as if your story came from someone else – here the English based Indian newspaper that did not research its source and story) 5. finding the confirmation. 6. denial of everything. 7. playing the long game 	<p>It has a highly active account in Twitter which pursues the following three strategies:</p> <ol style="list-style-type: none"> 1. Attacking news agencies with long and strong editorial histories, and established journalists who support diplomatic activities regarding Iran. 2. Attacking Iranian authorities such as Foreign Minister. 3. Promoting non-diplomatic actions such as war and sanction against Iran, by praising or spreading the contents from partisan accounts. 	<p>The campaign is composed of three main principles:</p> <ol style="list-style-type: none"> 1. Finding the sensitive and polarised topics (e.g. presidential candidates for election 2020, gun right, police violence, muslims, etc.) 2. Posing as Americans in order to make the campaign believable 3. Posting on both sides of divisive issues
<p>Plan</p>	<p>KGB</p>	<p>Team of people from the political wing of the MEK in Albania</p>	<p>IRA (Internet Research Agency)</p>
<p>ExecutionGroup</p>			

Table 7.3 continued from previous page

<p>ExecutionStrategy</p>	<ol style="list-style-type: none"> 1. finding an English based newspaper 2. finding the scientist that would confirm the story 3. denial defence mechanism 	<p>Publishing articles in American press such as Forbes, Hill, the Daily Caller, or the Diplomat, etc. in addition to gaining attention in online social media.</p>	<p>Providing content by:</p> <ol style="list-style-type: none"> 1. Posting very little linguistic materials and focusing more on memes or memes and hashtags 2. Recreating content by reusing the content (mostly memes) that were originally posted by IRA in previous campaigns 3. In a few cases the posts contained long text, but these were usually copy-pasted from other online sources. The copy-pasting was done with some care: often, posts would select paragraphs from separate sections of their sources and rearrange their order or omit large sections, rather than copying the entire text 4. leveraging authentic content from American users by taking screenshot from American tweets and posting them in Instagram
--------------------------	---	---	---

Table 7.3 continued from previous page

Information	AIDS is a biological weapon created by U.S. scientists which aimed to kill gay people and African Americans.	Insisting on the fallacy of relating the Iran regional activities to JCPOA	Posting on the following topics: Black activism, Confederate/Southern heritage, Environmentalism, feminism, Muslim/Islamic unity, LGBTQ+, Liberal, Anti-Trump, Anti-Bernie, Religious, Conservative, Gun right, thin blue line
Technique	Black propaganda (disinformation)	Propaganda	Disinformation
Effect	Manipulation and distrust.	Backing U.S. decision to withdraw from JCPOA	It is not known yet
QualityAspect	Perceptions in regards with AIDS as a disease and U.S.'s reputation damage as its creator.	Global perceptions regarding Iranian government long-term vision, and regional activities and actions	It is not known yet

As already explained above, the code snippet 5 shows that the offending actor that conducts the Alavi operation is MojahedinEKhalq. That means that MojahedinEKhalq is an individual for the class Offender.

Listing 5 Offender data in Alavi rumour campaign.

```
<owl:NamedIndividual rdf:about="...#MojahedinEKhalq">
  <rdf:type rdf:resource="...#Offender"/>
</owl:NamedIndividual>
```

The code snippet 6 shows that the effect of this operation is SupportUSDecisionToWithdraw. That means that SupportUSDecisionToWithdraw is an individual for the class Effect.

Listing 6 Effect data in Alavi rumour campaign.

```
<owl:NamedIndividual rdf:about="...#SupportUSDecisionToWithdraw">
  <rdf:type rdf:resource="...#Effect"/>
</owl:NamedIndividual>
```

Figure 7.11 shows the global view for the data associated to Alavi's manipulation campaign. This means that for each class marked with a yellow circle, one or more individuals marked with a purple rectangle are contained.

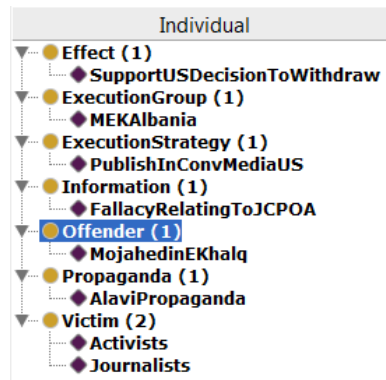


Figure 7.11: Data global view for Alavi rumour campaign.

7.5. CONCLUSION

In this research, we propose a fine-grained rumour model (DRSM) that benefits from both the misinformation machine and MCOM. Despite the common practice of the academia, which has a passive approach toward deliberate rumour spreading, in this work, we switched the perspective and looked at the rumour campaigns from the adversarial lenses. This approach helps us to capture the functional mechanisms underlying

rumour campaigns; and thus to take preemptive measures and develop resistance in advance. We evaluated the model through interviews with experts in this field. We also successfully tested our model in three notoriously famous misinformation campaigns. The deliberate rumour campaigns are often covert and their associated information are considered as classified. This engender in very few available case studies. Research on this domain requires more transparency. Perhaps with certain level of anonymization, the operation data could be released. More data would lead into more accurate models and more effective policies.

8

DISCUSSION & CONCLUSION

The ideal subject of totalitarian rule is ... people for whom the distinction between fact and fiction and the distinction between true and false no longer exist.

Hannah Arendt, *The Origins of Totalitarianism*

The emergence and spread of rumours is an ancient phenomenon with evolutionary origins. Human-beings benefit rumour spreading as a sense-making and threat management mechanism. It can also be used for the strategic purposes. The development of communication technologies provided an unprecedented opportunity for the vast spread of rumours. Despite the numerous advantages, the new communication technologies facilitate the emergence and circulation of the rumours. They increased the scale, speed, and scope of rumour spreading and turned it into a large-scale means of influence. Although charming from the offensive perspective, large-scale rumour spreading would be extremely concerning for the public as it may influence political, economic, and social well-being. Therefore it is of utmost importance to curb this phenomenon; otherwise, the repercussion could be far-reaching. To this end, this thesis aims at contributing to this subject domain and increasing the confrontation power against this wicked phenomenon.

This dissertation had the following research objective:

To systematically study the rumour confrontation within online social media.

We put forward the following research questions to accomplish the objective:

- RQ1. What is the rumour and how is it differentiated from its conceptual siblings?
- RQ2. To what extent social media streamline the spread of rumours?
- RQ3. What is the current status of rumour response strategies?
- RQ4. How ready is the academia regarding the spread of rumours?
- RQ5. How could we identify rumours in social networks automatically, consistently and in a timely manner?
- RQ6. How could we take preemptive measures regarding rumours in social media?

It was essential to lay down the building blocks of this dissertation at first. Thus we began with two concepts of rumour and social media. The notion of rumour refers to a controversial construct due to the lack of consensus regarding its conceptualisation. There is an epistemic crisis regarding the rumour and its conceptual siblings as they are often used interchangeably. This issue is addressed by looking at the rumour as a process instead of a product. We argue that the genesis of rumour and its conceptual siblings might be for different reasons; however, after the first generation, they become a very similar phenomenon. Therefore, different variations of false and unverified information such as misinformation, disinformation, propaganda, fake-news, pseudoscience, and conspiracy theory belong to the family of rumours. The second construct of this dissertation is social media. We investigate the role of different communication technologies and infer that the constellation of mediums together makes a sphere for the rumour circulation. In particular, we scrutinise the role of social media. They incorporate a set of successful features from other communication technologies as well as develop exclusive mechanisms that together greatly facilitate the emergence and spread of rumours. We

further study the role of recommendation systems as an exclusive feature of social media platforms via experimentation with real data extracted from YouTube. We observed a clear impact of this automation mechanism on the spread of rumours; however, this impact might be mediated by a variety of factors such as location, time, and rumour topic.

By having a clear understanding of rumour family and social media as its nurturing environment, then we can think of countering rumours. The first step is to understand what has been wrong so far in rumour control strategies and why the past counter strategies could not successfully control the surge of rumours. To address those concerns, we seek to understand the as-is circumstance of countering rumours. To this end, we review the organisational and governmental response to the rumour spreading since the second world war and analyse them based on the epidemic control framework. The reason behind choosing this framework is high degree of similarity between the spread of disease and dissemination of rumours. Based on this framework, the purpose of control strategies is either exposure minimisation, immunisation, or transmission reduction. After the analysis of the counter rumour strategies' primary goal, we concluded that despite the significance of immunisation strategy, it is the least practised one.

Thus, on the one hand, we have a sheer flow of rumours due to the social media features, and on the other hand, the response mostly involves a set of ephemeral and unreliable computational and non-computational strategies without creating long term immunity. To tackle this situation, our response should pursue the epidemic control framework, which combines the short-term strategies to take down the rumours instantly as well as strategies that train humans to not being deceived by the rumours.

To create immunity regarding rumours, the role of academia is crucial as it develops new methods and evaluates their effectiveness. Thus it is essential to understand the readiness of academia in this subject domain, as we would able to propel this field in case the growth is not proportionate to the current scale of rumour spreading. To this end, we use the emergence framework, which measures the novelty, growth, coherence, and impact of the scholars' attempts in the field of rumour studies. Our analysis shows an increasing trend for the growth, coherence, and impact and decreasing trend for the novelty. This is the result of organic growth in this subject domain so far; however, owing to the surge of rumours, the current degree of efforts seems insufficient, which entails considering external push strategy. This strategy could be implemented by establishing dedicated journals or conferences in order to create a stable community in this field. Fortunately, almost one year after this study, the journals of Harvard Misinformation Review and ACM Digital Threats from two highly respected publishers started their publications in this subject domain.

The other part of the response involves immediate action regarding rumours. Because of the scale and speed of rumour circulation in social media, it is not possible to manually identify rumours. This setting requires an algorithmic solution that can flag rumours automatically. Among the past computational approaches to the rumour resolution in social media, they mostly take the binary classification into account. This approach entails training the model based on two predefined classes of rumour and non-rumour. Although the rumour is most often defined similarly, non-rumour is defined arbitrarily. The lack of consistency in the definition of the classes in binary classifica-

tion makes the binary classifiers unreliable. To address those shortcomings regarding the computational rumour detection, we propose a novel approach based on one-class classification for automatic identification of the rumours.

Although the combination of rapid identification and immunisation makes a strong response to rumour circulation in social media, it is a passive approach to confront rumour. This means, our reaction to a rumour is based to past rumours. Thus, if a newly emerged rumour consists of novel elements, then most likely, we will not be able to counter it. To address this issue, we propose a data model with a pro-active perspective for rumour campaigns. In this approach, we look at the rumour from rumour-monger perspective, which allows us to develop preemptive measures. We build this model in an iterative manner. The skeleton of the model is based on the renowned model of misinformation machine. Then using other relevant models, expert's opinion, and real rumour campaigns, the model is operationalised. The last step is turning the model elements to the OWL code. Afterwards, the model could be populated based on real or synthetic data to simulate rumour campaigns.

8.1. SOCIETAL RELEVANCE

In order to discuss the social relevance of this thesis, I first quickly review two important aspects of rumour spreading. As it is explained before, the rumour spreading is a collective phenomenon arising when individuals participate in information circulation about an important topic. It does not appear in isolation, inside a single person's mind. Besides, it pertains to a topic that matters for all the rumours participants. Thus rumour is a social phenomenon about a topic that is instrumentally relevant for all of them. In fact, if we assume individuals in a society as threads, then rumour is like a fabric weaving the threads together. Along transferring information, rumours fulfil an implicit function of wielding influence on decision-making processes. It is a serious issue because rumour could be seen as a mediating factor on decisions regarding social, economic, and political issues. Rumours could mobilise people to withdraw their entire money and bankrupt the banks, change voters' opinion and influence the outcome of elections, and create traffic jam within an aid operation and jeopardise people's life. All these indicate what a significant phenomenon is a rumour at a societal level. Thus it is of utmost importance to develop resilience against something that can profoundly influence societal institutions. This thesis aims to address this potentially problematic phenomenon by proposing a response inspired by the epidemic control. The ultimate goal here is to make sure our society will not be manipulated like a puppet and will have a free will to decide among its options, and this thesis is a tiny step toward achieving that goal.

8.2. REFLECTION AND FUTURE RESEARCH

Besides the contributions that this dissertation puts forward, there are several critical challenges as well as multiple avenues for future research. One of the major limitations of this research pertains to data. There are very few numbers of publicly available annotated datasets in this field. This does not allow to capture the large volume and wide diversity of rumours in social media. As a result, the data-driven models are limited to specific time windows, incidents, and geographical locations and cannot be utilised fur-

ther than those contexts. The main reason behind the dearth of data is social media platforms. The content produced in those platforms is the core competency of social media companies. Without data, they cannot provide exclusive insights, and their revenue streams would completely drop out. Thus it is not surprising that they do not want to give away their main asset. They purport this is a necessary action to protect the users' privacy, although it is not difficult for social media companies to seal the back-doors and anonymise their data. One of the possible approaches to tackle this problem is getting the premium subscription to the platforms API. This is a costly solution which is only available for rich organisations or the ones in close partnership with social media companies. Because academic institutions barely fall into either of those eligible cases for the premium subscription, it is often not an option within the scientific community. The other approach could be producing synthetic data using generative adversarial networks (GANs). GANs are algorithmic architectures composed of two neural networks competing with each other in order to generate new, synthetic instances of data that resemble the training data. Although they can increase the volume of the data, they cannot address the diversity issue. The other approach to tackle the data problem is transfer learning (TL). In TL, an already trained deep neural networks on a particular task, is applied to a different but related problem. Despite GANs, TL can solve the diversity issue by creating a generalised model across different domains; however, it does not produce a large volume of data. Hence, one possible avenue for future research would be a combination of GANs and TL to address the volume and variety of problems at the same time. It helps to create a large diversified dataset out of small training samples.

The other limitation of this research which is somehow related to the previous one is about the accountability of social media platform companies regarding rumour spreading. They are basically the problem owner here, as a significant part of rumour spreading takes place in their platforms. They increase the users' engagement by manipulation of information circulation through mechanisms such as recommendation system and social bot, which have been shown their effectiveness in rumour promotion. Social media platforms are accountable regarding rumour dissemination, and they should take responsibility of it. However, they have not devoted sufficient attention to this problem as every now and then a major rumour emerges, circulates and is exposed to millions of users. The spread of rumour is not an internal organisational problem which should be resolved within the organisation. It is a problem with societal implications. Currently, quite a few scholars all around the world are working on countering rumours, but none or maybe very few numbers of them find the opportunity to test their ideas and proposed models in a real environment. Social media companies do not even allow external scientists to get involve the rumour confrontation process. Those companies often have a group or task-force composed of a few numbers of researchers working on this issue, while ironically, they can see the floating rumours in front of their eyes. In this vein, maybe one future avenue would be organising annual conferences by the social media companies regarding tackling rumours. In such venues, they could invite all the researchers working on this topic to present their studies and see how they can incorporate the novel proposed approaches into their platforms. The other way to make social media platforms accountable is called the co-regulation. Unlike statutory regulation (i.e., when state actor regulates the behaviour of organisations and its members

by implementing and enforcing legislation) and self-regulation (i.e., when non-state actor regulates the behaviour of its members), co-regulation says, the platforms will do X, or the state actors will do Y. In other words, platforms will demonstrate their ability to regulate rumours or state-actors will do it for them. It is threatening the companies with action if they do not actually engage in proper regulation themselves. Co-regulation benefits self- and statutory-regulation at the same time. It has a perfect sense of problem and enforcement guarantee due to engagement of both platforms and government which bring domain information and legislative support, respectively.

The domain dependency is one of the other limitations in this research. Dependency on specific subjects precipitates in less generic outcomes which can reduce the applicability of research. Rumour spreading is a domain-agnostic phenomenon which could emerge in any circumstance. Although not all the subject domains are the same, and some tend to provide a fertile ground for nurturing rumours, this does not mean to leave some domains for some others. Maybe some particular domains require more attention as they may lead to severe consequences; however, what should not be sacrificed here is justice about subject domains. Hence, on the one hand, we want to cover all subject domains, and on the other hand, we should note that some domains may lead to more adverse consequences. An ideal counter strategy takes rumours from all the subject domains into account with respect to their potential for emergence and circulation. To tackle this issue, what is essentially required is a better understanding of the rumour domain sensitivity. We need to know which subject-domains are more rumour prone. We also need to agree on a framework that allows us to grade rumours based on their severity. Clearly, some rumours are more dangerous than the other. By knowing the volume and severity of rumours per subject domain, we would be able to keep our counter-rumour attempts domain-independent and effective.

The other limitation is about technology emergence framework. As Rotolo explains in the technology emergence framework, the technology's performance in all dimensions shows the overall status of that technology in the emergence framework. Nonetheless, trying to pin down the absolute values for emergence dimensions is relatively meaningless. Besides, an emergent technology may only be compared with the other technologies if they are also in the same domain. This limitation clarifies that the comparative analysis with a baseline model must be in the same domain. To the best of our knowledge, there is no other study in this domain that could be used to compare. One way to tackle this issue is to do the retrospective analysis, which allows us to compare this research field with itself over time. The other limitation is about the inter-relation of dimensions. There are five dimensions in this framework which may co-evolve with each other. In other words, it is not clear whether and to what extent those dimensions influence each other. This might impact the framework outcome and needs to be investigated. The other limitation is about the fifth dimension, namely, uncertainty and ambiguity. Despite the thorough conceptualisation of this dimension, its operationalisation is largely unexplored which makes the framework outcome less accurate.

The next limitation is the lack of clear ethical guideline regarding social experiments in social media platforms. One of the salient research methodologies in the field of rumour studies is social experimentation. Specifically, when we are interested in measuring the impact of an intervention, this method is highly relevant. However, the number of

guidelines or sometimes hypothetical guidelines hinders the progress. Without a doubt, social experimentation is a sensitive topic which entails a comprehensive guideline to avoid repeating unpleasant experiences such as Stanford prison experiment. However, currently, there are so many of those guidelines in different levels (e.g., EU-, national-, university-, department-, professor-, and platform-level) that cannot fulfil its initial purposes and only works as a source of confusion which might eventually lead in the cancellation of the research project. Additionally, this flurry of guidelines could become a source of injustice as if a novel social experiment is considered ethical in the U.S while the same experiment is regarded as unethical within EU, and then a U.S. based researcher could proceed and publish his/her study while the EU based student cannot do that.

The rumour spreading is a moving target, and there is no silver bullet to get rid of it once and for all; however, by creating immunity as well as reducing the circulation rate, we could hope to build a resilient society against this obnoxious phenomenon.

A

APPENDIX

A.1. CHAPTER 5

Table A.1: Queries for data collection from Web of Science

	Query	Returned results	Description
1	TS = ((rumour) or ("blind item*") or (snopes and fact) or (hoax) or ("fake-news") or ("fake news") or ("false dilemma") or ("common misconceptions*") or ("character assassination") or (defamation) or (disinformation) or (factoid) or (fallacy) or (propaganda) or ("Active measures") and (russia or soviet) or ("denial and deception") or ("false flag") or ("information warfare") or (kompromat) or ("post truth") or ("alternative fact*") or ("urban-legend*") or ("urban legend*") or ("media manipulation*") or (truthiness) or ("fact-check*") or ("fact check") or ("factcheck*") or ("smear campaign*") or (pseudoscience) or (gossip) or (misinformation))	25769	Some of the returned manuscripts are about computer networks which is irrelevant to the field of unverified information

2	<p>TS = ((rumour) or ("blind item*") or (snopes and fact) or (hoax) or ("fake-news") or ("fake news")) or ("false dilemma") or ("common misconceptions*") or ("character assassination") or (defamation) or (disinformation) or (factoid) or (fallacy) or (propaganda) or ("Active measures") and (russia or soviet) or ("denial and deception") or ("false flag") or ("information warfare") or (kompromat) or ("post truth") or ("alternative fact*") or ("urban-legend*") or ("urban legend*") or ("media manipulation*") or (truthiness) or ("fact-check*") or ("fact check*") or ("factcheck*") or ("smear campaign*") or (pseudoscience) or (gossip) or (misinformation) NOT TS=(WiFi OR "ad hoc" OR ad-hoc OR "gossip protocol*" OR "gossiping protocol*" OR "gossip algorithm*" OR "gossiping algorithm*" OR "cellular network*" OR "mobile network*" OR VANET OR "sensor network*" OR "radio network*" OR "Wireless network*" OR "Push Pull" OR Push-pull OR "Push vs. Pull" OR "Push & Pull" OR "Push & Pull protocol*" OR quasirandom OR Quasi-random OR "acoustic network*" OR "peer-to-peer network*" OR "peer to peer network*" OR "p2p network*" OR "distributed system*" OR "communication network*" OR "communication-network*" OR "gossip-based" OR "randomized gossip" OR "periodic gossip*" OR microgrid)</p>
24021	
	<p>Surprisingly, many of the returned manuscripts are in the field of Oncology which seemed a little bit strange. After careful examination, we discovered WoS OCR algorithm does not work accurately and significant part of the returned papers contain term of "tumour" instead of "rumour"</p>

3	<p>TS = ((rumour) or ("blind item*") or (snopes and fact) or (hoax) or ("fake-news") or ("fake news") or ("false dilemma") or ("common misconceptions*") or ("character assassination") or (defamation) or (disinformation) or (factoid) or (fallacy) or (propaganda) or ("Active measures") and (russia or soviet) or ("denial and deception") or ("false flag") or ("information warfare") or (kompromat) or ("post truth") or ("alternative fact*") or ("urban-legend*") or ("urban legend*") or ("media manipulation*") or (truthiness) or ("fact-check*") or ("fact check*") or ("factcheck*")) or ("smear campaign*") or (pseudoscience) or (gossip) or (misinformation) NOT TS=(WiFi OR "ad hoc" OR ad-hoc OR "gossip protocol*" OR "gossiping protocol*" OR "gossip algorithm*" OR "gossiping algorithm*" OR "cellular network*" OR "mobile network*" OR VANET OR "sensor network*" OR "radio network*" OR "Wireless network*" OR "Push Pull" OR Push-pull OR "Push vs. Pull" OR "Push & Pull" OR "Push & Pull protocol*" OR quasirandom OR Quasi-random OR "acoustic network*" OR "peer-to-peer network*" OR "peer to peer network*" OR "p2p network*" OR "distributed system*" OR "communication network*" OR "communication-network*" OR "gossip-based" OR "randomized gossip" OR "periodic gossip*" OR microgrid OR tumour)</p>	21675	<p>Although, a considerable noise are removed from the dataset, there are still some irrelevant results, because for example there is a person called "Rumer", but WoS OCR algorithm recognized at as "rumor", or there is Russian physicist called "Rumour" which apparently is confused by the concept of "rumour", or there is a geological site in Australia called "Rumour" which will be confused by the rumour concept again</p>
---	---	-------	---

<p>4</p>	<p>TS = ((rumour) or ("blind item*") or (snopes and fact) or (hoax) or ("fake-news") or ("fake news") or ("false dilemma") or ("false dilemma") or ("common misconceptions*") or ("character assassination") or (defamation) or (disinformation) or (factoid) or (fallacy) or (propaganda) or ("Active measures") and (russia or soviet) or ("denial and deception") or ("false flag") or ("information warfare") or (kompromat) or ("post truth") or ("alternative fact*") or ("urban-legend*") or ("urban legend*") or ("media manipulation*") or (truthiness) or ("fact-check*") or ("fact check") or ("factcheck*") or ("smear campaign*") or (pseudoscience) or (gossip) or (misinformation)) NOT TS=(WiFi OR "ad hoc" OR ad-hoc OR "gossip protocol*" OR "gossiping protocol*" OR "gossip algorithm*" OR "gossiping algorithm*" OR "cellular network*" OR "mobile network*" OR VANET OR "sensor network*" OR "radio network*" OR "Wireless network*" OR "Push Pull" OR Push-pull OR "Push vs. Pull" OR "Push& Pull" OR "Push & Pull protocol*" OR quasirandom OR Quasirandom OR "acoustic network*" OR "peer-to-peer network*" OR "peer to peer network*" OR "P2P network*" OR "distributed system*" OR "communication network*" OR "communication-network*" OR "gossip-based" OR "randomized gossip" OR "periodic gossip*" OR microgrid OR tumour) AND Manual examination</p>	<p>21571</p>	<p>This is the most accurate dataset on rumour dissemination</p>
----------	--	--------------	--

A**A.2. CHAPTER 7****A.2.1. DATA MODEL**

Table A.2: Classes of the model (E).

Entity number	Entity name	Entity definition
1	Actor	Refers to different classes of actors that are either conducting an operation, the ones targeted by the operation, or other collateral actors (other actor(s) than the intended victim) that have not been directed and intentionally targeted, but are impacted. In this research, the actors can be either nation-state actors, non-state actors (groups or organisations that are not associated with any state actor) or hybrid actors (groups or organisations that are linked and are supported by state actors or a combination between state actors and non-state actors) [243].
2	Offender	Refers to the actor(s) that plan(s) and execute(s) an operation.
3	ExecutionGroup	Refers to the actor(s) that execute(s) an operation based on a specific order. However, in case that the Offender acts directly, then the ExecutionGroup does not exist.
4	Victim	Refers to the actor(s) targeted by the offender(s) actor(s) through an operation.
5	Unknown	Refers to other collateral actor(s) (other actor(s) than the victim) that can be impacted (directly or indirectly) by the operation (Maathuis et al., 2018).
6	Aim	Refers to the objective or goal of the operation that helps reaching a desired end state [258].
7	Plan	Refers to the plan that an offender actor has to be able to further execute an operation.
8	ExecutionStrategy	Refers to the strategy that an offender actor has to be able to execute an operation [259].
9	Anonymous account	Refer to the account that hide their real identity.
10	Hacked or stolen account	Refers to the accounts that belong to others and have been taken over through stealing or hacking.
11	Impersonated account	Refers to the accounts that pretend their account belong to others.
12	Known account	Refers to the accounts that reveal their real identity.
13 - 15	Human/Bot/Cyborg	Refers to the accounts that are driven by humans, bots or cyborgs.

Table A.2 continued from previous page

16	Technique	Refers to the technique that an offender actor uses in an operation, and is classified as: Propaganda, Fake-news Conspiracy Theory, HybridTechnique and UnknownTechnique.
17-21	Propaganda, Fake-news Conspiracy Theory, HybridTechnique and UnknownTechnique.	Refers to different manipulation techniques.
22	Other media	Refers to other media channels that could be used in an operation such as state-sponsored televisions/radio, independent televisions/radio etc.
23	Information	Refers to the content, message or topic that is used to develop a narrative using a technique in an operation in order to achieve its aim.
24	Post	Refers to the post that contains the information used in an operation.
25	Effect	Refers to the impact of an operation on both the victim actor (intended target) or other collateral actor(s).
26	Quality/Aspect/Value	Refers to the qualities, aspects or values (e.g. stability, trust, democracy) of actors that are being impacted through this operation.

Table A.3: Attributes/Data properties of the model (D).

Data number	Data property name	Data property Definition	Characterising Entity	Data Type
1	PartOfBroaderIO	Refers to checking or assessing if an operation is part of a broader operation or campaign.	Aim	Boolean
3	StrategicNarrative	Refers to the strategic narratives behind an operation, and that can imply considering political, economic, military narratives etc.	Offender and Execution Group	String
4	BuildFromScratch	Refers to checking or assessing if the content used in this operation (e.g. Information entity) is built from scratch or not (i.e. reused).	Information	Boolean
5	BuildFromOther	Refers to pointing or referring to the initial source of the content used or reused and that is integrated in the content of an operation (e.g. Information entity).	Information	String
6	ExecutedCrossPlatform	Refers to checking or assessing if an operation and its corresponding technique(s) are conducted on multiple platforms.	Technique	Boolean
7	ExecutedOnPlatform	Refers to the name of the platform where an operation and its corresponding technique(s) are conducted.	Technique	String
8	HighGradeProfile	Refers to checking or assessing if the targeted victim of an operation has a high grade profile implying to be (very) significant to the operation in achieving its desired aim(s).	Actor	Boolean
9	UserAccountLanguage	Refers to the language used by a user.	Actor	String

Table A.3 continued from previous page

10	UserProfileURL	Refers to the URL (Uniform Resource Locator) used by a user. The type of this attribute is String.	Actor	String
11	PostCreationTime	Refers to the exact moment when the post was released in an operation.	Post	DateTimeStamp
12	PostLanguage	Refers to the language used by a post in an operation.	Post	String
13	ReplyCount	Refers to checking or assessing to number of replies that a post has.	Post	Integer
14	ViewStatus	Refers to checking or assessing if a post of an operation was viewed or not.	Post	Boolean

Table A.4: Relations/Object properties of the model (R).

Object property number	Object property name	Object property Source	Object property Destination	Object property definition
1	has Aim	Offender	Aim	Implies that one or more actors have specific aim(s) or goal(s) to achieve by conducting an operation.
2	isFulfilledBy	Plan	Aim	Implies that the aim is fulfilled by the plan.
3	isTargeting	Offender / ExecutionGroup	Victim	Reflects the direct link between the offender that is conducted an operation in order to target a victim in an operation.
4	isUtilising	ExecutionGroup	Other media	Reflects that the Execution Group benefits Other media (TV, Radio, ...) for the operation.
5	isExecutedByGroup	Plan	Offender / Execution-Group	Reflects that a well-defined and structured plan is designed and followed in order to achieve well defined effects (i.e. aim) in an operation.
6	isEngagingAudience	Aim	Victim	Reflects how the aim of an actor/some actors implies influencing specific victim(s).
7	isActingThrough	Offender	ExecutionGroup	Raptures the executing actor(s) that act on behalf of the Offender actor(s).
8	isContributingToAchieving	Effect	Aim	Raptures the direct relation between Effect and Aim in order to reflect how different effects contribute to the achievement of the aim of one or more actors.

Table A.4 continued from previous page

9	hasStrategy	Offender / ExecutionGroup	ExecutionStrategy	Reflects that a well-defined and well-structured strategy is thought and built in order to achieve the aim(s) of an operation.
10	isStartedFrom	Other media	Information	Reflects use the Information to contribute in the operation.
11	isUsing	Execution Strategy	Technique	Determines which technique is used in the operation.
12	isPlanting	Offender / ExecutionGroup	Post	Plants the message into the social media
13	isAdopting	Post	Propaganda / FakeNews / ConspiracyTheory / HybridTechnique / UnknownTechnique	Illustrates what are the techniques used in a specific post in an operation.
14	isDrivenBy	ExecutionGroup	Human / Bot / Cyborg	Reflects who is behind the social media accounts.
15	isEmbedding	Post	Information	Reflects the social media message incorporates the Information
16	isStartedFrom	Post	Other media	Points to the fact that the Information Operation initiates in social media.
17	isStartedFrom	Other media	Post	Points to the fact that the Information Operation initiates in conventional media such as TV, radio, or newspaper.

Table A.4 continued from previous page

18	isUsedInCombinationWith	Post	Other media	Implies the fact that an operation is used with other social media manipulation campaigns or means outside social media such as classical journalistic means e.g. newspapers and TV.
19	isApplying	Offender / ExecutionGroup	Propaganda / FakeNews / ConspiracyTheory / HybridTechnique / UnknownTechnique	Reflects what kind of technique(s) are used by specific actor(s).
20	isEmploying	Propaganda / FakeNews / ConspiracyTheory / HybridTechnique / UnknownTechnique	Information	Depicts what kind of techniques engages the information in an operation (weaponization).
21 – 30	isViewedBy / isSharedBy / isCommentedBy / hasEmojiFrom / isReportedBy / isLikedBy / isDislikedBy / isBlockedBy / isFollowedBy	Post	Victim / Unknown	Reflects that the social media post was [viewed / shared / commented / produced emoji / is reported / produced a positive emotion / produced a negative emotion / is blocked by / is followed by / is unfollowed by] by the targeted actor or other collateral actor(s).
31	isImpacting	Effect	Offender / ExecutionGroup / Victim / Unknown	Shows what kind of effects is impacting the actors (the targeted ones, the ones conducting and responsible for the operation, as well as other collateral actors).

Table A.4 continued from previous page

32	hasTypeEffect	Propaganda / FakeNews / ConspiracyTheory / HybridTechnique / UnknownTechnique	Effect	Specifies what techniques have specific effects in an operation.
33	hasQualityOrAspect	Offender / ExecutionGroup / Victim / Unknown	QualityAspectOrValue	Illustrates the qualities, aspects or values affected in an operation.
34	hasEffectOn	Effect	QualityAspectOrValue	Illustrates which effect impact different qualities, aspects or values of actors affected in an operation.
35	isTrappedBy	Unknown	Post	Reflects how other collateral actor(s) (i.e. actors different than the targeted ones - Victim) can be impacted in an operation.
36	isA	ParentNode	ChildNode	Captures the inheritance relation between a parent node and a child node.

A.2.2. MODEL EVALUATION

Table A.5: The list of news outlets that published Heshmat Alavi's articles.

News outlet	First Article Date	Last Article Date	Article Count	Availability
Alarabiya	13 December 2016	7 January 2019	108	The articles are still available.
Forbes	22 December 2016	21 December 2017	In a span of a year, between April 2017 and April 2018, Alavi published a staggering 61 articles for the Forbes website [106]. However, by searching at Forbes website only 6 articles with Heshmat Alavi as the author appear. According to [256] Forbes has removed Alavi pieces from its website.	The articles content are no longer available ¹ .
Daily Caller	28 November 2016	9 February 2017	5	The articles are still available.
The Hill	26 October 2014	3 January 2017	6	The articles are still available.
The Diplomat	13 March 2017	13 May 2017	5	The articles content are no longer available ² .

¹ Editor Note: "This page is no longer active".

² Editor's Note: "A June 2019 investigation revealed that this article was authored by a false persona. Accordingly, this article has been retracted in full for not meeting our standards on authorship disclosure. The Diplomat regrets this situation."

Table A.5 continued from previous page

VOA	-	-	-	<p>After Washington Post reporter, Jason Rezaian piece on Heshmat Alavi, VOA sent him the following statement about about #HeshmatAlavi and articles attributed to that name.: Regarding your opinion piece in the Washington Post, the Voice of America has removed the articles and is in the process of identifying any other references to “Heshmat Alavi” on its digital platforms, intending to delete those as well. The Voice of America makes every effort to serve as a consistently reliable and authoritative source of news and information and to fulfill its mission of telling America’s story and providing factual, truthful information to those who have no other access to it</p>
-----	---	---	---	--

A.2.3. INTERVIEW SETTING

Interview Questions

- Activity 1: introduce ourselves and the project + explain the ethical protocol (privacy, anonymous) + explain that our model is not yet published (2 min)
- Activity 2: introduction questions
 1. Q1. What is your name and function? (2 min)
 2. How familiar are you with Information Operations and covert Information Operations in social media? (2~3 min)
- Activity 3: introduce the model in general without showing the model yet (2~3 min)
- Activity 4: share desktop, show and explain the model (5 min)
 1. Do you think that real offensive Information Operations in social media work / function based on the mechanism captured in the model? (5 min)
 2. Do you think that the model concisely describes offensive Information Operations in social media? (5 min)
 3. Do you think that the model can be applied in different contexts, by different actors having specific aims and techniques in social media? (5 min)
 4. Do you think that the model is clear, understandable, and easy to communicate to others? (5 min)
 5. Do you have other remarks or suggestions regarding the discussed model? (5 min)
- Activity 5: additional questions
 1. Could you please recommend us someone else to discuss with our model? (4 min)
 2. Would you be available in the future for another interview or survey? (1 min)
- Activity 6: concluding remarks from our side (including that we can share our published with them) and from their side => thank you, bye! (2~3 min)

Research Interview Consent Form (Ethical Protocol)
--

Research Title: Capturing the Sense of Using Information as a Weapon in Social Media: a Computational Approach

Name of Researchers (Interviewers): Amir Ebrahimi Fard and Clara Maathuis

Name of Research Participant (Interviewee):

Interview Date: . .2019

Estimation time: 45 – 60 minutes.

Thank you for agreeing to be interviewed as part of this research. This document relies on ethical procedures for academic research and aims at informing the participants of this interview and explicitly requires their agreement. Would you therefore read the following information and sign this form to certify that you approve and agree with the following conditions:

- You have been informed by the Researchers of the purpose of this interview.
- Your participation is voluntary.
- The interview will not be recorded, but a resume transcript will be produced during the interview.
- Upon request the transcript can be provided to the Interviewee.
- The transcript will be analysed only by the Researchers (Interviewers).
- All or parts of the interview information will be integrated in further academic publications, reports or presentations.
- Any summary or direct quotation from this interview will be made available through academic publications, reports, or presentations.
- In the produced academic publications, reports or presentations your identity will be fully anonymised so that you cannot be identified.
- You have the right to do not answer to one or more of the questions or to stop this interview in any moment.
- Any variations or modifications of the above conditions will only occur with your explicit request or approval.

By signing this form you agree with all mentioned conditions.

Research Participant name:

Research Participant signature:

Date:

Researchers signature:

Date:

If you have any further questions or concerns, please contact: Amir Ebrahimi Fard (a.ebrahimifard@tudelft.nl) or Clara Maathuis (clara.maathuis@tudelft.nl)

BIBLIOGRAPHY

- [1] Lee Howell et al. "Digital wildfires in a hyperconnected world". In: *WEF report 3.2013* (2013), pp. 15–94.
- [2] Bente Kalsnes. "Fake news". In: *Oxford Research Encyclopedia of Communication*. 2018.
- [3] Fredrick. Koenig. *Rumor in the marketplace : the social psychology of commercial hearsay*. Auburn House Pub. Co, 1985, p. 180.
- [4] G. W. Allport and L. Postman. *The Psychology of Rumor*. 1st. Henry Holt and Company, 1947.
- [5] T. A. Knopf. *Rumors, race, and riots*. Transaction Publishers, 1975, p164. ISBN: 1412805570.
- [6] N. DiFonzo. *The watercooler effect : an indispensable guide to understanding and harnessing the power of rumors*. New York, NY, USA: Avery, 2009, p. 291.
- [7] Nicholas. DiFonzo and Prashant. Bordia. *Rumor psychology : social and organizational approaches*. American Psychological Association, 2007, p. 292.
- [8] Nicholas Difonzo and Prashant Bordia. *Rumors influence: Toward a dynamic social impact theory of rumor*. Frontiers of social psychology. New York, NY, US: Psychology Press, 2007, pp. 271–295.
- [9] R. H. Knapp. "A Psychology of Rumour". In: *Public Opinion Quarterly* 8.1 (1944), pp. 22–37. DOI: 10.1086/265665.
- [10] Adam B Ellick and Adam Westbrook. *Scopus vs. Web of Science vs. Google Scholar*. 2018. URL: <https://www.nytimes.com/2018/11/12/opinion/russia-meddling-disinformation-fake-news-elections.html>.
- [11] Max Fisher. *Syrian hackers claim AP hack that tipped stock market by \$136 billion. Is it terrorism*. 2013. URL: <https://www.washingtonpost.com/news/worldviews/wp/2013/04/23/syrian-hackers-claim-ap-hack-that-tipped-stock-market-by-136-billion-is-it-terrorism/>.
- [12] J. Farrell, K. McConnell, and R. Brulle. "Evidence-based strategies to combat scientific misinformation". In: *Nature Climate Change* 9.3 (Mar. 2019), pp. 191–195. ISSN: 1758-678X. DOI: 10.1038/s41558-018-0368-6. URL: <http://www.nature.com/articles/s41558-018-0368-6>.
- [13] Soroush Vosoughi, Deb Roy, and Sinan Aral. "The spread of true and false news online." In: *Science* 359.6380 (Mar. 2018), pp. 1146–1151.
- [14] Tamotsu Shibutani. *Improvised News: A Sociological Study of Rumor*. Ardent Media, 1966.

- [15] Bradley Franks and Sharon Attia. "Rumours and Gossip as genres of communication". In: *The Social Psychology of Communication*. Springer, 2011, pp. 169–186.
- [16] Nicolas Turenne. "The rumour spectrum". In: *PLOS ONE* 13.1 (Jan. 2018). Ed. by Frederic Amblard.
- [17] D. M. J. Lazer et al. "The science of fake news." In: *Science (New York, N.Y.)* 359.6380 (Mar. 2018), pp. 1094–1096. DOI: 10.1126/science.aao2998.
- [18] Jo Fox. "'Fake news'—the perfect storm: historical perspectives". In: *Historical Research* 93.259 (2020), pp. 172–187.
- [19] Silvio Waisbord. "Truth is what happens to news: On journalism, fake news, and post-truth". In: *Journalism studies* 19.13 (2018), pp. 1866–1878.
- [20] Nicholas A Christakis and James H Fowler. *Connected: The surprising power of our social networks and how they shape our lives*. Little, Brown Spark, 2009.
- [21] Heidi Oi-Yee Li et al. "YouTube as a source of information on COVID-19: a pandemic of misinformation?" In: *BMJ Global Health* 5.5 (2020).
- [22] Spring, Marianna. *Coronavirus: False claims viewed by millions on YouTube*. 2020. URL: https://www.bbc.com/news/technology-52662348?utm_source=dlvr.it&utm_medium=twitter.
- [23] Paul Lewis. *Fiction is outperforming reality: how YouTube's algorithm distorts truth*. 2018. URL: <https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth>.
- [24] C. Shao et al. "The spread of low-credibility content by social bots". In: *Nature Communications* 9.1 (2018), p. 4787. DOI: 10.1038/s41467-018-06930-7.
- [25] Samuel C Woolley and Philip N Howard. *Computational propaganda: political parties, politicians, and political manipulation on social media*. Oxford University Press, 2018.
- [26] Andrew M Guess et al. "A digital media literacy intervention increases discernment between mainstream and false news in the United States and India". In: *Proceedings of the National Academy of Sciences* 117.27 (2020), pp. 15536–15545.
- [27] Yariv Tsfati et al. "Causes and consequences of mainstream media dissemination of fake news: literature review and synthesis". In: *Annals of the International Communication Association* (2020), pp. 1–17.
- [28] Chris Marsden, Trisha Meyer, and Ian Brown. "Platform values and democratic elections: How can the law regulate digital disinformation?" In: *Computer Law & Security Review* 36 (2020), p. 105373.
- [29] Sinan Aral and Dean Eckles. "Protecting elections from social media manipulation". In: *Science* 365.6456 (2019), pp. 858–861.
- [30] Nir Grinberg et al. "Fake news on Twitter during the 2016 US presidential election". In: *Science* 363.6425 (2019), pp. 374–378.
- [31] Alexandre Bovet and Hernán A Makse. "Influence of fake news in Twitter during the 2016 US presidential election". In: *Nature communications* 10.1 (2019), pp. 1–14.

- [32] Samuel Carruthers. “Countering Disinformation: A Case Study of Government Responses to Russian Information Warfare”. In: (2019).
- [33] Freja Hedman et al. “News and political information consumption in Sweden: Mapping the 2018 Swedish general election on Twitter”. In: *Data Memo 2018.3*. Project on Computational Propaganda, 2018.
- [34] Emilio Ferrara. “Disinformation and social bot operations in the run up to the 2017 French presidential election”. In: *arXiv preprint arXiv:1707.00086* (2017).
- [35] Philip N Howard and Bence Kollanyi. “Bots, # StrongerIn, and # Brexit: computational propaganda during the UK-EU referendum”. In: *Social Science Research Network* (2016). URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2798311.
- [36] Przemyslaw M Waszak, Wioleta Kasprzycka-Waszak, and Alicja Kubanek. “The spread of medical fake news in social media—the pilot quantitative study”. In: *Health policy and technology* 7.2 (2018), pp. 115–118.
- [37] J. Rick Ponting. “Rumor Control Centers”. In: *American Behavioral Scientist* 16.3 (Jan. 1973), pp. 391–401.
- [38] Philipp Lorenz-Spreen et al. “Accelerating dynamics of collective attention”. In: *Nature communications* 10.1 (2019), pp. 1–9.
- [39] Deen Freelon and Chris Wells. “Disinformation as political communication”. In: *Political Communication* 37.2 (2020), pp. 145–156.
- [40] Claire Wardle. “Fake news. It’s complicated”. In: *First Draft* 16 (2017).
- [41] Priyanka Meel and Dinesh Kumar Vishwakarma. “Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities”. In: *Expert Systems with Applications* (2019), p. 112986.
- [42] Cailin O’Connor and James Owen Weatherall. *The misinformation age: How false beliefs spread*. Yale University Press, 2019.
- [43] Francesco Pierri and Stefano Ceri. “False news on social media: a data-driven survey”. In: *ACM Sigmod Record* 48.2 (2019), pp. 18–27.
- [44] Maria D Molina et al. ““Fake news” is not simply false information: a concept explication and taxonomy of online content”. In: *American Behavioral Scientist* (2019), p. 0002764219878224.
- [45] Srijan Kumar and Neil Shah. “False Information on Web and Social Media: A Survey”. In: *ArXiv* (Apr. 2018).
- [46] S Kumar, R West, and J Leskovec. “Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes”. In: *Proceedings of the 25th international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016.
- [47] Xichen Zhang and Ali A Ghorbani. “An overview of online fake news: Characterization, detection, and discussion”. In: *Information Processing & Management* 57.2 (2020), p. 102025.

- [48] Nicholas Difonzo. "Conspiracy rumor psychology". In: *Conspiracy theories and the people who believe them* (2018).
- [49] Nicholas DiFonzo. "Rumor". In: *The Corsini Encyclopedia of Psychology* (2010), pp. 1–2.
- [50] Prashant Bordia et al. "Rumor as Revenge in the Workplace". In: *Group & Organization Management* 39.4 (2014), pp. 363–388.
- [51] Robin IM Dunbar. "Gossip in evolutionary perspective". In: *Review of general psychology* 8.2 (2004), pp. 100–110.
- [52] Gary Alan Fine and Ralph L Rosnow. "Gossip, gossipers, gossiping". In: *Personality and social psychology bulletin* 4.1 (1978), pp. 161–168.
- [53] Nicholas DiFonzo and Prashant Bordia. "Rumor, gossip and urban legends". In: *Diogenes* 54.1 (2007), pp. 19–35.
- [54] Jowett, G. S., & Heath, R. L. "Propaganda". In: *The Encyclopedia of Public Relations* (2004), pp. 652–656.
- [55] Samantha Bradshaw and Philip N Howard. *The global disinformation order: 2019 global inventory of organised social media manipulation*. Project on Computational Propaganda, 2019.
- [56] Sarah J Jackson and Brooke Foucault Welles. "Hijacking# myNYPD: Social media dissent and networked counterpublics". In: *Journal of Communication* 65.6 (2015), pp. 932–952.
- [57] Nicholas DiFonzo. "Propaganda". In: *The Corsini Encyclopedia of Psychology* (2010), pp. 1–3.
- [58] Truda Gray and Brian Martin. "Backfires: white, black and grey". In: *Journal of Information Warfare* 6.1 (2007), pp. 7–16.
- [59] David W Guth. "Black, white, and shades of gray: The sixty-year debate over propaganda versus public diplomacy". In: *Journal of Promotion Management* 14.3-4 (2009), pp. 309–325.
- [60] L John Martin. "Disinformation: An instrumentality in the propaganda arsenal". In: *Political Communication* 2.1 (1982), pp. 47–64.
- [61] Herbert Romerstein. "Disinformation as a KGB Weapon in the Cold War". In: *Journal of Intelligence History* 1.1 (2001), pp. 54–67.
- [62] Nicholas O'Shaughnessy. "From Disinformation to Fake News: Forwards into the Past". In: *The SAGE Handbook of Propaganda* (2019), p. 55.
- [63] Britt Paris and Joan Donovan. "Deepfakes and Cheap Fakes". In: *United States of America: Data & Society* (2019).
- [64] Jan-Willem Van Prooijen. *The psychology of conspiracy theories*. Routledge, 2018.
- [65] Michael Barkun. *A culture of conspiracy: Apocalyptic visions in contemporary America*. Vol. 15. Univ of California Press, 2013.
- [66] Eirikur Bergmann. *Conspiracy & populism: The politics of misinformation*. Springer, 2018.

- [67] Vincent F Hendricks and Mads Vestergaard. “Fact Resistance, Populism, and Conspiracy Theories”. In: *Reality Lost*. Springer, 2019, pp. 79–101.
- [68] Yochai Benkler, Robert Faris, and Hal Roberts. *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press, 2018.
- [69] Cook, John. *A history of FLICC: the 5 techniques of science denial*. 2020. URL: <https://www.skepticalscience.com/history-FLICC-5-techniques-science-denial.html>.
- [70] Alberto Acerbi. “Cognitive attraction and online misinformation”. In: *Palgrave Communications* 5.1 (2019), pp. 1–7.
- [71] Standage, Tom. *The true history of fake news*. 2017. URL: <https://www.1843magazine.com/technology/rewind/the-true-history-of-fake-news>.
- [72] Julie Posetti and Alice Matthews. “A short guide to the history of ‘fake news’ and disinformation”. In: *International Center for Journalists* 7 (2018).
- [73] CiTS. *A Brief History of Fake News*. 2018. URL: <https://www.cits.ucsb.edu/fake-news/brief-history>.
- [74] Sven Ove Hansson. “Science denial as a form of pseudoscience”. In: *Studies in History and Philosophy of Science Part A* 63 (2017), pp. 39–47.
- [75] Philipp Schmid and Cornelia Betsch. “Effective strategies for rebutting science denialism in public discussions”. In: *Nature Human Behaviour* 3.9 (2019), pp. 931–939.
- [76] J Cook et al. *Denial101x: Making sense of climate science denial*. edX. 2015.
- [77] Sven Ove Hansson. “Dealing with climate science denialism: experiences from confrontations with other forms of pseudoscience”. In: *Climate Policy* 18.9 (2018), pp. 1094–1102.
- [78] Steven J Frenda, Rebecca M Nichols, and Elizabeth F Loftus. “Current issues and advances in misinformation research”. In: *Current Directions in Psychological Science* 20.1 (2011), pp. 20–23.
- [79] Michael J Mazarr et al. *Hostile Social Manipulation Present Realities and Emerging Trends*. Tech. rep. RAND National Defense Research Inst Santa Monica Ca Santa Monica United States, 2019.
- [80] Prashant Bordia and Nicholas DiFonzo. “Problem solving in social interactions on the Internet: Rumor as social cognition”. In: *Social Psychology Quarterly* 67.1 (2004), pp. 33–49.
- [81] Derek Ruths. “The misinformation machine”. In: *Science* 363.6425 (2019), pp. 348–348.
- [82] Zohar Erez et al. “Communication between viruses guides lysis–lysogeny decisions”. In: *Nature* 541.7638 (2017), pp. 488–493.

- [83] Dario-Marcos Bayani, Michael Taborsky, and Joachim G Frommen. "To pee or not to pee: urine signals mediate aggressive interactions in the cooperatively breeding cichlid *Neolamprologus pulcher*". In: *Behavioral Ecology and Sociobiology* 71.2 (2017), p. 37.
- [84] H Kasozi and RA Montgomery. "How do giraffes locate one another? A review of visual, auditory, and olfactory communication among giraffes". In: *Journal of Zoology* 306.3 (2018), pp. 139–146.
- [85] Robbins, S and Judge, T. *Organisational Behaviour*. Pearson Prentice Hall, 2007.
- [86] John Fiske. *Introduction to communication studies*. Routledge, 2010.
- [87] Sheila Steinberg. *An introduction to communication studies*. Juta and Company Ltd, 2007.
- [88] Stephen W Littlejohn and Karen A Foss. *Theories of human communication*. Waveland press, 2010.
- [89] F Dance and C Larsen. *The functions of Human Communication. A Theoretical Approach*. Holt, Rinehart and Winston, 1976.
- [90] Sune Lehmann. "Fundamental Structures in Temporal Communication Networks". In: *Temporal Network Theory*. Springer, 2019, pp. 25–48.
- [91] Prakash Chakravarthi. "The history of communications-from cave drawings to mail messages". In: *IEEE Aerospace and Electronic Systems Magazine* 7.4 (1992), pp. 30–35.
- [92] Eltjo Buringh and Jan van Zanden. "Charting the "Rise of the West": Manuscripts and Printed Books in Europe, a long-term Perspective from the Sixth through Eighteenth Centuries". In: *The Journal of Economic History* 69.2 (2009), pp. 409–445.
- [93] Soll, Jacob. *The Long and Brutal History of Fake News*. 2016. URL: <https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535>.
- [94] Joachim Neander and Randal Marlin. "Media and Propaganda: The Northcliffe Press and the Corpse Factory Story of World War I". In: *Global Media Journal* 3.2 (2010), p. 67.
- [95] BBC. *The corpse factory and the birth of fake news*. 2017. URL: <https://www.bbc.com/news/entertainment-arts-38995205>.
- [96] Ellic Howe. *The black game: British subversive operations against the Germans during the Second World War*. Michael Joseph, 1982.
- [97] Egon Larsen. "Now it can be told". In: *AJR Information* 40.9 (1985).
- [98] Nicholas Rankin. *A genius for deception: how cunning helped the British win two world wars*. Oxford University Press, 2009.
- [99] BBC World. *Britain's secret propaganda war*. 2019. URL: <https://www.bbc.co.uk/programmes/w3csyx52>.

- [100] Mann, R. *Daisy petals and mushroom clouds: LBJ, Barry Goldwater, and the ad that changed American politics*. 2011. URL: <https://www.bbc.com/news/entertainment-arts-38995205>.
- [101] Mann, R. *How the "Daisy" Ad Changed Everything About Political Advertising*. 2017. URL: <https://www.smithsonianmag.com/history/how-daisy-ad-changed-everything-about-political-advertising-180958741/>.
- [102] CBS. *Memorable Campaign Ads*. 2017. URL: <https://www.cbsnews.com/pictures/memorable-campaign-ads/5/>.
- [103] Cris, D. *This is the 30-year-old Willie Horton ad everybody is talking about today*. 2018. URL: <https://edition.cnn.com/2018/11/01/politics/willie-horton-ad-1988-explainer-trnd/index.html>.
- [104] FiveThirtyEight. *How To Destroy A Presidential Candidate*. 2018. URL: <https://fivethirtyeight.com/features/film-how-to-destroy-a-presidential-candidate/>.
- [105] Dowe, Tom. *News You Can Abuse*. 1997. URL: <https://www.wired.com/1997/01/netizen-6/>.
- [106] Hussain, Murtaza. *An Iranian Activist Wrote Dozens of Articles for Right-Wing Outlets. But Is He a Real Person?* 2019. URL: <https://theintercept.com/2019/06/09/heshmat-alavi-fake-iran-mek/>.
- [107] Jos Van Bommel. "Rumors". In: *The Journal of Finance* 58.4 (2003), pp. 1499–1520.
- [108] Daniel B Hoch et al. "Information exchange in an epilepsy forum on the World Wide Web". In: *Seizure* 8.1 (1999), pp. 30–34.
- [109] Rajesh K Aggarwal and Guojun Wu. "Stock market manipulations". In: *The Journal of Business* 79.4 (2006), pp. 1915–1953.
- [110] William M Silberg, George D Lundberg, and Robert A Musacchio. "Assessing, controlling, and assuring the quality of medical information on the Internet". In: *Jama* 277.15 (1997), pp. 1244–1245.
- [111] Lisa A Burke and Jessica Morris Wise. "The effective care, handling and pruning of the office grapevine". In: *Business Horizons* 46.3 (2003), pp. 71–76.
- [112] M Earley. "Report On Cyberstalking: A New Challenge For Law Enforcement And Industry". In: *Washington DC: United States Department of Justice. Retrieved January 4* (1999), p. 2005.
- [113] Haewoon Kwak et al. "What is Twitter, a social network or a news media?" In: *Proceedings of the 19th international conference on World wide web*. 2010, pp. 591–600.
- [114] Meeyoung Cha et al. "Measuring user influence in twitter: The million follower fallacy." In: *Icwsm* 10.10-17 (2010), p. 30.
- [115] Johan Ugander et al. "The anatomy of the facebook social graph". In: *arXiv preprint arXiv:1111.4503* (2011).
- [116] Albert-László Barabási et al. *Network science*. Cambridge university press, 2016.

- [117] Rowena L Briones et al. "Keeping up with the digital age: How the American Red Cross uses social media to build relationships". In: *Public relations review* 37.1 (2011), pp. 37–43.
- [118] Edward L Schor. "Developing communality: family-centered programs to improve children's health and well-being." In: *Bulletin of the New York Academy of Medicine* 72.2 (1995), p. 413.
- [119] J Scott Brennen et al. "Types, sources, and claims of Covid-19 misinformation". In: *Reuters Institute* 7 (2020), pp. 3–1.
- [120] Tarlach McGonagle. "'Fake news' False fears or real concerns?" In: *Netherlands Quarterly of Human Rights* 35.4 (2017), pp. 203–209.
- [121] Kimberly M Christopherson. "The positive and negative implications of anonymity in Internet social interactions: 'On the Internet, Nobody Knows You're a Dog'". In: *Computers in Human Behavior* 23.6 (2007), pp. 3038–3056.
- [122] Zeynep Tufekci. *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press, 2017.
- [123] Kalliopi Kyriakopoulou. "Authoritarian states and internet social media: Instruments of democratisation or instruments of control?" In: *Human Affairs* 21.1 (2011), pp. 18–26.
- [124] Roth, Yoel and Pickles, N. *Updating our Approach to Misleading Information*. 2020. URL: https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html.
- [125] Dietmar Jannach and Michael Jugovac. "Measuring the business value of recommender systems". In: *ACM Transactions on Management Information Systems (TMIS)* 10.4 (2019), pp. 1–23.
- [126] Google. *Recommendations: What and Why?* 2019. URL: <https://developers.google.com/machine-learning/recommendation/overview>.
- [127] Shlomo Berkovsky and Jill Freyne. "Web personalization and recommender systems". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015, pp. 2307–2308. URL: <https://doi.org/10.1145/2783258.2789995>.
- [128] Rajnikant Bhagwan Wagh and Jayantrao Bhaurao Patil. "Web personalization and recommender systems: An overview". In: *Proceedings of the 18th International Conference on Management of Data*. 2012, pp. 114–114.
- [129] Emilio Ferrara et al. "The rise of social bots". In: *Communications of the ACM* 59.7 (2016), pp. 96–104.
- [130] Sandra C Matz et al. "Psychological targeting as an effective approach to digital mass persuasion". In: *Proceedings of the national academy of sciences* 114.48 (2017), pp. 12714–12719.
- [131] Centre for Data Ethics and Innovation. *Review of online targeting: Final report and recommendations*. 2020. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/864167/CDEJ7836-Review-of-Online-Targeting-05022020.pdf.

- [132] van der Linden, S and Yasseri T and Watts D. *Discussion on Twitter*. 2020. URL: https://twitter.com/Sander_vdLinden/status/1247146391002132480.
- [133] Zeynep Tufekci. *YouTube, the Great Radicalizer*. 2018. URL: <https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html>.
- [134] Max Fisher and Amanda Taub. *On YouTube's Digital Playground, an Open Gate for Pedophiles*. 2019. URL: <https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html>.
- [135] Manoel Horta Ribeiro et al. "Auditing radicalization pathways on YouTube". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 131–141.
- [136] Chaslot, G. *Exploring YouTube recommendations*. 2016. URL: <https://github.com/pnbt/youtube-explore>.
- [137] Colin Klein, Peter Clutton, and Adam G Dunn. "Pathways to conspiracy: The social and linguistic precursors of involvement in Reddit's conspiracy theory forum". In: *PLoS one* 14.11 (2019), e0225098.
- [138] A Campbell. *Proud Boy Allegedly Murders Brother With A Sword Thinking He's Killing A Lizard*. 2019. URL: https://www.huffpost.com/entry/proud-boy-allegedly-murders-brother-with-a-sword-thinking-hes-a-lizard_n_5c36042ee4b05b16bcfcb3d5?guccounter=1.
- [139] Sergey Brin and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine". In: *Proceedings of the Seventh World Wide Web Conference*. 1998.
- [140] Edward Loper and Steven Bird. "NLTK: the natural language toolkit". In: *arXiv preprint cs/0205028* (2002).
- [141] D. J. Daley and D. G. Kendall. "Epidemics and Rumours". In: *Nature* 204.4963 (Dec. 1964), pp. 1118–1118. DOI: 10.1038/2041118a0.
- [142] DW. *German justice minister to set up task force on Internet hate speech*. 2015. URL: <https://www.dw.com/en/german-justice-minister-to-set-up-task-force-on-internet-hate-speech/a-18714334>.
- [143] DW. *German justice minister: AfD uses hate speech online*. 2016. URL: <https://www.dw.com/en/german-justice-minister-afd-uses-hate-speech-online/a-35961820>.
- [144] DW. *German opposition parties call to replace online hate speech law*. 2018. URL: <https://www.dw.com/en/german-opposition-parties-call-to-replace-online-hate-speech-law/a-42058030>.
- [145] DW. *German populist groups AfD, Pegida look for common ground*. 2018. URL: <https://www.dw.com/en/german-populist-groups-afd-pegida-look-for-common-ground/a-18180581>.
- [146] Connolly, Kate. *Germany greets refugees with help and kindness at Munich central station*. 2015. URL: <https://www.theguardian.com/world/2015/sep/03/germany-refugees-munich-central-station>.

- [147] DW. *Charlie Hebdo attack: German lawmakers warn against xenophobia*. 2015. URL: <https://www.dw.com/en/charlie-hebdo-attack-german-lawmakers-warn-against-xenophobia/a-18180826>.
- [148] Barzic, Gwenaelle and Kar-Gupta, Sudip. *Facebook, Google join drive against fake news in France*. 2017. URL: <https://www.reuters.com/article/us-france-election-facebook/facebook-google-join-drive-against-fake-news-in-france-idUSKBN15LOQU>.
- [149] The Local. *Battle begins to stop 'fake news' from impacting the French presidential election*. 2017. URL: <https://www.thelocal.fr/20170228/can-could-fake-news-influence-the-french-election>.
- [150] Bashir, wais. *Fact-checking fake news in the French election*. 2017. URL: <https://www.bbc.com/news/world-europe-39495635>.
- [151] Funke, Daniel and Flamini, Daniela. *A guide to anti-misinformation actions around the world*. 2018. URL: <https://www.poynter.org/ifcn/anti-misinformation-actions/>.
- [152] Fung, Brian and Garcia, Ahiza. *Facebook has shut down 5.4 billion fake accounts this year*. 2019. URL: <https://edition.cnn.com/2019/11/13/tech/Facebook-fake-accounts/index.html>.
- [153] Elizabeth Anne Bodine-Baron et al. *Countering Russian social media influence*. RAND Corporation Santa Monica, 2018.
- [154] Herbert A Simon. "Designing organizations for an information-rich world". In: *International Library of Critical Writings in Economics* 70 (1996), pp. 187–202.
- [155] Lyons, Tessa. *Hard Questions: What's Facebook's Strategy for Stopping False News?* 2018. URL: <https://newsroom.fb.com/news/2018/05/hard-questions-false-news/>.
- [156] Harvey, Del. *Serving healthy conversation*. 2018. URL: https://blog.twitter.com/en_us/topics/product/2018/Serving_Healthy_Conversation.html.
- [157] Google. *How Google Fights Disinformation*. 2018. URL: https://storage.googleapis.com/gweb-uniblog-publish-prod/documents/How_Google_Fights_Disinformation.pdf.
- [158] Silverman, Henry. *The Next Phase in Fighting Misinformation*. 2019. URL: <https://newsroom.fb.com/news/2019/04/tackling-more-false-news-more-quickly-https-newsroom-fb-com-news-2019-04-tackling-more-false-news-more-quickly/>.
- [159] Mosseri, Adam. *Working to Stop Misinformation and False News*. 2017. URL: <https://newsroom.fb.com/news/2017/04/working-to-stop-misinformation-and-false-news/>.
- [160] Policy team. *Update on Twitter's review of the 2016 US election*. 2018. URL: https://blog.twitter.com/en_us/topics/company/2018/2016-election-update.html.

- [161] White, Karen. *Improving health during Global Media and Information Literacy Week 2018*. 2018. URL: https://blog.twitter.com/en_us/topics/company/2018/Global-MIL-Week-2018.html.
- [162] Governance team. *We're focused on serving the public conversation*. 2018. URL: https://about.twitter.com/en_us/values/elections-integrity.html#partnerships.
- [163] *Google News Initiative*. URL: <https://newsinitiative.withgoogle.com/>.
- [164] Claes Wohlin et al. "The success factors powering industry-academia collaboration". In: *IEEE software* 29.2 (2011), pp. 67–73.
- [165] Omar Al-Tabbaa and Samuel Ankrah. "Engineered University-Industry Collaboration: A Social Capital Perspective". In: *European Management Review* 16.3 (2019), pp. 543–565.
- [166] Archibong, Ime. *API and Other Platform Product Changes*. 2018. URL: <https://developers.facebook.com/blog/post/2018/04/04/facebook-api-platform-product-changes/>.
- [167] Jacob Kastrenakes. *WhatsApp limits message forwarding in fight against misinformation*. 2019. URL: <https://www.theverge.com/2019/1/21/18191455/whatsapp-forwarding-limit-five-messages-misinformation-battle>.
- [168] Fabry, Merrill. *Here's How the First Fact-Checkers Were Able to Do Their Jobs Before the Internet*. 2017. URL: <https://time.com/4858683/fact-checking-history/>.
- [169] Lucas Graves. *Deciding what's true: The rise of political fact-checking in American journalism*. Columbia University Press, 2016.
- [170] N. DiFonzo, P. Bordia, and R. L. Rosnow. "Reining in rumors". In: *Organizational Dynamics* 23.1 (June 1994), pp. 47–62. DOI: 10.1016/0090-2616(94)90087-6.
- [171] Cathy Faye. "Governing the grapevine: The study of rumor during World War II." In: *History of psychology* 10.1 (2007), p. 1.
- [172] Stephen Young, Alasdair Pinkerton, and Klaus Dodds. "The word on the street: Rumor, "race" and the anticipation of urban unrest". In: *Political Geography* 38 (2014), pp. 57–67.
- [173] Marin Lessenski. *Common Sense Wanted: Resilience to 'Post-Truth' and its Predictors in the New Media Literacy Index 2018*. Tech. rep. Open Society Institute, 2018. URL: https://osis.bg/wp-content/uploads/2018/04/MediaLiteracyIndex2018_publishENG.pdf.
- [174] Sander Van der Linden et al. "Inoculating the public against misinformation about climate change". In: *Global Challenges* 1.2 (2017), p. 1600008.
- [175] Sander Van Der Linden et al. "Inoculating against misinformation". In: *Science* 358.6367 (2017), pp. 1141–1142.
- [176] Jon Roozenbeek and Sander van der Linden. "The fake news game: actively inoculating against the risk of misinformation". In: *Journal of Risk Research* 22.5 (2019), pp. 570–580.

- [177] John Cook, Peter Ellerton, and David Kinkead. “Deconstructing climate misinformation to identify reasoning errors”. In: *Environmental Research Letters* 13.2 (2018), p. 024018.
- [178] Adam Kucharski. “Study epidemiology of fake news”. In: *Nature* 540.7634 (Dec. 2016), pp. 525–525.
- [179] William Goffman and Vaun A. Newill. “Generalization of Epidemic Theory: An Application to the Transmission of Ideas”. In: *Nature* 204.4955 (Oct. 1964), pp. 225–228.
- [180] Yamir Moreno, Maziar Nekovee, and Amalio F Pacheco. “Dynamics of rumor spreading in complex networks”. In: *Physical Review E* 69.6 (June 2004), p. 066130.
- [181] Daryl J. Daley and Joe. Gani. *Epidemic modelling : an introduction*. Cambridge University Press, 2001, p. 213.
- [182] Shuja Shafi et al. “Hajj: health lessons for mass gatherings”. In: *Journal of infection and public health* 1.1 (2008), pp. 27–32.
- [183] Gabriel M Leung and Angus Nicoll. “Reflections on pandemic (H1N1) 2009 and the international response”. In: *PLoS Med* 7.10 (2010), e1000346.
- [184] Fuminori Kato et al. “Combined effects of prevention and quarantine on a break-out in SIR model”. In: *Scientific reports* 1 (2011), p. 10.
- [185] Roger Magnusson. “Advancing the right to health: the vital role of law”. In: *Advancing the Right to Health: The Vital Role of Law, World Health Organization, Switzerland* (2017).
- [186] TA Ruff. “Immunisation strategies for viral diseases in developing countries”. In: *Reviews in Medical Virology* 9.2 (1999), pp. 121–138.
- [187] Dietram A Scheufele and Nicole M Krause. “Science audiences, misinformation, and fake news”. In: *Proceedings of the National Academy of Sciences* 116.16 (2019), pp. 7662–7669.
- [188] D Rotolo and D Hicks. “What is an emerging technology?” In: *Research Policy* 10.44 (2015), pp. 1827–1843.
- [189] Thomas S. Kuhn. *The structure of scientific revolutions*. University of Chicago Press, 1970, p. 210. ISBN: 0226458083.
- [190] Donald E. Stokes. *Pasteur’s quadrant : basic science and technological innovation*. Brookings Institution Press, 1997, p. 180.
- [191] J Alexander, J Chase, and N Newman. “Emergence as a conceptual framework for understanding scientific and technological progress”. In: *Technology Management for Emerging Technologies (PICMET)*. IEEE, 2012.
- [192] Stephen F. Carley et al. “An indicator of technical emergence”. In: *Scientometrics* 115.1 (Apr. 2018), pp. 35–49.
- [193] Philip Shapira, Seokbeom Kwon, and Jan Youtie. “Tracking the emergence of synthetic biology”. In: *Scientometrics* 112.3 (Sept. 2017), pp. 1439–1469.

- [194] Iowa State University Library. *Scopus vs. Web of Science vs. Google Scholar*. 2018. URL: <http://instr.iastate.libguides.com/c.php?g=120420&p=785310>.
- [195] M. Callon et al. "From translations to problematic networks: An introduction to co-word analysis". In: *Social Science Information* 22.2 (Mar. 1983), pp. 191–235.
- [196] E Garfield, IH Sher, and RJ Torpie. *The use of citation data in writing the history of science*. 1964, p. 75.
- [197] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3.Jan (2003), pp. 993–1022.
- [198] Santo Fortunato. "Community detection in graphs". In: *Physics Reports* 486.3-5 (Feb. 2010), pp. 75–174.
- [199] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (Oct. 2008).
- [200] Scott Emmons et al. "Analysis of Network Clustering Algorithms and Cluster Quality Metrics at Scale". In: *PLOS ONE* 11.7 (July 2016). Ed. by Constantine Dovrolis, e0159161.
- [201] Pravin Chopade and Justin Zhan. "A Framework for Community Detection in Large Networks Using Game-Theoretic Modeling". In: *IEEE Transactions on Big Data* 3.3 (Sept. 2017), pp. 276–288.
- [202] S. Kwon, M. Cha, and K. Jung. "Rumor Detection over Varying Time Windows". In: *PLOS ONE* 12.1 (Jan. 2017). Ed. by Zhong-Ke Gao, e0168344. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0168344.
- [203] G. Liang et al. "Rumor Identification in Microblogging Systems Based on Users' Behavior". In: *IEEE Transactions on Computational Social Systems* 2.3 (Sept. 2015), pp. 99–108. ISSN: 2329-924X. DOI: 10.1109/TCSS.2016.2517458. URL: <http://ieeexplore.ieee.org/document/7419281/>.
- [204] A. Zubiaga, M. Liakata, and R. Procter. "Learning Reporting Dynamics during Breaking News for Rumour Detection in Social Media". In: (Oct. 2016). URL: <http://arxiv.org/abs/1610.07363>.
- [205] F. Yang et al. "Automatic detection of rumor on Sina Weibo". In: *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*. ACM, 2012.
- [206] A. Zubiaga et al. "Detection and Resolution of Rumours in Social Media". In: *ACM Computing Surveys* 51.2 (Feb. 2018), pp. 1–36. DOI: 10.1145/3161603.
- [207] R. Sicilia et al. "Twitter rumour detection in the health domain". In: *Expert Systems with Applications* 110 (Nov. 2018), pp. 33–40. ISSN: 0957-4174. DOI: 10.1016/J.ESWA.2018.05.019.
- [208] A. Zubiaga et al. "Crowdsourcing the Annotation of Rumourous Conversations in Social Media". In: *Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion*. New York, New York, USA: ACM Press, 2015, pp. 347–353. ISBN: 9781450334730. DOI: 10.1145/2740908.2743052.

- [209] A. Zubiaga et al. “Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads”. In: *PLOS ONE* 11.3 (Mar. 2016). Ed. by Naoki Masuda, e0150989. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0150989.
- [210] J. H. M. Janssens. “Outlier selection and one-class classification”. PhD thesis. Tilburg University, 2013.
- [211] S. Khan and M. G. Madden. “One-class classification: taxonomy of study and review of techniques”. In: *The Knowledge Engineering Review* 29.03 (June 2014), pp. 345–374. DOI: 10.1017/S026988891300043X.
- [212] D. M. J. Tax. “One-class Classification: concept-learning in the absence of counter-examples”. PhD thesis. Delft University of Technology, 2001.
- [213] S. Kwon et al. “Prominent features of rumor propagation in online social media”. In: *International Conference on Data Mining*. IEEE, 2013.
- [214] C. Castillo, M. Mendoza, and B. Poblete. “Information credibility on twitter”. In: *Proceedings of the 20th international conference on World wide web - WWW '11*. New York, New York, USA: ACM Press, 2011, p. 675. ISBN: 9781450306324. DOI: 10.1145/1963405.1963500.
- [215] J. Ma et al. “Detecting Rumors from Microblogs with Recurrent Neural Networks”. In: *IJCAI*. 2016, pp. 3818–3824.
- [216] V. C. Raykar et al. “Learning From Crowds”. In: *Journal of Machine Learning Research* 11.Apr (2010), pp. 1297–1322.
- [217] Sardar Hamidian and Mona T Diab. “Rumor detection and classification for twitter data”. In: *arXiv preprint arXiv:1912.08926* (2019).
- [218] S. Vosoughi, M. N. Mohsenvand, and D. Roy. “Rumor gauge: predicting the veracity of rumors on twitter”. In: *ACM Transactions on Knowledge Discovery from Data* 11.4 (July 2017), pp. 1–36. ISSN: 15564681. DOI: 10.1145/3070644.
- [219] J. Kim et al. “Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation”. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM '18*. New York, New York, USA: ACM Press, 2018, pp. 324–332. ISBN: 9781450355810. DOI: 10.1145/3159652.3159734.
- [220] S. A. Alkhodair et al. “Detecting breaking news rumors of emerging topics in social media”. In: *Information Processing & Management* (Feb. 2019). DOI: 10.1016/J.IPM.2019.02.016.
- [221] O. Ajao, D. Bhowmik, and S. Zargari. “Fake News Identification on Twitter with Hybrid CNN and RNN Models”. In: *Proceedings of the 9th International Conference on Social Media and Society - SMSociety '18*. New York, New York, USA: ACM Press, 2018, pp. 226–230. ISBN: 9781450363341. DOI: 10.1145/3217804.3217917.

- [222] S. Das Bhattacharjee, A. Talukder, and B. V. Balantrapu. "Active learning based news veracity detection with feature weighting and deep-shallow fusion". In: *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, Dec. 2017, pp. 556–565. ISBN: 978-1-5386-2715-0. DOI: 10.1109/BigData.2017.8257971.
- [223] V. Qazvinian et al. "Rumor has it: identifying misinformation in microblogs". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2011), pp. 1589–1599.
- [224] Sardar Hamidian and Mona Diab. "Rumor identification and belief investigation on twitter". In: *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 2016, pp. 3–8.
- [225] M Johnson Vioulès et al. "Detection of suicide-related posts in Twitter data streams". In: *IBM Journal of Research and Development* 62.1 (2018), pp. 7–1.
- [226] Antigoni Maria Founta et al. "A unified deep learning architecture for abuse detection". In: *Proceedings of the 10th ACM Conference on Web Science*. ACM, 2019, pp. 105–114.
- [227] Myle Ott et al. "Finding deceptive opinion spam by any stretch of the imagination". In: *arXiv preprint arXiv:1107.4557* (2011).
- [228] A Zubiaga, M. Liakata, and R. Procter. "Exploiting context for rumour detection in social media". In: *International Conference on Social Informatics*. Springer, 2017, pp. 109–123.
- [229] D. Crystal. *Language and the Internet*. Cambridge: Cambridge University Press, 2006. ISBN: 9780511487002. DOI: 10.1017/CB09780511487002.
- [230] Q. Zhang et al. "Automatic detection of rumor on social network". In: *Natural Language Processing and Chinese Computing*. 2015, pp. 113–122.
- [231] S. Wijeratne et al. "Feature Engineering for Twitter-based Applications". In: *Feature Engineering for Machine Learning and Data Analytics*. 2017, pp. 359–384.
- [232] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer. "Psychological Aspects of Natural Language Use: Our Words, Our Selves". In: *Annual Review of Psychology* 54.1 (Feb. 2003), pp. 547–577. ISSN: 0066-4308. DOI: 10.1146/annurev.psych.54.101601.145041.
- [233] K. Wu, S. Yang, and K. Q. Zhu. "False rumors detection on Sina Weibo by propagation structures". In: *2015 IEEE 31st International Conference on Data Engineering*. IEEE, Apr. 2015, pp. 651–662. ISBN: 978-1-4799-7964-6. DOI: 10.1109/ICDE.2015.7113322.
- [234] O. Varol et al. "Feature Engineering for Social Bot Detection". In: *Feature Engineering for Social Bot Detection*. CRC Press, Mar. 2018, pp. 311–334. DOI: 10.1201/9781315181080-12.
- [235] D. A. Bird. *Rumor as Folklore: An Interpretation and Inventory*. 1979, p. 648.
- [236] D. A. Bird, S. C. Holder, and D. Sears. "Walrus is Greek for Corpse: Rumor and the Death of Paul McCartney". In: *The Journal of Popular Culture* X.1 (June 1976), pp. 110–121. ISSN: 00223840. DOI: 10.1111/j.0022-3840.1976.1001{_}110.x.

- [237] B. Schölkopf et al. "Support vector method for novelty detection". In: *Advances in neural information processing systems*. 2000, pp. 582–588.
- [238] D. M. J. Tax and R. P. W. Duin. "Support Vector Data Description". In: *Machine Learning* 54.1 (Jan. 2004), pp. 45–66. ISSN: 0885-6125. DOI: 10.1023/B:MACH.0000008084.60811.49.
- [239] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.Oct (2011), pp. 2825–2830. ISSN: ISSN 1533-7928. URL: <http://www.jmlr.org/papers/v12/pedregosa11a.html>.
- [240] R. P. W. Duin et al. *PRTTools: A matlab toolbox for pattern recognition*. 2000.
- [241] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008, p. 482. ISBN: 0521865719.
- [242] Amir Ebrahimi Fard et al. "Computational Rumor Detection Without Non-Rumor: A One-Class Classification Approach". In: *IEEE Transactions on Computational Social Systems* 6.5 (2019), pp. 830–846.
- [243] Clara Maathuis, Wolter Pieters, and Jan Van Den Berg. "Cyber weapons: a profiling framework". In: *2016 International Conference on Cyber Conflict (CyCon US)*. IEEE. 2016, pp. 1–8.
- [244] C Maathuis, W Pieters, and J van den Berg. "Developing a Cyber Operations Computational Ontology". In: *Journal of Information Warfare* 17.3 (2018), pp. 32–49.
- [245] Nimmo, Ben., Francois, Camille., Eib, Shawn., Ronzaud, Lea., Ferreira, Rodrigo., Hernon., Chris, and Tim Kostelancik. *Exposing Secondary Infection*. 2020. URL: <https://secondaryinfection.org/>.
- [246] Saddique, Hussein. *Factory of Lies Democracy Under Attack*. 2018. URL: <https://www.nbcnews.com/video/nbc-news-signal-presents-factory-of-lies-democracy-under-attack-1362496579619>.
- [247] Ellick, Adam B and Westbrook, Adam. *Operation InfeKtion: Russian Disinformation from Cold War to Kanye*. 2018. URL: <https://www.nytimes.com/2018/11/12/opinion/russia-meddling-disinformation-fake-news-elections.html>.
- [248] C Maathuis, W Pieters, and J van den Berg. "A Knowledge-Based Model for Assessing the Effects of Cyber Warfare". In: *Proceedings of the 12th NATO Conference on Operations Research and Analysis*. 2018.
- [249] Sean Bechhofer et al. "OWL web ontology language reference". In: *W3C recommendation* 10.02 (2004).
- [250] Ahlam Sawsaa and Joan Lu. "Building information science ontology (OIS) with methontology and protégé". In: *Journal of Internet Technology and Secured Transactions (JITST)* 1.3/4 (2012).
- [251] Denny Vrandečić. "Ontology evaluation". In: *Handbook on ontologies*. Springer, 2009, pp. 293–313.

- [252] Active Measures Working Group et al. "Soviet influence activities: a report on active measures and propaganda, 1986-87". In: *Washington, DC: US Department of State* (1987), pp. 33–35.
- [253] Richard H Shultz and Roy Godson. *Dezinformatsia: Active measures in Soviet strategy*. Potomac Books, 1984.
- [254] Thomas Rid. "Disinformation: A primer in Russian active measures and influence campaigns". In: *Hearings before the Select Committee on Intelligence, United States Senate, One Hundred Fifteenth Congress*. Vol. 30. 2017.
- [255] Al Jazeera News. *US-Iran tensions, trolls and the dubious case of Heshmat Alavi*. 2019. URL: <https://www.aljazeera.com/programmes/listeningpost/2019/06/iran-tensions-trolls-dubious-case-heshmat-alavi-190616075641073.html>.
- [256] Rezaian, J. *Why does the U.S. need trolls to make its Iran case?* 2019. URL: <https://www.washingtonpost.com/opinions/2019/06/11/why-does-us-need-trolls-make-its-iran-case/>.
- [257] Camille François, Ben Nimmo, and C Shawn Eib. "The IRA CopyPasta Campaign". In: *Graphika, okt* (2019).
- [258] *Information Operations*. Tech. rep. Joint Chiefs of Staff, 2014.
- [259] Jon Roozenbeek and Sander van der Linden. "Fake news game confers psychological resistance against online misinformation". In: *Palgrave Communications* 5.1 (2019), pp. 1–10.

LIST OF PUBLICATIONS

The followings list presents my publications during PhD, some of which formed the major parts of this dissertation.

PUBLICATIONS

- **Fard, A. E.**, Mohammadi, M., Chen, Y., & Van de Walle, B. (2019). *Computational Rumor Detection Without Non-Rumor: A One-Class Classification Approach*. IEEE Transactions on Computational Social Systems, 6(5), 830-846.
- **Fard, A. E.**, & Cunningham, S. (2019). *Assessing the Readiness of Academia in the Topic of False and Unverified Information*. ACM Journal of Data and Information Quality (JDIQ), 11(4), 1-27.
- Alfano, M., **Fard, A. E.**, Carter, J. A., Clutton, P., & Klein, C. (2020). *Technologically scaffolded atypical cognition: The case of YouTube's recommender system*. Synthese.
- **Fard, A. E.**, & Lingeswaran, S. (2020, April). *Misinformation Battle Revisited: Counter Strategies from Clinics to Artificial Intelligence*. In Companion Proceedings of the Web Conference 2020 (WWW '20) (pp. 510–519). Association for Computing Machinery (ACM).
- **Fard, A. E.**, Mohammadi, M., Cunningham, S., & Van de Walle, B. (2019, June). *Rumour As an Anomaly: Rumour Detection with One-Class Classification*. In 2019 IEEE International Conference on Engineering, Technology and Innovation (pp. 1-9). IEEE.
- **Fard, A. E.**, Mercuur, R., Dignum, V., Jonker, C. M., & van de Walle, B. (2020). *Towards Agent-based Models of Rumours in Organizations: A Social Practice Theory Approach*. In Advances in Social Simulation (pp. 141-153). Springer, Cham.
- **Fard, A. E.**, Mohammadi, M., Cunningham, S., & Van de Walle, B. *Detecting Rumours in Disasters: An Imbalanced Learning Approach*. In International Conference on Computational Science. Springer.
- Hazenburg, H., van den Hoven, M. J., Cunningham, S., Alfano, M., Ashgari, H., Sullivan, E., **Fard, A. E.**, & Rodriguez, E. T. (2018). *Micro-Targeting and ICT media in the Dutch Parliamentary system: Technological changes in Dutch Democracy*.
- Sebastian, A. G., Lending, K. T., Kothuis, B. L. M., Brand, A. D., Jonkman, S. N., van Gelder, P. H. A. J. M., ... & **Fard, A. E.** (2017). *Hurricane Harvey Report: A fact-finding effort in the direct aftermath of Hurricane Harvey in the Greater Houston Region*.

FORTHCOMING

- Alfano, M., Sullivan, E., & **Fard, A. E.**. *Ethical pitfalls for natural language processing in psychology*. In M. Dehghani & R. Boyd (eds.), *The Atlas of Language Analysis in Psychology*. Guilford Press.

UNDER REVIEW

- **Fard, A. E.**, & Maathuis, C. *Capturing the Sense of Using Information as a Weapon in Social Media: a Computational Approach*.
- Mohammadi, M., **Fard, A. E.** *Ontology Alignment Revisited: A Bibliometric Narrative*.
- **Fard, A. E.**, & Verma, T. *A Comprehensive Review on Countering Rumours in the Age of Online Social Media Platforms*. In *Causes and Symptoms of Socio-Cultural Polarization: Role of Information and Communication Technologies*, Springer (Under Review).

WORK IN PROGRESS

- **Fard, A. E.**, Mohammadi, M., Verma, T. *TwifEX: A Unified Library for Twitter Feature Extraction*.

