

Perspective Discovery in Controversial Debates

An exploration of unsupervised topic models

Jody Liu

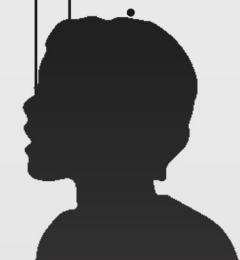
Technische Universiteit Delft

- [Redacted]
- [Redacted]
- [Redacted]
- [Redacted]
- [Redacted]

- [Redacted]
- [Redacted]
- [Redacted]
- [Redacted]

- [Redacted]
- [Redacted]
- [Redacted]
- [Redacted]

- [Redacted]
- [Redacted]
- [Redacted]
- [Redacted]



Perspective Discovery in Controversial Debates

An exploration of unsupervised topic models

by

Jody Liu

in partial fulfillment of the requirements for the degree of

Master of Science
in Computer Science

at the Delft University of Technology,
to be defended publicly on Thursday August 27, 2020 at 13:00 PM.

Supervisors:	Dr. N. Tintarev	
	T. Draws, MSc	
Thesis committee:	Prof. Dr. Ir. G. J. Houben,	TU Delft
	Dr. M. Aniche,	TU Delft
	Dr. N. Tintarev,	TU Delft

This thesis is confidential and cannot be made public until August 26, 2020.

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

Preface

This thesis marks the end of my time as a master student at TU Delft. For me, the ten month during thesis project is a reflection of how I have spent my time at TU Delft and what I have learned. These ten months gave me a glimpse of how researchers think 'outside-the-box', how they act on it and how to critically evaluate their own assumptions. But most importantly, I learned valuable lessons that I bring with me as I start my new chapter. In that respect I do not see this project done by one, but is rather a collection of multiple people that has shaped the thesis as it is.

It was near the end of the first year at TU Delft when it all started and that I have approached Nava for a thesis topic. At that time a thesis felt new and exciting, but daunting. It is unlike a course with predefined steps and I was unsure how I could best approach it. I am grateful for Nava's patience throughout this time and the constant guidance in how to shape the focus of my research. This was also one of the first learning points for me: *Research is an area with unknown variables, but with the liberty to explore them.* From giving me space to choose my own dataset to the opportunity to conduct a real user study. She encouraged me to scope my own research and more importantly gave me freedom to do so. Moreover, she was there for the support when needed, gave critical feedback and was always quick in her responses. With that I am also thankful for the whole Epsilon group. Their feedback gave me new angles to look at the problem and has become a safe place to ask questions and to be curious. This also made me accept better that, *It is okay to ask questions and it is okay to not know the answer at first.* I am also grateful for my daily supervisor Tim and his care for the project. I could fire any crazy question and we would brainstorm together about the possibilities. As research constantly made me question myself and to be in doubt, Tim helped me in building trust in my own work. Moreover, his knowledge in writing has been a valuable element and made me learn that: *Being a researcher is not only about conducting the research, but also about properly formulating this in an understandable way to an audience.* Another exciting moment was when this third point was being put to the test as I was starting to approach the Green Light. Geert-Jan, Maurício and Nava, who were in my thesis committee, were given the opportunity to read the thesis report. I am thankful for their time to evaluate the work. Lastly, my friends and family who have supported me in this journey by taking me out on (virtual) dinners and be the listening pole when I stumbled on frustrating moments. Consequently I was hit with my last realization point: *No matter the situation, allow yourself to have fun and that you are not alone.*

My time at TU Delft has taught me more than only study-related knowledge. If I could summarize my time at university in one sentence then it would be: *A chapter with opportunities to think freely but critically and regardless of the ups or downs there is always place to add some joy and laughter.* I also hope that you as a reader will find the research topic equally interesting and that you have learned a little bit more in the realms of this big world.

Jody Liu
Delft, August 2020

Contents

List of Figures	7
List of Tables	9
1 Introduction	1
1.1 Problem Statement	1
1.2 Research objectives	2
1.3 Research questions	2
1.4 Contributions	2
1.5 Thesis Outline	3
2 Literature background	5
2.1 Methodology of the literature study	5
2.2 Definitions and concepts	5
2.2.1 Defining perspective	5
2.2.2 Theoretical framework for perspectives	6
2.2.3 Limitations of the theoretical framework	6
2.3 Sentiment analysis	7
2.3.1 Introduction	7
2.3.2 Methods for sentiment analysis	7
2.3.3 Limitations	8
2.4 Aspect extraction	8
2.4.1 Introduction	8
2.4.2 Methods for aspect extraction	8
2.4.3 Topic models	8
2.4.4 Joint topic models	10
2.4.5 Limitations	13
2.5 Evaluation methods for topic models	13
2.5.1 Introduction	13
2.5.2 Topic coherence	13
2.5.3 User study	13
2.5.4 Experimental setup of the joint topic models	14
2.6 Summary	14
2.7 Research gap and motivation	15
3 Methodology	17
3.1 Introduction	17
3.2 Formulation of the research questions	17
3.3 Method of evaluation	18
3.4 Summary	19
4 Dataset	21
4.1 Introduction	21
4.2 Data assembling process	21
4.2.1 Data retrieval	21
4.2.2 Data cleaning	22
4.2.3 Data annotation	22
4.3 Exploratory study of the data	24
4.4 Data input for the models	25
4.5 Summary	26

5	Topic models for perspective discovery	27
5.1	Introduction	27
5.2	Preliminaries	27
5.3	Selection of joint topic models	28
5.4	Pipeline of the topic models	28
5.5	Summary	29
6	Evaluating the clustering ability of topic models	31
6.1	Problem formulation	31
6.2	Evaluation setting	31
6.2.1	Method	31
6.2.2	Data preprocessing and tokenization	31
6.2.3	Evaluation metrics	32
6.3	Finding the most optimal preprocessing techniques	33
6.4	Qualitative analysis of the final topic model setup	35
6.5	Summary	35
7	Human understandability of topic models	37
7.1	Problem formulation	37
7.2	Experimental setting	37
7.2.1	User task	38
7.2.2	Pilot study	39
7.2.3	Participants	39
7.2.4	Procedure	39
7.2.5	Statistical analyses	40
7.3	Results	42
7.4	Summary	46
8	Discussion	47
8.1	Introduction	47
8.2	Dataset	47
8.3	Topic model performance	48
8.4	User experience	49
8.5	Limitations	50
8.6	Summary	50
9	Conclusion	53
9.1	Conclusion	53
9.2	Future work	54
A	Data and libraries	57
A.1	List of perspectives	57
A.2	Preprocessing of the data	57
B	User study	59
B.1	Results	59
B.2	Interface design	59
	Bibliography	65

List of Figures

2.1	An example of an opinion quintuple extraction on a controversial claim in text. Documents taken from [1]	6
2.2	Visualization of LDA	9
2.3	The plate notation of the Latent Dirichlet Allocation model. α = Dirichlet prior for the per-document topic distribution, β = the Dirichlet prior for the per-topic word distribution. Given M documents, θ_m = the topic distribution for document m . Given N tokens z_{mn} = topic for the n -th token in m , and w_{mn} = specific token.	9
2.4	Plate notation of TAM. It includes two multinomial distributions τ and ψ , and binomial distribution σ . A token has thus two variables z and τ to denote the assignment of a topic and an aspect respectively. And also two binary variables l and ψ to denote if the token is 1) topic or background dependent, and 2) if the token is aspect dependent or not. For example, $l = 0$ is a background token, $l = 1$ a topic token. A token can therefore be background-related, an aspect-independent topic or an aspect-dependent topic.	11
2.5	Plate notation of JST. JST provides the element of sentiment to describe different topics. Drawn from distribution π , switching variable s determines whether a token is from a $topic_z$ -positive-word-distribution or $topic_z$ -negative-distribution.	11
2.6	Plate notation of VODUM. Drawn from distribution π , switching variable v determines whether a word is a noun (topic word) or a verb, adjective, adverb (perspective word). A token is then drawn from a $topic_z$ -noun-word-distribution ϕ or the $topic_z$ -non-noun-word-distribution ϕ_2	12
2.7	Plate notation of LAM. It uses a combination of POS-tags and subjectivity lexicon in which multiple types of tokens can be drawn. A switching variable x determines whether a token is a noun drawn from a background distribution ϕ_b or non-nouns from a topic-specific argument distribution ϕ_a . Tokens from the latter distribution can also be divided into two types of arguments drawn from either $topic_z$ -positive-arguments distribution or $topic_z$ -negative-arguments distribution.	12
3.1	Illustration of how each research question is answered	18
3.2	Pipeline of the thesis structure	18
4.1	The interface of <i>Debate.org</i>	22
4.2	Distribution of number of perspectives per document	24
4.3	Distribution of the perspective indices present in the corpus	24
4.4	The six expressed perspectives	25
4.5	Word clouds on the complete data and when data is split based on stance	25
4.6	Zip's law distribution of the top 50 unique tokens in the corpus. The token 'abortion' is removed here and is seen as a stopword.	26
5.1	Pipeline of how the topic model are being used	28
6.1	The iterative process of finding the most suitable tokenized data input to answer RQ2	32
6.2	Top 10 tokens per topic per topic model	35
7.1	Snippet of the second page of the user study. Participants had to assign a perspective to the correct topic. An error was given if the criteria stated in the instruction has not been met. Here the viewpoints refer to perspectives.	41
7.2	Distribution of chosen perspectives by the users across all the models. Perspective labels are on the x-axis and frequency on the y-axis.	42

7.3	Normalized distribution of how often each perspective in V (excluding honeypot checks) has been chosen. The true perspective labels can be found in Table 7.1. The red line equals 0.0625 (= 1/16) and is the probability of an arbitrary chosen perspective in list V . For a perfect model, the six correct perspectives should be chosen more often than any other random perspective with probability 0.0625.	42
7.4	The difference between LDA and TAM in the user's chosen labels on a given topic. Users gave 10 different answers for LDA and 3 for TAM.	43
7.5	Average number of correct perspectives found by users, with standard error	44
8.1	The six correct perspectives that are present in the final corpus, together with their baseline tokenization.	48
A.1	List of perspectives to annotate the documents, according to <i>abortion.procon.org</i>	57
B.1	Informed consent	61
B.2	First page of the user study after the informed consent	62
B.3	Snippet of the second page of the user study. The participant first needs to read the instructions.	63
B.4	Final part of the user study where the participant reviews the topic models.	64

List of Tables

2.1	Terminology related to perspectives	6
2.2	Overview of Joint Topic Models	11
2.3	The experimental setup in existing literature	14
4.1	Attributes of the dataset upon retrieval	22
4.2	Interrater-reliability score between main annotator and two independent annotators, using Krippendorff's Alpha	23
5.1	The key differences between each topic model	29
6.1	Influence of the data input on the adjusted rand index and topic coherence scores per model	33
6.2	Influence of combining preprocessing techniques on the adjusted rand index and topic coherence scores per model	34
7.1	The total list of perspectives V shown to a participant. p_1 to p_6 are the correct perspectives that are present in the corpus. $p_{honeypot1}$ and $p_{honeypot2}$ are the honeypot checks and the other eight perspectives are the randomly chosen perspectives from the <i>Procon</i> list. List V has an even number of pro- and con-perspectives.	38
7.2	Frequencies based on stance	39
7.3	An overview of the questions being asked in the user study	39
7.4	Descriptive table of the topic models. $N_{participants}$ = number of participants, $\mu_{perspectives}$ = average number of perspectives chosen per topic, $\mu_{correct}$ = average number of total correct perspectives found per topic model, SE = standard error of $\mu_{correct}$	44
7.5	ANOVA score - The impact of topic models on the number of correct perspectives	44
7.6	Post-hoc comparison - The impact of topic models on the number of correct perspectives	45
7.7	Correlation table between stance and number of chosen pro-perspectives (pro_n).	45
7.8	Mean and standard deviation of the answers given on Q_6 , Q_7 and Q_8 of the user study. The answers are a 5-point likert scale with 1 being strongly disagree and 5 strongly agree.	46
A.1	The libraries used for this research	58
B.1	Open feedback comments from participants on the user study	59

1

Introduction

1.1. Problem Statement

Over the years, researchers within the computational fields have focused more on extracting useful information from online opinionated text sources (e.g., debate fora or social networks) as the Web has provided richer user content [2]. As such, researchers have explored methods to perform sentiment analysis with an increasing interest in stance classification on controversial debates (= is a document supporting, opposing or neutral towards a controversial claim?) [3] [4] [5] [6] [7]. However, aside from classifying documents based on stance it would be equally interesting to automatically extract the underlying reasons behind a stance in order to truly understand a controversial debate. We may call these underlying reasons as perspectives [8]. In controversial debates with strong dividing opinions there may exist a large spectrum of supporting- and opposing perspectives. For example, a person supporting the legalization of abortion can take the perspective that *women should have the right to make their own choice about their own body*. Other people can oppose to abortion with the perspective that *a mother should take responsibility when producing a child instead of performing abortion*. Few researchers have focused on extracting such perspectives because it may require an available labelled dataset on perspectives. Such a labelled dataset will help in assessing whether the perspectives are correctly extracted from text. However, textual discussions can easily expand to a large number of documents in which labelling their main underlying reasons becomes a time-consuming task [8]. Moreover, controversial debates may require sufficient knowledge in order to understand the diverse set of perspectives and also depends on how annotators perceive the world [9]. This can make the annotation process a bias-prone task. With no such labelled dataset, it is unclear how accurate models can identify and cluster perspectives from opinionated text on controversial debates. Additionally, it is unclear whether the models' output is understood by humans for usability.

Whereas few have focused on extracting perspectives and no usable dataset is known to date, finding and extracting perspectives can be especially important today. Online debate platforms have become accessible platforms for people to freely voice their opinion [10]. Such accessibility provides rich textual information. However, it also renders information overload because users have only limited capacity to read the available information. Especially when a debate starts to grow to hundreds or thousands of opinionated documents without any structure in the discussion, it can be a time-consuming task to read each single document. For example, in 2019 alone the number of documents related to the abortion debate hits over 227.000 search results ¹. Only a few may have the patience to read through all the search results. Consequently, high cognitive information overload can lead to phenomena as media bias by journalists and confirmation bias in which one selectively reads information in line with their own belief [11] [12]. This leads to the likelihood of narrow-minded discussions and low acceptance of other perspectives [13]. Consequently, it may further increase unwanted conflicts and leads to a too fast judgement in decision-making without considering other ideas or thoughts. For example, the debate about the legalization of abortion has become an unending legal and moral issue among the American public [14]. It addresses the ethical issue of when life starts and the moral status of a fetus, and has made the abortion debate an unsolved political discussion between Democrats and

¹We used search query: *abortion debate after:2019-01-01 before:2019-12-31* in search engine www.google.com

Republicans [15]. Instead of trusting their own ideology with blind faith, considering a broad spectrum of perspectives nurtures analytical and critical thinking [16]. When taking different perspectives into account, people will question the credibility of the source, re-think their own arguments and try to identify and challenge existing assumptions.

To this end it would be useful to automatically get an overview of 1) the controversial topic and 2) the different perspectives for being for or against a controversial claim. Such an overview transforms content from large unstructured documents to a structured set of perspectives, reducing cognitive information overload. We may call this transformation process as *perspective discovery*. We define perspective discovery as the process of automatically finding and extracting a structured overview of perspectives from unstructured text. Such perspective discovery may be useful for journalists to write objective articles quicker, citizens could use it to develop a better informed opinion or legislators may use it for policy-making decisions. The research of this thesis therefore focuses on closing the gap between stance classification and perspective discovery. To this end, we research existing methods that can potentially provide a human understandable overview of perspectives from large unstructured controversial debates.

1.2. Research objectives

A family of methods potentially supporting perspective discovery is the concept of topic modelling. A topic model is an unsupervised method designed to find underlying structure that is hidden in unstructured text. Based on multinomial distributions, a topic model gives such a structure by computing a set of main topics. Additionally, each topic can be described by a set of (e.g., 10) representative top words. Topic models have been used for information extraction tasks. For example, they can be used to distinguish and describe documents with distinguishable characteristics such as a corpus of articles about politics, sports and business [17].

To the best of our knowledge, topic models have not been used for perspective discovery and on one specific controversial debate. We view the topic model's topics as perspectives in which each perspective is described by a set of words. With online platforms creating large unstructured discussions with unlabeled text, we explore if unsupervised topic models can be used to create a structured overview of a corpus on controversial debates. In that respect we quantify the unsupervised topic models on how well they discover correct perspectives and on how human understandable the perspectives are.

1.3. Research questions

Based on the research objective, we have defined the following main research question:

What topic model is able to discover human understandable perspectives on controversial issues?

There are a few sub questions in order to answer the main research question.

- **RQ1:** What topic models exist in relation to perspective discovery?
- **RQ2:** What topic model is most suitable to correctly cluster documents based on their perspective label?
- **RQ3:** What topic model is most informative to correctly explain perspectives to humans?
- **RQ4:** Do users choose more perspectives in line with their own pre-existing stance?

1.4. Contributions

The contributions for this research are highlighted below. These are based on the identified research gap from the literature study.

- We explore if topic models can be used to discover perspectives in online controversial debates.
- We compare potential topic models by using one evaluative framework leading to a better comparison and understanding of existing methods.

- As part of the evaluative framework, we create a new labelled dataset that has a perspective label assigned to each document. With this dataset we ensure that all topic models are evaluated with the same data input as well as providing an evaluation against a ground truth.
- The evaluative framework emphasizes on quantifying the correctness level of the topic model's output. For this we introduce a metric to evaluate how well topic models can cluster documents based on perspectives and compare this with a ground truth. Moreover, we introduce a user study to evaluate how well the topic model's output is also human understandable.

1.5. Thesis Outline

Chapter 2 defines the concept of perspectives and explores the closely related field of opinion mining. Here we focus on what information needs to be extracted for perspective discovery and what techniques currently exist to perform this task. Chapter 3 explains why the four sub research questions were formulated which stemmed from the identified research gap in the previous chapter and the main research question. We also explain the approach to answer these research questions. Chapter 4 introduces the dataset and how the dataset has been assembled. This is followed by Chapter 5 answering **RQ1** and explains the models used for the experiments. Chapter 6 and 7 answer **RQ2, RQ3** and **RQ4**. We then discuss these results in a subsequent chapter followed by the conclusion in Chapter 9.

2

Literature background

2.1. Methodology of the literature study

As mentioned in Chapter 1 the aim of this thesis is to discover perspectives in text. The literature study first defines the concept of perspectives and how this differs from stance. We then focus on two main areas: perspectives on conceptual level and the computational methods for perspective discovery. A computational field closely related to perspectives is opinion mining. With opinion mining we can understand what type of textual information needs to be extracted for perspective discovery and what computational methods exist to perform this task. These methods are also known as sentiment analysis and aspect extraction. The last part looks into research about topic models as they enable the discovery of hidden structures in text in an unsupervised way. This may enable the combination of both sentiment analysis and aspect extraction at once to automatically discover perspectives from unstructured documents.

2.2. Definitions and concepts

2.2.1. Defining perspective

A controversial issue or claim is a concept that often invokes two strong polarities. A person can support a claim or can oppose to a claim. This classification is also known as stance. The notion of stance is closely linked to sentiment, but is not always the same. For example, *“Not legalizing abortion is bad. Abortion is not the killing of a person because the fetus is not alive yet”* expresses a negative sentiment whereas it still takes a pro-abortion stance [18].

Given a claim, people therefore hold a position (= stance) towards a claim. However, stances usually have underlying reasons; for example *“Abortion is not the killing of a person because the fetus is not alive yet”*, illustrates a positive stance towards abortion. This stance is invoked by the topical issue (= controversial claim) where the source of its stance is explained by someone’s underlying reason [19]. We define these underlying reasons as perspectives. The perspective for supporting abortion is *the belief that abortion is not murder because a fetus is not yet alive*.

Table 2.1 adheres to terminology used within this section. This terminology may be used for sentiment, stance or perspectives and may lead to confusion for the reader. The list therefore highlights these concepts to establish a common understanding throughout this thesis.

Term	Definition
Controversial issue	A concept of public interest invoking two or more conflicting stances or views [20] [21]
Sentiment	A state associated with one's evaluation or judgment often categorized in three polarity types: positive, neutral and negative [20]
Stance	A favorability towards a specific topic, idea, object or proposition and takes either a neutral, supporting or opposing position [22]
Perspective	The underlying arguments behind a person's stance [8]

Table 2.1: Terminology related to perspectives

2.2.2. Theoretical framework for perspectives

A research field closely linked to perspectives is opinion mining. The area of opinion mining may use a theoretical framework in which opinionated documents can be viewed as a quintuple of five attributes [23]. These attributes are {*Entity*, *Aspect*, *Opinion Orientation*, *Opinion Holder*, *Time*}. Figure 2.1 illustrates what the different attributes are for a controversial issue such as abortion and shows how the framework can be used for finding our perspective. The claim is about whether abortion should be allowed or not, with two opinionated documents.

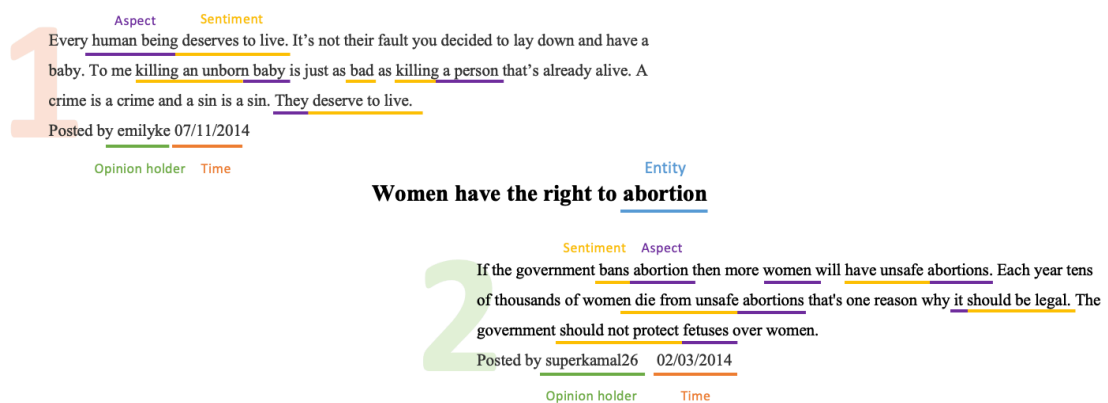


Figure 2.1: An example of an opinion quintuple extraction on a controversial claim in text. Documents taken from [1]

The entity (shown in blue) is the main controversial topic in which an opinion is directed at and is in this case abortion. An opinion always holds an opinion orientation (indicated in yellow in Figure 2.1) and can be positive and negative (e.g. "bad" is negative, "protect" is positive). These so-called sentiment words are often a set of adjectives, verbs, verb phrases and adverbs.

The third attribute of the framework is called aspects and are defined as the main subject in which the sentiment may be linked to (shown in purple). For example, killing an unborn 'baby', or die from unsafe 'abortions'. As shown, these relevant aspects are often the nouns and noun phrases in a text. We also see that merely nouns or noun phrases are not meaningful enough to understand a person's mental position towards a claim. We thus require a combination of both nouns and sentiment words.

The fourth and fifth attributes (indicated in green and orange) are the opinion holder - is the person who expresses the opinion - and time. These are often the metadata of a text. Based on this theoretical framework we can discover perspectives from text by representing a perspective as a combination of opinion orientation and aspect.

2.2.3. Limitations of the theoretical framework

Based on the theoretical framework for opinion mining, perspectives can be linked to opinions. Here a perspective may be formed by the framework's opinion orientation and the framework's aspects. The methods to extract them from text are also known as sentiment analysis for the opinion orientation

and aspect extraction for the aspects. With these two methods one could go through each sentence and identify the necessary aspects and its sentiment. These methods are mostly done separately and are later combined to fill in the attributes in the theoretical framework [18]. However, this becomes a difficult task when the sentences are becoming too complex.

First, there are cases in which sentences are not lengthy enough to identify a sentiment or aspect at all [24]. (E.g. *"I understand the problem"* (= what is the sentiment?), *"I like it"* (= where does 'it' refer to?)). A second issue is that sentences may hold a sentiment, but does not necessarily be an actual sentiment. An example is in conditional sentences - consisting of two clauses: condition clause and consequent clause - *"If being against abortion increases the likelihood of less death rates, then I will vote in the next campaign for Trump"*. Such a construction might convey a sentiment but is actually non-sentimental due to the if-statement. Additionally, sentences are of too complex structures in which multiple expressions are present [23]. (E.g. *"I understand that killing innocent lives is wrong, but I don't believe that a life begins when a baby is still in a woman's fetus"*). Lastly, there may be comparative phrases which may express a feeling but does not necessarily convey a negative or positive sentiment towards either one of two targets. (E.g. *"Medical abortion is better than surgical abortion"*, does the person think abortion is good or wrong?).

In other words, understanding perspectives on sentence-level is possible in case of simple sentences, but has its limitations when the sentences become complex. To possibly overcome these issues we first need to understand what methods currently exist for sentiment analysis and aspect extraction separately. We then identify a set of methods that may find a perspective by ignoring this separation which we call joint topic models. These joint topic models are an enhanced version of aspect extraction that combines the aspects (= nouns) and sentimental words (= verbs, adverbs, adjectives).

2.3. Sentiment analysis

2.3.1. Introduction

Sentiment analysis can be used to determine the perspective's sentiment and it has been widely used to automatically determine a sentiment in product reviews, news articles or debate fora. This has been shown to be useful as more organisations want to know how users feel about a service, product or event - allowing for targeted recommendations and fulfilling user needs [25] [26] [27].

Researchers have explored the use of sentiment lexicons and natural language processing features such as Part-of-Speech tags (POS-tags) to enhance the automatic sentiment classification from text.

2.3.2. Methods for sentiment analysis

Current methods to detect sentiment are mostly supervised [28]. The supervised approach can be formulated as a categorical classification problem in which the classification of the model is either objective, positive or negative [29]. Although the order might change, the main steps for these classification tasks are the creation of dependency trees to parse the sentences and the use of a sentiment lexicon to categorize these words based on sentiment [24] [30] [23].

For the first task dependency parsing trees, as the Stanford NLP Library¹ are applied to understand the words' grammatical relationship in sentences. These sentences are represented by a tree in which sub-trees are clauses focused on a specific aspect. Having this structure aids in understanding which words in the text should be extracted for sentiment. Because adjectives, verbs or adverbs mostly portray a sentiment we can use this parsing tree to identify these specific sentiment words.

Sentiment lexicons are then on these relationships used to determine the overall binary-valued sentiment in a given text. Sentiment lexicons such as SentiWordNet [31] are dictionaries of sentiment words categorized on their sentimental orientation of being positive, negative or objective. For example, (+1)-score can be given to adjectives related to positive sentiment appearing in the lexicon and (-1)-score on adjectives appearing in the negative in the lexicon. One could then perform computations to determine the overall sentiment of a text by taking the average, multiplication or addition of these scores.

Over the years, researchers have further optimized traditional sentiment analysis techniques by adding extra features. An example is the realization of term presence and term frequency taken from the field of Information Retrieval and Text Classification [32]. Both term presence and term frequency have been adapted for sentiment analysis tasks in which it has been shown that term presence have

¹<https://nlp.stanford.edu/software/lex-parser.shtml>

a higher impact than term frequency. For example, taking the occurrence of rare words into account instead of frequently occurring words has been shown to contain more information - also known as the Hapax Legomena [32]. The term presence is also highlighted in case of sentiment at the end of sentences: it has been observed that the occurrence of negative sentiment words at the end of a sentence determines the overall sentiment when classifying the sentiment regardless of how many times a positive is observed in a text [32].

2.3.3. Limitations

Although these methods mostly focus on the occurrence or presence of sentiment words, it does not always take syntactic structure or grammatical dependencies into account. This means that certain valence shifters such as intensifiers, negations or downtoners are not able to be identified [24] [29]. For instance, "not great" is less negative than "not good", whereas "not true" is more negative than "hardly true". These are still hard to identify in current methods, because there is a high variety in how humans can express their sentiment. Due to this high variety machine learning algorithms have the ability to learn only a portion of the possible text patterns as there is not enough data to train on [23]. Even for humans it is not always as easy. A research showed that the average human achieves a 77% accuracy in determining a sentiment in political debates [33]. Moreover, current classification tasks could handle around 60% of the cases [23].

2.4. Aspect extraction

2.4.1. Introduction

The extraction of aspects (e.g. organizations, people and locations) can be used for finding the perspective's aspects. The existing methods for this task fall within the traditional information extraction domain [17]. In the traditional sense these techniques have been used to find main topics or categories in a corpus. This can be useful to categorize news articles in their categories such as sports, culture or politics or to summarize the main issues of a large set of articles. Consequently it helps in document browsing, text mining and document retrieval [34] [23]. As mentioned earlier, most aspects are nouns or noun phrases [35]. In case of perspective discovery, the difficulty with aspect extraction lies in finding the correct aspects because not all nouns or noun phrases in a text represent an actual perspective.

2.4.2. Methods for aspect extraction

To the best of our knowledge, no supervised methods exist to find the aspects targeted for perspective discovery. This may be a consuming task for large sets of documents because each document needs to be read in order to identify which aspects are of importance and which are not.

There are three main unsupervised approaches to perform aspect extraction. The first method is to create POS-tags to identify nouns and then apply language rules in which explicit structures in the text are used to find the targeted nouns [23]. A second technique is the use of sequence models such as Hidden Markov Models (HMM) to learn and predict patterns in the text. When a sequence is learned HMM may predict the next subsequent word. If it predicts the targeted noun, the word will be extracted. The third line of methods is topic modeling. This method takes the probabilistic property of word occurrence into account and is more descriptive than the other two methods as related words are being grouped together [18]. What makes topic model different from the other methods is that it ignores the necessity of going through each sentence individually. Instead, it looks at the complete corpus to find the main aspects which are also called the topic model's *topics*. Such a flexibility overcomes the difficulty of finding the correct aspects when the sentences are becoming too complex as described in Section 2.2.3. There is a rich variety in how humans can formulate a perspective into words and these variations cannot always be covered by using strict, pre-fixed language rules or by learning patterns as in HMMs. A topic model could find aspects regardless of how a text is syntactically written.

2.4.3. Topic models

Introduction

Topic models are originally designed to discover short descriptions of hidden elements (topics) in a collection of text [36]. It enables faster processing of a large corpus while maintaining the statistical

relationships between documents. For perspective discovery each aspect is seen as a topic k with each k being described through the top N tokens from a per-topic word distribution. This makes it possible to find aspects regardless of the text structure, writing style or use of language. Unlike language rules or HMM, we do not need to know specific syntactical structures of sentences or word probabilities after a sequence of words. Losing this specificity allows for more flexibility in which one aspect is described by a set of tokens. This comes however at the cost of interpretation as the aspect is now not explicitly given. The aspects are now expressions grouped together by the co-occurrence of tokens.

Topic models: Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is one of the most common topic models. LDA is a Bayesian statistical model. Given a corpus, LDA generates two distributions from documents as seen in Figure 2.2. The first, shown at the bottom of the image, is per-document topics distribution θ to draw a mixture of underlying topics K for each document. The other is per-topic words distribution ϕ to draw a mixture of tokens for each topic k in K .

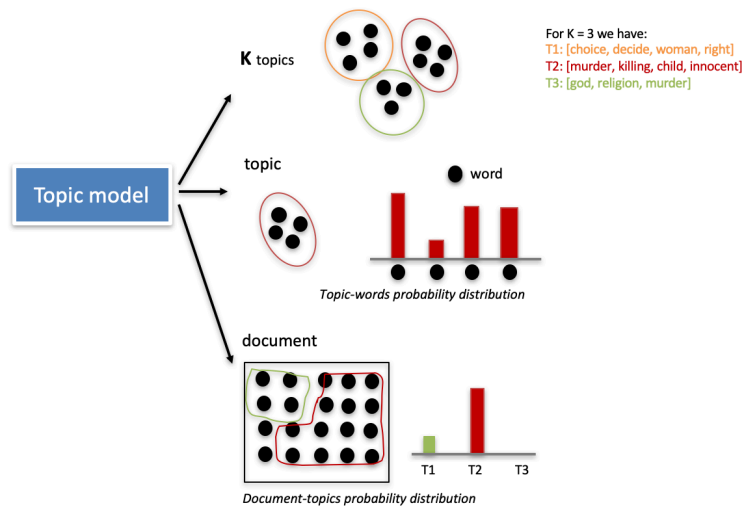


Figure 2.2: Visualization of LDA

A topic model is often illustrated through a plate notation as seen in Figure 2.3.

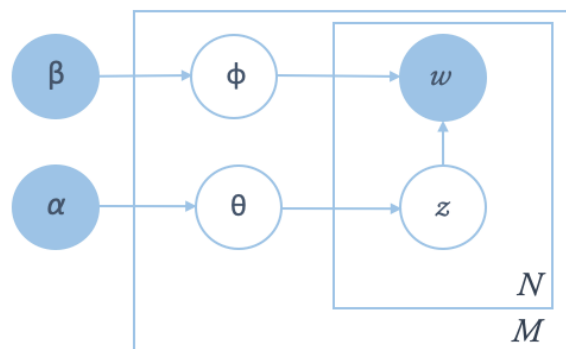


Figure 2.3: The plate notation of the Latent Dirichlet Allocation model. α = Dirichlet prior for the per-document topic distribution, β = the Dirichlet prior for the per-topic word distribution. Given M documents, θ_m = the topic distribution for document m . Given N tokens z_{mn} = topic for the n -th token in m , and w_{mn} = specific token.

LDA has three hyperparameters to estimate the distributions θ and ϕ . The number of K topics the model needs to generate, the Dirichlet prior α for the per-document topics distribution θ and Dirichlet prior β for the per-topic words distribution ϕ . Given M documents the model would have θ_m as the topics-distribution for document m . Given N tokens z_{mn} is a topic for the n -th word in document m ,

and w_{mn} is the specific word. With these computed distributions we can determine the K topics where each topic k in K is described with the top N tokens from a topics-word distribution.

Prior to LDA there were models such as Latent Semantic Indexing (LSI) and Probabilistic LSI (pLSI). However, the needed parameters for the latter two models grow linearly with the number of training documents, suggesting that these models are prone to overfitting [36]. LDA has a fixed parameter K which is irrespective of the corpus size. Moreover, LDA assumes a document can have multiple topics drawn from a per-document topics distribution, whereas its predecessor is one-dimensional assuming a document can only have one topic.

Limitations of Latent Dirichlet Allocation model for perspective discovery

Given a set of documents, LDA may perform well in distinguishing main topics when each entity have their own distinguishable jargon. For example, a set of documents with topics as sports, economics and culture can be distinguished. It becomes a more difficult task when the model needs to discover perspectives from a set of documents that focuses on one specific topic.

First, an LDA with a corpus focusing on purely one topic such as abortion would have difficulty discovering the distinct perspectives because the documents are quite homogeneous. This means that all documents cover the entity abortion and use similar words which makes it difficult to find distinguishing perspectives between these documents [37]. For example, *"this is wrong, a child has the right to choose"* opposes abortion, whereas *"this is wrong, a woman has the right to choose"* supports abortion. Both use many similar words, but the meaning is different.

The second observation is that LDA by itself is not able to distinguish neutral words and sentimental words. LDA would not make such a distinction and would put all these words together as long as they are related to each other. This would also mean that negative or positive words would be grouped together due to relatedness. For example, given two documents with one about *"murder"* and the other about *"not murder"* will be put together because both are about murder. Instead, the first document might be against abortion because abortion is seen as murder and the other supports abortion because it is not seen as murder. This makes LDA less useful when it needs to be applied on perspectives. Differences between perspectives are established because people hold different stances but people can also have the same stance but have their own reasons. LDA would have difficulty in finding these two separation types.

For the first problem, iterative LDAs have been proposed. Preliminary topics are being generated by the topic model. A following iteration with this preliminary knowledge can then be used to find more fine-grained topics [38]. Such iterations have been shown to improve the overall quality of extracting deeper levels of aspects in comparison to the standard LDA. This problem does however not solve the second problem. A possible direction is then to create joint topic models (JTMs). Instead of categorizing words based on only its statistical relatedness, these models incorporate features used within sentiment analysis to create stronger distinctions.

2.4.4. Joint topic models

We use the term Joint Topic Models (JTM) to describe topic models that use LDA as a base with additional components. Moreover, these models are designed with the believe that documents about the same subject are written differently based on the author's mental position. For example, research abstracts about computational linguistics would be written differently by a computer scientist versus a linguist. Although these JTMs have not been used for perspective discovery on controversial issues, we can investigate whether this is possible. In this case we view the documents as being written by authors that hold different stances on a controversial topic. In other words, the mental position an author could have is being for or against a controversial topic where JTMs can potentially discover the perspectives between and within these stances from text.

An overview of JTMs is shown in Table 2.2. The four models are *Topic-Aspect Model* (TAM), *Joint-Sentiment Topic Model* (JST), *Viewpoint-Opinion Discovery Unified Model* (VODUM) and *Latent-Argument Model* (LAM), and are based on LDA. As seen in Table 2.2, most models incorporate elements from sentiment analysis to create more granular distinctions. This is also apparent in the computation of multiple per-topic words distributions for JTMs than only one per-topic words distribution in LDA. For instance, JST and LAM use prior knowledge to first separate words in either being positive or negative. This prior knowledge can be viewed as an external source that a topic model may use to compute their distributions. The prior knowledge that JST and LAM uses is a subjectivity lexicon where negative or

positive words may be drawn from a per-topic negative words distribution or per-topic positive words distribution respectively. All in all, their results show the importance of providing more granularity in words to generate more diverse topics.

Model	Use of prior knowledge?	Added component to LDA
Topic-Aspect Model (TAM) [39] (EMNLP '10)	No	Separates tokens based on stronger granular word distribution levels
Joint-Sentiment Topic model (JST) [40] (CIKM '09)	Yes, subjectivity lexicon	Separates tokens based on sentiment
Viewpoint-Opinion Discovery Unified Model (VODUM) [41] (ECIR '16)	No	Separate tokens based on POS-tags
Latent Argument Model (LAM) [8] (EMNLP '17)	Yes, subjectivity lexicon	Separates tokens through POS-tags and subjectivity lexicon

Table 2.2: Overview of Joint Topic Models

The differences between the JTMs are shown in Figure 2.4, 2.5, 2.6 and 2.7. Each model is illustrated through a plate notation. The blue component is LDA showing that each model uses LDA as a base. Furthermore, the plate notations highlight the main differences between each model that is determined through a switching variable (shown in yellow). For each model this switching variable has a different meaning and is a key element in what distinguishes them from each other. For each model the switching variable explains how the distributions from a Dirichlet such as α and β are drawn. These Dirichlet are the filled-in circles.

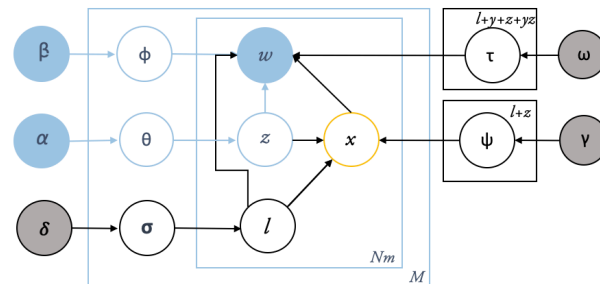


Figure 2.4: Plate notation of TAM. It includes two multinomial distributions τ and ψ , and binomial distribution σ . A token has thus two variables z and τ to denote the assignment of a topic and an aspect respectively. And also two binary variables l and ψ to denote if the token is 1) topic or background dependent, and 2) if the token is aspect dependent or not. For example, $l = 0$ is a background token, $l = 1$ a topic token. A token can therefore be background-related, an aspect-independent topic or an aspect-dependent topic.

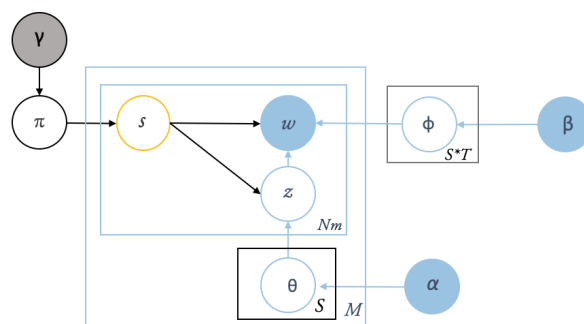


Figure 2.5: Plate notation of JST. JST provides the element of sentiment to describe different topics. Drawn from distribution π , switching variable s determines whether a token is from a $topic_z$ -positive-word-distribution or $topic_z$ -negative-distribution.

Topic-Aspect Model (TAM)

The focus of TAM is that the tokens in a document are drawn from multiple distributions instead of one [39]. TAM is a model that has an aspect covering an entire document. The model assumes that

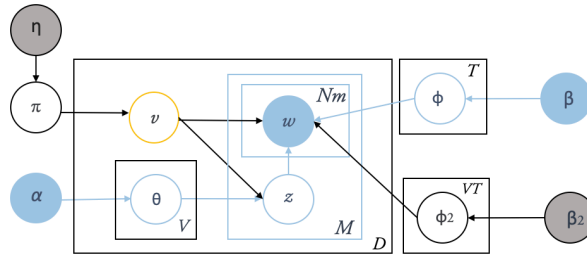


Figure 2.6: Plate notation of VODUM. Drawn from distribution π , switching variable v determines whether a word is a noun (topic word) or a verb, adjective, adverb (perspective word). A token is then drawn from a $topic_z$ -noun-word-distribution ϕ or the $topic_z$ -non-noun-word-distribution ϕ_z

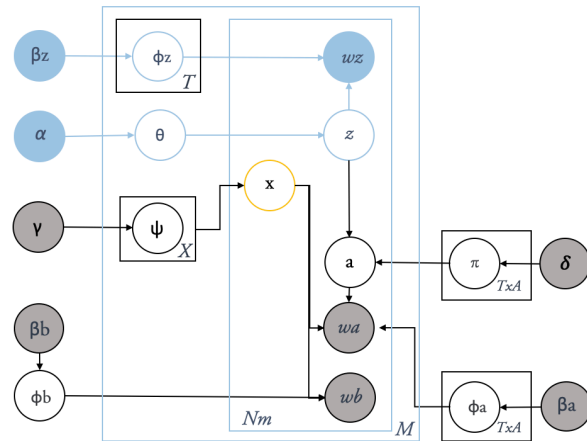


Figure 2.7: Plate notation of LAM. It uses a combination of POS-tags and subjectivity lexicon in which multiple types of tokens can be drawn. A switching variable x determines whether a token is a noun drawn from a background distribution ϕ_b or non-nouns from a topic-specific argument distribution ϕ_a . Tokens from the latter distribution can also be divided into two types of arguments drawn from either $topic_z$ -positive-arguments distribution or $topic_z$ -negative-arguments distribution.

tokens can be either or both a background-, aspect- or topic word. The model's approach gives a more granular representation of a corpus in which tokens can describe an entire corpus (= background tokens), a certain topic (= aspect-independent topics) or an aspect within a topic (= aspect-dependent topics). A switching variable x determines whether a word is topic-dependent or independent, as well as aspect-dependent or independent. This means that tokens can be drawn from multiple distributions and give more granularity than other joint topic models. For perspective discovery, we could see background tokens as all neutral tokens, topic tokens as tokens about the complete controversial topic and aspect-dependent topic tokens as the perspective specific tokens.

Joint Sentiment Topic Model (JST)

The goal of JST is to categorize topics and simultaneously classify them based on sentiment. For this it uses prior knowledge that is fed into the model [40]. This prior knowledge is the MPQA-lexicon which consists of sentimental tokens classified as being either positive or negative. As a result JST partitions tokens based on being positive or negative and gives for each sentiment a different set of topics. This separation is done through switching variable s . In our case, this means that a perspective's sentiment may be JST's sentiment label and the perspective's aspect JST's topic.

Viewpoint and Opinion Discovery Unification Model (VODUM)

VODUM separates tokens based on its Part-of-Speech tags where tokens can be drawn from either per-topic noun tokens distribution or from per-topic sentiment tokens distribution (= verbs, adverbs, adjectives). Tokens that are drawn from these distributions are called by VODUM as topic tokens and viewpoint tokens respectively. For this no prior knowledge is needed as the authors assume that they are too costly and not always available. Their aim is to create an unsupervised model that is applicable within multiple domains. The model is able to create multiple distributions where a topic may have

sentiment-independent topic tokens and sentiment-dependent viewpoint tokens. These topic tokens are formed through nouns and the viewpoint-tokens through sentiment tokens. Here, VODUM's topic tokens can be seen as the perspective's aspects and VODUM's viewpoint tokens as the perspective's sentiment.

Latent Argument Model (LAM)

LAM is closely in line with our research domain in which they aim to find arguments given multiple topics. The model could be seen as a combination of JST and VODUM since it uses both a subjectivity lexicon to determine the stance of documents and also POS-tags to determine if tokens describe a certain topic or a topic-dependent argument. This may make the separation of documents stronger than for the other models. For our experiments LAM's arguments can be seen as our perspective and LAM's topic as a specific controversial topic as whole.

2.4.5. Limitations

Most discussed models focus on explicit language in which words describe the exact meaning. Because these models are built on probability and word frequency, it may have difficulty in grouping coherent, understandable implicit words. For example, if we have "This phone is too small" - or "This phone easily fits in my pocket" - then in both cases 'small' and 'fit in pocket' implies 'size'. Mapping these implicit words requires an understanding of the context domain and there is still limited research done to properly map these in an automatic way [23]. The focus of this research is therefore on explicit target extraction only.

2.5. Evaluation methods for topic models

2.5.1. Introduction

The choice for topic model evaluation methods depends on what the goal of the evaluation is and can cover evaluations such as the extend of interpretability, generalizability, reliability, clustering, word coherence or stability [42]. Because topic models are unsupervised models there are no metrics focused on the accuracy level of such models. The most often used metric is topic coherence. Another possibility is to employ user studies. However, as seen in Section 2.4.3 few focus on user evaluation as they are also expensive in time.

2.5.2. Topic coherence

An evaluation metric widely used for topic models is the topic coherence score. This value may lie between 0 and 1 [43]. It measures how interpretable a topic is and whether the top N tokens in a topic are collectively a theme [44]. A high coherence score indicates that the tokens inside a topic are well-supported by each other.

There are several implementations for the topic coherence which have been examined in [43]. The authors explored five different topic coherence techniques where the c_v measure has the highest correlation with human judgement (= 0.7). The c_v measure has been proposed by [43] and is based on a pipeline of four steps. It segments topic T into set S and measures for all the possible pairs in S the probability that they co-occur in the complete corpus. Taking the mean of these probabilities result in a one-value topic coherence score.

In case of perspective discovery this metric can be used to evaluate how well topic models give coherent words to describe a perspective. A higher topic coherence score would mean a topic model is able to better create a coherent set of perspectives in a complete corpus.

2.5.3. User study

Another way to measure the quality of the topic model is to perform a user study. Because a topic model is unsupervised it is unknown whether the output is in line with the corpus. This means that in a user evaluation participants are asked for the human-identifiable semantic coherence of the model. A user study can be designed to perform the task of *word intrusion* and *topic intrusion* [45]. For word intrusion, the user has to identify a word that does not match in a topic (e.g. a user should recognize that 'tennis' does not belong in the word set *cat, dog, tennis, horse*). In topic intrusion we test whether a given topic associates with a document. This is done by letting the user identify a topic that is not in line with a given document.

2.5.4. Experimental setup of the joint topic models

Paper	Dataset	Model comparisons w/	Experiment goal	Evaluation method
TAM [39]	Research abstracts on computational linguistics Israel-Palestinian conflict articles	LDA, cLDA	Topic coherence Document classification	Use of user study Use model output as input for supervised classifier to determine accuracy of correctly classifying a document based on the stance
JST [40]	Free format movie reviews	SVM	Impact of prior knowledge on sentiment classification Topic coherence	Accuracy score of sentiment classification but not perspective Manual evaluation
VODUM [41]	News articles of Israel-Palestinian conflict	TAM, JTV, LDA	Topic coherence Document classification	Manual evaluation Accuracy of aspect identification, but not perspectives
LAM [8]	Parliament discussions of House of Commons in the UK	CPT, JTV, LDA	Topic coherence Topic clustering Perspective coherence Perspective clustering	Use of a specific topic coherence metric - see Section 2.5 for explanation Take top tokens within cluster and see if they have same topic label Use of a specific topic coherence metric Take top tokens within cluster and see if they have same topic label

Table 2.3: The experimental setup in existing literature

We examine the different experimental setups that authors have used to evaluate the JTMs mentioned in Section 2.4.3. From Table 2.3 most JTMs are compared with LDA, but few compare other JTMs found in Section 2.4.3².

As shown in Table 2.3 there are mostly two types of evaluation: topic coherence and document classification. For the first one the method to calculate the coherence score is different for each study. Combined with different datasets, it is unclear what kind of model performs better. Moreover, the limitation of such a metric is that it provides little judgement in the value to a human [46] [43]. For example, a high topic coherence score may be given with topical tokens [‘dog’, ‘cat’, ‘mouse’, ‘snake’] which are all animal tokens. Such a set would make sense when the article is about animals, but if the corpus is about abortion then the above topic output is unrelated to the corpus. It is therefore uncertain whether the topic coherence score relates to the correctness of the output and what the corpus entails.

As for the latter evaluation type we see that classification is merely based on the author’s angle (or in our case, stance) and not perspectives. For example, TAM has been used on computational linguistics research abstracts written by either a computer scientist or linguist. This means that the classification is based on whether documents can be separated based on these two angles. It is uncertain whether the topic model can also distinguish documents within one angle. It is possible that various computer scientists share different ideas (or in our case, perspectives), but the ability to distinguish these perspectives within one angle have not been measured. Similarly, when evaluating the classification performance for VODUM it was about whether documents can be separated between an Israeli or Palestinian author. There was no evaluation on the differences between several Israeli authors or between Palestinian authors. As none of the JTMs focus on this matter it is uncertain whether joint topic models are able to capture subtleties between opinionated documents when focused on one particular topic and if it can distinguish documents within one angle and if the topic model’s output is explainable enough to describe documents.

2.6. Summary

We defined perspectives as the underlying reasons behind a stance. A theoretical framework within opinion mining could help in understanding how perspectives from text can be extracted. Here, a theoretical framework is a quintuple of five attributes: $\{Entity, Aspect, Opinion Orientation, Opinion Holder \text{ and } Time\}$. The theoretical framework shows that a perspective holds a sentiment with regards to an aspect. As a result, this observation has been used to further understand current computational methods to find perspectives from text.

We see that literature often separates the task in finding sentiment and aspects as sentiment analysis and aspect extraction. Here, sentiment analysis may be found using parsing trees and sentiment lexicons to identify the overall sentiment. Aspect extraction can be found through unsupervised methods such as syntactic rules, Hidden Markov Models or topic models. However, separating both sentiment analysis and aspect extraction makes it difficult to identify a well-structured overview of perspectives as sentences can become too complex to find a clear sentiment or aspect.

We also found methods creating a mixture of both sentiment analysis and aspect extraction at once.

²CPT and JTV in Table 2.3 are also topic models but have not been introduced in previous sections as they are not available for use. This is one of the criteria when choosing JTMs.

These are joint topic models which is an unsupervised approach with the ability to find descriptive topics regardless of syntactical text structure. We looked at these joint topic model approaches to understand how perspectives could possibly be extracted from text. We saw that most of these aspect extraction methods use elements of sentiment analysis to find more granular descriptive aspects in text. Although these models have not been experimented on controversial issues with strong polarities, they may have the ability to find perspectives given their model's design. It is however unclear which joint topic model performs best as few joint topic models are being compared with one another. Moreover, all models have been evaluated under a different framework which makes comparison difficult. We also do not know how well these models can discover the correct perspectives as they are all unsupervised methods.

As most research focuses on explicit opinion expressions instead of implicit opinion expression, we also focus on explicit opinion expressions. This means that implicit opinions through verb phrases are not taken into account.

2.7. Research gap and motivation

Based on the literature study, a research gap has been identified. This gap presents the focus of the thesis.

- **Gap 1:** Literature on stance classification has been explored but few focus on how a perspective can be computationally modelled.

Contribution: We view perspectives as a combination of sentiment and aspects where we explore the use of an unsupervised approach called joint topic models. Instead of separating both elements into sentiment analysis and aspect extraction, we explore the use of joint topic models that may give more descriptive perspectives by incorporating sentiment analysis features in their aspect extraction approach.

- **Gap 2:** Joint topic models have not been used on documents focusing on one specific controversial debate. This questions whether distinct perspectives *between* but also *within* one stance can be found.

Contribution: In order to understand whether joint topic models can be used for perspective discovery in controversial debates, we need to know what perspectives are present in the input dataset. Because there is no dataset of labelled documents on controversial debates, we create our own annotated dataset consisting of perspective labels that are annotated on controversial forum posts (in this case the controversial topic is abortion).

- **Gap 3:** Most identified joint topic models made comparisons with LDA and did not perform evaluations with other joint topic models. This led to the question which joint topic model is most effective in finding perspectives.

Contribution: We employ an experimental setting in which all models are evaluated with the same metrics and dataset for better comparison.

- **Gap 4:** The joint topic models have mostly been evaluated on the basis of topic coherence. This does not mean that the found topics by a topic model are relevant and correct.

Contribution: In this research we evaluate the correctness of the joint topic models. We evaluate whether an unsupervised approach can correctly cluster documents based on ground truth labels and if the computed topics are descriptive enough to explain to humans. This approach allows us to quantify the correctness and usability of unsupervised joint topic models. As this quantification has not been done before, we introduce new types of evaluation on the basis of correctness.

3

Methodology

3.1. Introduction

In Chapter 2 we have identified four research gaps. To accompany each of these four research gaps we pose four research questions that also answer the main research question of the thesis. This chapter explains each research question, how it relates to the research gap and the method to answer each research question.

3.2. Formulation of the research questions

In total we have four research questions to answer the main research question: *What topic model is able to discover human understandable perspectives on controversial issues?* These four questions are explained below.

- **RQ1:** *What topic models exist in relation to perspective discovery?*

Explanation: Recall that we view perspective as a combination of sentiment and aspect. In order to explore the use of topic models for perspective discovery we first need to know what topic models exist that could potentially perform this task. **RQ1** therefore centers around this question and fills **Gap 1** introduced in Section 2.7.

- **RQ2:** *What topic model is most suitable to correctly cluster documents based on their perspective label?*

Explanation: The identified JTM's have not been evaluated in line with the specific task of perspective discovery. This means that there is no quantification of how well the JTM's can find the desired perspectives from text. For this we create a labelled dataset for perspective discovery, but also introduce new metrics to accompany this gap. This is in line with **Gap 2**, **Gap 3** and **Gap 4**. **RQ2** questions how well the topic model output groups documents together compared to a ground truth as to quantify a correctness level of topic models.

- **RQ3:** *What topic model is most informative to correctly explain perspectives to humans?*

Explanation: This research question also fills the last three research gaps. Both **RQ2** and **RQ3** are similar in that they focus on quantifying the topic model's performance. However, **RQ2** focuses on the clustering ability of topic models regardless of how perspectives are being described by the topic models. **RQ3** centers around how well topic models describe human understandable perspectives. A good clustering score for a model would not immediately mean that this is well understandable for humans, and vice versa. We therefore have both **RQ2** and **RQ3** to evaluate topic models based on their correctness but with a different approach.

- **RQ4:** Do users choose more perspectives in line with their own pre-existing stance?

Explanation: As mentioned in Chapter 1 the main research question arose from the problem that humans may have certain biases when reading articles due to high cognitive load. **RQ4** investigates how users with different stances interpret a topic model. An effect we question here is the *false consensus effect* [47]. This effect states that people tend to believe that their own behaviour is seen as common among the general public. This could mean that a person's stance on a controversial topic is seen as a common belief which influences how they perceive the potentially ambiguous output of a topic model. It is possible that users interpret the computed topics to be in line with their own stance which may not coincide with its intended meaning. Ideally, a topic model should not be subject to any form of personal preferences and we investigate whether this is the case.

3.3. Method of evaluation

Figure 3.1 illustrates how each research question is answered.

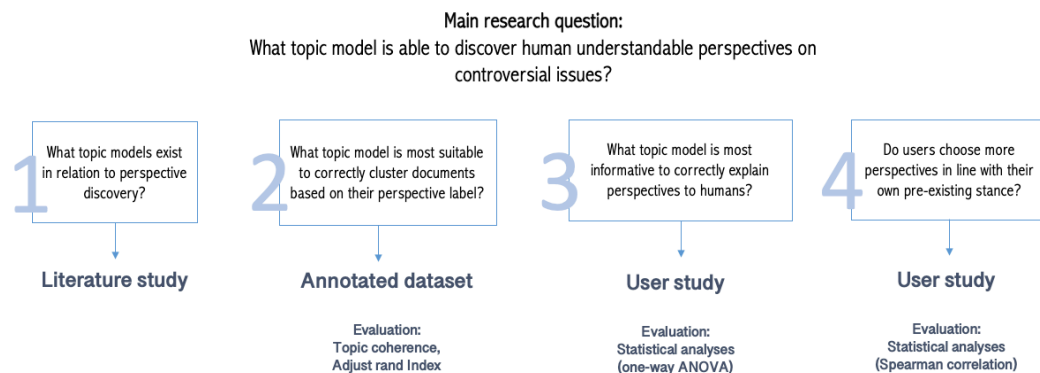


Figure 3.1: Illustration of how each research question is answered

We have answered **RQ1** by the literature study where we investigated a set of JTMs that can potentially be used for perspective discovery. Chapter 5 provides an overview of these JTMs and further explains how this can be used for perspective discovery. **RQ2** is an evaluation that uses the ground truth labels and two metrics: topic coherence and adjusted rand index. The latter metric has not been used yet to evaluate the identified JTMs and is explained in more detail in Chapter 6. **RQ3** and **Q4** are answered by conducting a user study. We evaluate these results through statistical analyses.

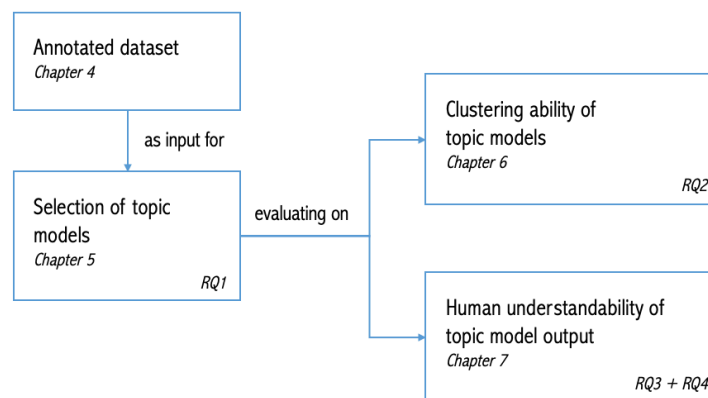


Figure 3.2: Pipeline of the thesis structure

Figure 3.2 shows the pipeline of the research and is an illustration of how the remainder of this thesis is structured. Chapter 4 introduces the dataset. This is followed by Chapter 5 where we explain

the topic models used in this research which answers **RQ1**. We then focus in Chapter 6 on **RQ2**. The next chapter centers around the last two research questions.

3.4. Summary

We posed four research question in order to answer the main research question. These four research questions are based on the research gap identified in the literature study. We also explained how we approach each of the four research questions and provided an overview of how the subsequent chapters are structured.

4

Dataset

4.1. Introduction

The goal of this study is to evaluate whether topic models are able to discover perspectives from a large set of documents. In order to perform an evaluation we need a set of labelled documents listing the perspectives present in each document. If there is no ground truth, it becomes difficult to evaluate if a topic model gives the correct output. To the best of our knowledge there exists no labelled dataset that is large enough and includes perspective annotations on documents. This led to the decision to assemble our own annotated dataset. These data are opinionated documents on the highly controversial topic *abortion* with documents having strong dividing opinions. Moreover, all documents are written in English.

To obtain opinionated documents we retrieve documents from an online debate forum called *Debate.org*¹. Moreover, this forum ensures that all documents are focused on a specifically chosen topic. This reduces noise present in other sources such as social media platforms where documents focused on one subject are more difficult to filter and the document length can be too short to hold a clear opinion. The choice of abortion as topic is based on the source *ProCon*². *ProCon* provides the most popular controversial issues in the United States where each issue has a list of opposing and proposing arguments. From the list of controversial topics on *ProCon* we chose abortion as topic because it gave one of the highest search results in *Debate.org*, as well as being the most balanced controversial discussion. The debate about abortion has 51% for-votes and 49% against-votes in *Debate.org*, highlighting how controversial the claim is with no definite right or wrong answer. Moreover, we use the opposing- and proposing arguments from *ProCon* to give a perspective label to each document in the dataset.

Section 4.2 explains how the data has been assembled and Section 4.3 explores this data.

4.2. Data assembling process

4.2.1. Data retrieval

The documents are retrieved from *Debate.org*. This debate forum lets users start an online discussion where any user can express their opinion on a topic. In our debate, all documents relate to the debated question of whether abortion should be legal.

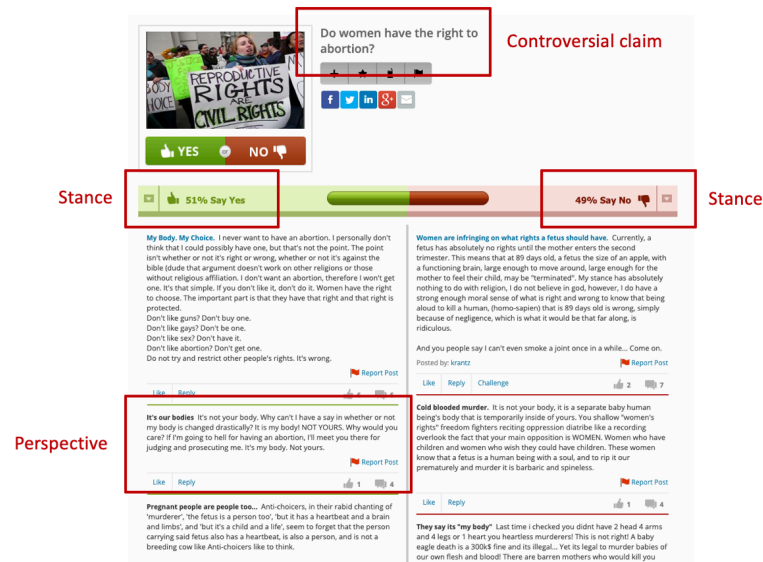
Figure 4.1 shows how such a debate looks like. At the top of the webpage there is the controversial claim. There are two possible stances: for and against, and under each stance a user's perspective. This is similar to how we view controversial claims where people hold a certain stance and have a perspective to express this stance. From the example we can also see that a stance may contain multiple perspectives.

We used a data extraction tool in the Chrome browser called *webscraper*³ to automatically retrieve the documents from *Debate.org*. The tool lets a user create a plan (sitemap) to navigate through

¹<https://www.Debate.org/>

²www.ProCon.org

³<https://webscraper.io/>

Figure 4.1: The interface of *Debate.org*

a website and asks the user to select the website fields needed for data extraction. The data are then stored in CSV format. Because the forums are led by users, any user may open a discussion. This also means that there are multiple online discussions on *Debate.org* regarding the topic abortion. We therefore searched for all the abortion debates on the website and scraped all their open online discussions. After retrieval the raw data consist of 2,934 documents with the attributes shown in Table 4.1. Each document has a unique document identifier and the textual opinion. As metadata a document has the URL of where the document has been retrieved from, the title of the document and the stance.

Attribute	Data format	Description
Web-scraped-order	String	Document ID
Web-scraped-start-url	String	URL link from where the document has been retrieved
Article	String	Full text of the document
Title	String	Title of the document given by the author
Stance	String	Author's stance of the document (<i>for</i> or <i>against</i>)

Table 4.1: Attributes of the dataset upon retrieval

4.2.2. Data cleaning

We retrieved 2,934 raw documents. Because the debate website is accessible for any user there may be documents not useful for our study. Each document is manually read to remove documents either not containing a clear stance, containing only arbitrary characters (= spam) or is unrelated to the abortion debate.

4.2.3. Data annotation

As we are interested in the topic model's ability to correctly discover perspectives from text, we need to know what perspectives are present in the corpus. For this, we assign a perspective label to each document. The possible perspectives to annotate each document are based on *ProCon*. This source lists 31 possible perspectives on abortion. As observed in Chapter 2 there may be unlimited ways to express a thought to text. This also means that *ProCon* may not cover all the possible perspectives a person has. A perspective-value 'Other' with index 32 has been given if a document cannot be labelled

with one of the 31 perspectives in *ProCon*. With this list of possible perspectives we read each document and label each document based on the present perspectives in the text. Here a document may have one or multiple perspectives. This means that the label is a list of one or multiple perspectives that has been identified in the document.

To ensure that the annotated input data for the topic models is reliable, we ask two independent annotators to annotate a subset of the data. We then use an interrater-reliability measure to calculate the agreement among each independent annotator and the main annotator. If there is an agreement score higher than 0.8 between the main annotators and the two independent annotators, we can assume the annotated data to be reliable. The choice of sample subset and evaluation metric influence the outcome of the reliability score and are further explained below.

Sample data for interrater-reliability score

According to [48] and [49] the following criteria should be met to calculate the interrater-reliability score:

1. Diversity and variability in the data: the sample needs to be representative for the complete data
2. No default category values such as 'Can't answer', 'Not sure': this leads to an easy-way-out in difficult decision-making
3. Large enough sample size: there should be enough data such that each possible value has a fair chance of occurring. Either use a minimum of 10 – 20% datapoints of the complete data or use a sample size summary table.

The created sample data is one that meets the three criteria above and is a sample representative for the input data. This input data is one where each document can have only one perspective-value (Section 4.4 explains this choice). In total there are 1,900 documents containing only one perspective-value. The sample size is based on a combination of the summary table given in [48] and the 10 – 20% rule. We thus create a sample of 400 documents from the 1,900 documents. This sample is then split and annotated by two separate annotators.

Interrater-reliability score

In order to measure the agreement between the annotators we need an interrater-reliability metric. As the values per article are nominal we can use Krippendorff's Alpha. We use the Krippendorff's Alpha in which $\alpha = 1$ means a total agreement among the annotators, $\alpha = 0$ is a random agreement and $\alpha < 0$ is worse than random but may give structural errors. As suggested in [50] a minimum value of 0.8 for α is desired.

In order to measure the reliability of the main annotator, we calculate twice the interrater-reliability between the main annotator and one of the independent annotators. This gives us the average reliability score of the main annotator as seen in Table 4.2.

Main annotator	
Annotator 1	0.73
Annotator 2	0.81

Table 4.2: Interrater-reliability score between main annotator and two independent annotators, using Krippendorff's Alpha

Combining perspectives based on the interrater-reliability score

Table 4.2 shows an agreement score of 0.73 between the first independent annotator and the main annotator. Preferably, we would want to have a data input with an agreement above 0.80. When taking a closer look at the documents with disagreement, the confusion mainly stemmed from using two perspective-values interchangeably. These two are: 1) *Abortion is murder* 2) *Life begins at conception, so unborn babies are human beings with a right to life*. Both perspectives could hold a similar meaning in which both oppose abortion with the believe that an unborn baby is a human being. Because they both hold a strong relation and often led to confusion when annotating documents, we decided to combine these two perspectives into one perspective. This perspective is: *Abortion is murder, because*

unborn babies are human beings with a right to life. To validate the assumption that the agreement will be increased, we asked the second independent annotator to annotate a sample set with this modification. This led to an agreement score of 0.81. Because we value a high reliability score of the data, we combined two confusing perspectives into one. This means that for the remainder of this thesis we have a dataset with in total 30 possible perspective labels instead of 31.

4.3. Exploratory study of the data

There are in total 2,454 annotated documents after removing documents containing spam, inappropriate language or with indifferent stance.

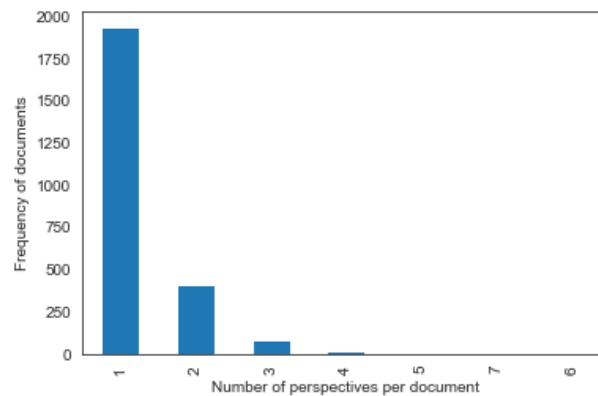


Figure 4.2: Distribution of number of perspectives per document

Because a document may contain multiple perspectives, Figure 4.2 highlights the distribution of number of perspectives per document. It is shown that majority of the corpus holds one perspective, with approximately half of the corpus containing more than one.

To investigate what the distribution of the different perspectives are for the cleaned corpus, we computed a histogram as seen in Figure 4.3. The distribution shows the perspectives through an index-number and how often they appear in the corpus. We notice a large variety in which almost all perspectives from *ProCon* are found. We also find that a small number of perspectives cover the complete corpus. This may affect the topic model as perspectives appearing less frequently may not be found by the model.

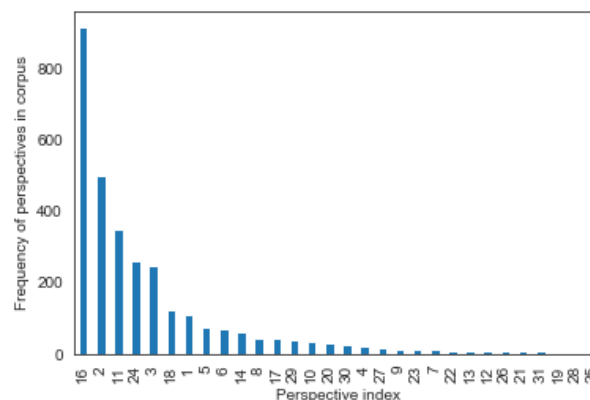


Figure 4.3: Distribution of the perspective indices present in the corpus

For a closer look of the type of perspectives, Figure 4.4 shows the six most expressed perspectives for being for or against abortion.

PRO	CON
<p>Pro 2 Reproductive choice empowers women by giving them control over their own bodies.</p> <p>Pro 3 Personhood begins after a fetus becomes “viable” (able to survive outside the womb) or after birth, not at conception.</p> <p>Pro 11 A baby should not come into the world unwanted.</p>	<p>Con 16 Abortion is murder because unborn babies are human beings with a right to life.</p> <p>Con 18 Abortion is the killing of a human being, which defies the word of God.</p> <p>Con 24 If women become pregnant, they should accept the responsibility that comes with producing a child.</p>

Figure 4.4: The six expressed perspectives



(a) Tokens based on all documents (b) Tokens on all for-documents (c) Tokens on all against-documents

Figure 4.5: Word clouds on the complete data and when data is split based on stance

Figure 4.4 illustrates that each perspective may have a counter-perspective. For example, perspective 2 can go against perspective 18 in which one believes in the woman’s independence and the other may disagree in which God should play the upper-hand. The differences are perhaps even stronger when comparing 3 with 16, both highlight the concept of life in which one believes that life begins at conception and the other believes that life begins after being born. Similarly, perspective 11 goes against perspective 24 in which one can interpret it as a child should not be born unwanted and the other as a child should be born regardless of being unwanted or not. One emphasizes on the woman in which an unwanted child should not be born and the other emphasizes on the child where the woman needs to take responsibility. This shows how controversial this topic is with three pro-and con perspectives being closely related. However, this also questions whether a topic model is able to discover these six distinct perspectives or perhaps is better in discovering the three overarching topics.

Figure 4.5a shows a word cloud of the corpus after tokenization and removing stopwords. The word cloud shows the most frequent tokens present in the corpus. The font size indicates the frequency of the word. The bigger the font size, the more it appears in the corpus. We may derive that many words are quite general with not a strong indication of which words could possibly belong to which perspective. Figure 4.5b and Figure 4.5c are word clouds when splitting documents based on stance. When we separate tokens based on stance, there is not a big difference in word use. This shows that the most occurring words in the corpus are used across both stances, highlighting how homogeneous the corpus is.

When we look at the distribution of unique token frequency we notice a strong Zip’s law distribution. Figure 4.6 shows a snippet of the first 50 tokens in the corpus where a small subset takes up most of the total number of tokens. Of the 3,343 unique tokens, almost 80% only appear a few times. As mentioned in [8] this is a common issue within linguistics and may need to be taken into consideration when computing the topic models.

4.4. Data input for the models

From Figure 4.2 and Figure 4.3 documents may contain one or multiple perspectives with the perspective-values not being evenly balanced. Some documents contain more perspectives than other documents and some perspectives appear more often than other perspectives. For the topic model input data we aim for a dataset where documents are equally distributed over the available perspectives. To ensure that each perspective has an equal chance of appearing, we reduce the datasize to a corpus of documents containing only one perspective and has an equal perspective frequency. To create such a

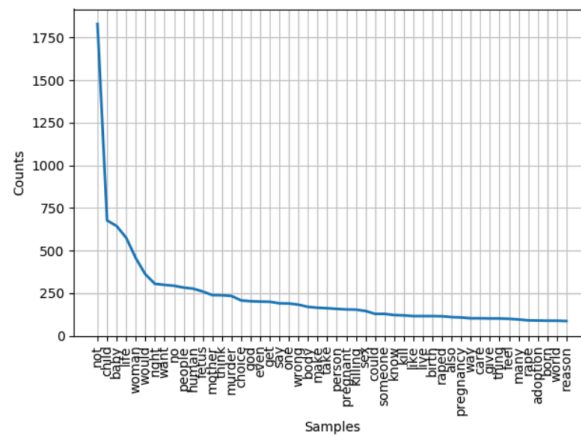


Figure 4.6: Zip's law distribution of the top 50 unique tokens in the corpus. The token 'abortion' is removed here and is seen as a stopword.

dataset, we first select documents with labels being one of the top six most expressed perspectives. This creates a balanced dataset with three pro- and three con perspectives. When we computed the distribution over this dataset there are more than 750 documents that have perspective 16 whereas 100 documents have perspective 18. A topic model may not represent these six perspectives well when some perspectives appear more often in the corpus than others. We therefore randomly select 100 documents per perspective to make a final balanced dataset. We chose 100 because it is the smallest document size per perspective. To conclude, the final corpus consists of 600 documents where each document has only one perspective-value and is one of the six most expressed perspectives presented in Figure 4.4.

4.5. Summary

The raw dataset is a set of 2,934 opinionated documents retrieved from online debate forum *Debate.org*. This dataset is further reduced to 2,454 documents after cleaning the data. The content of the dataset are opinionated user forum entries on the topic abortion. Because the objective is to evaluate several topic models with a ground truth, we annotated each document based on its perspective. The 30 possible annotation values are from an outside source called *ProCon*. Each document may contain one or multiple perspective labels. To make sure this annotated dataset is reliable, we also asked two independent annotators to annotate a subset of the data. We then calculated the interrater-reliability score between the main annotator and two independent annotators. This resulted in a usable dataset with a score above 0.8, indicating the reliability of the annotations.

We also explored the dataset where most documents have one perspective in which the top six expressed perspectives are highly controversial. These perspectives are three pro-perspectives and three con-perspectives that can be viewed as three overarching topics. We created word clouds by counting the frequency of each unique token. From there it is shown that the word use for all documents seem quite similar regardless of stance. This raises the question whether topic models can find the six distinct perspectives, only the three overarching topics or none. For the actual topic model's data input we conclude that we aim for an evenly distributed dataset. We thus adapt the data to a final corpus that is a set of 600 documents. This is a balanced final corpus with three pro- and three con perspectives as possible document labels and where each of these six perspectives is represented by 100 unique documents.

5

Topic models for perspective discovery

5.1. Introduction

The main objective of this thesis is to understand what topic model is able to discover human understandable perspectives on controversial issues. In order to answer this question we need to know what topic models exist and how they can be used for perspective discovery as defined in Chapter 2. This chapter therefore answers **RQ1** *What topic models exist in relation to perspective discovery?*

5.2. Preliminaries

The final corpus for the topic models is a set of documents D about the topic abortion. With $D = \{d_1, \dots, d_m\}$ and $m = 600$, each document d_i is one opinion on abortion labelled with a stance (= *for* or *against*) and a perspective. We define n possible perspective-values. $P = \{p_1, \dots, p_n\}$ is the set of perspective labels with $n = 6$. The value of p_i corresponds to perspective a_i which is one of the perspectives in the set A taken from *ProCon*.

The possible perspective label for d_i is one of the following six perspectives in P :

- p_1 = Reproductive choice empowers women by giving them control over their own bodies.
- p_2 = Personhood begins after a fetus becomes 'viable' (able to survive outside the womb) or after birth, not at conception.
- p_3 = A baby should not come into the world unwanted.
- p_4 = Abortion is murder, because unborn babies are human beings with a right to life.
- p_5 = Abortion is the killing of a human being, which defies the word of God.
- p_6 = If women become pregnant, they should accept the responsibility that comes with producing a child.

The topic models LDA, TAM, JST, LAM and VODUM are used for the experiments where LDA is used as baseline model. Each topic model gives K topics in which each topic k should correspond to p . This means that $K = n = 6$. Each topic k is described with the top 10 words from the per-topic word distribution. We kept the hyperparameters for each model fixed to ensure equal comparison of the models. These hyperparameter values are similar to literature [51] [52] and are: 1,000 for the number of iterations, 0.01 for β , 6 for the number of topics K and $50/K$ for α . Due to the Gibbs sampler having variance in the data, we also ran each model 5 times and computed their average scores.

5.3. Selection of joint topic models

We identified a set of JTM (joint topic models) in Chapter 2 that could potentially be used to find perspectives. The four JTMs that are used in this research are listed in Table 5.1 and are: TAM, JST, VODUM and LAM. The reasons for choosing these four JTMs answers **RQ1** and are listed below.

1. *Better comparison of topic models:* Each JTM has as foundation LDA. This allows us to evaluate what additional component on top of LDA would have more impact for perspective discovery and perhaps also understand why this occurs. Moreover, we can use LDA as the baseline model to evaluate if a JTM would also perform better than a more standard topic model such as LDA in case of perspective discovery.
2. *Topic model's design:* Each chosen JTM has been designed according to the same principle: an author writes a document about the same subject but may write it from a different angle. For example, for the dataset used to evaluate VODUM these angles may be based on the author's demographics (articles on Israel-Palestine conflicts written by Israeli-an author vs Palestinian author) or for TAM's dataset on the author's expertise (research articles on computational linguistics by computer scientists vs linguists). Although they do not explicitly focus on opinionated text from a political angle, we could extend their notion of 'angle' as one based on political belief. In our case we can evaluate whether these models can be used for opinionated documents in which an author may take pro-life angle or pro-choice angle. Moreover, it is not only about using JTMs to distinguish documents *between* pro-life versus pro-choice. We foremost want to evaluate whether these same JTMs can also distinguish documents *within* one stance and find distinctive perspectives. This has not been explored yet by the authors of the existing JTMs.
3. *Code availability:* Our focus lies on understanding and comparing existing topic models, not on building one. This means that we excluded models where no code was available (also after contacting the author). Moreover, this ensures we are comparing models that are created by the original authors and are not subject to possible errors.

5.4. Pipeline of the topic models

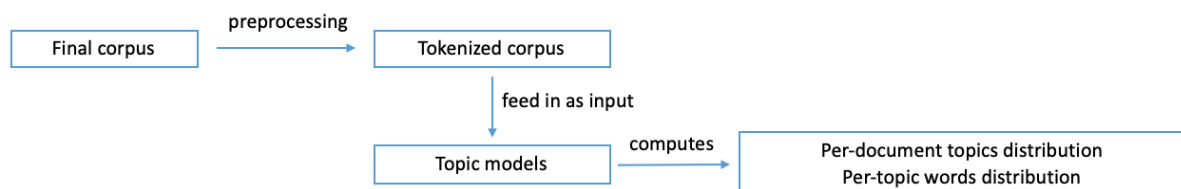


Figure 5.1: Pipeline of how the topic model are being used

Figure 5.1 illustrates the pipeline when running a topic model. Each topic model requires a tokenized corpus. This means that a corpus would need to be transformed into a set of tokens. Recall that the final corpus is one with documents containing only one perspective label and is one of the six perspective values as explained in Chapter 4. To transform these documents we tokenize them according to existing literature for topic models. The following tokenization order has been used: for each document all contractions are removed, then we apply spelling corrections and remove any punctuation and digits. We then lowercase all words and remove stopwords, followed by lemmatization. Section A.2 lists the libraries used to perform this tokenization. The tokenization process is also called *preprocessing* as seen in Figure 5.1. After the final corpus is preprocessed into a tokenized corpus they are feed into the model. After a topic model has processed the tokenized corpus, the model gives a set of multinomial distributions from where we can draw K per-document topics and a top N tokens per topic k in K .

Table 5.1 summarizes the key differences between each topic model and what types of tokens may be drawn for perspective discovery.

Model	Prior knowledge?	Contribution according to literature	per-topic token types
LDA	No	Compute descriptive bag-of-words topics to describe main themes in text in an unsupervised way	baseline tokens
TAM	No	No prior knowledge needed, separating tokens by more granular token frequency levels	background-tokens, topic-independent tokens, topic-dependent perspective tokens
JST	Yes	Separating topics based on sentiment	positive tokens and negative tokens
VODUM	No	No prior knowledge needed, is too costly and domain-dependent Separating tokens based on POS-tags	topic-neutral tokens and perspective-non-neutral tokens
LAM	Yes	Combines POS-tags and sentiment to separate tokens	background-neutral tokens, perspective-positive tokens and perspective-negative tokens

Table 5.1: The key differences between each topic model

5.5. Summary

In order to answer the main research question, **RQ1** questions what topic models currently exist to perform perspective discovery. Based on the literature study we have identified four joint topic models that may perform the task of perspective discovery. These are TAM, JST, VODUM and LAM and we compare each model with each other and the baseline model LDA. These four joint topic models are chosen because of the same foundation being based on LDA. This allows for a better understanding and comparison of each joint topic model. Other reasons are code availability and that they have the same design objective. All models are created with the assumption that a corpus may contain documents about the same subject but could be written from a different angle. Although they have not been used on controversial topics, we extend their reasoning to see whether they can be used for perspective discovery on controversial issues and can distinguish perspectives between and within stances.

6

Evaluating the clustering ability of topic models

6.1. Problem formulation

This chapter focuses on answering **RQ2**: *What topic model is most suitable to correctly cluster documents based on their perspective label?* For this we use the labels of the dataset. The objective is to evaluate whether topic models are able to correctly group documents based on their perspective. A perfect clustering would mean no cluster overlap between documents and that all documents with the same perspective label has been assigned the same most-probable topic according to the topic model's per-document topic distribution. The final corpus and the evaluation pipeline for the topic models are explained in Chapter 5. We made the code and data available and can be found at <https://osf.io/uns63/>.

6.2. Evaluation setting

6.2.1. Method

The objective is to find the topic model that is able to cluster the documents best given ground truth labels. There are five topic models to evaluate: LDA, TAM, JST, VODUM and LAM. For all these models we first preprocess the corpus into a corpus of tokenized documents D . We then run the models and compare the given output with two metrics: Topic Coherence (TC) and adjusted Rand Index (aRI). The latter metric is used to evaluate the clustering performance.

A preliminary study that used the preprocessing techniques as described in Section 5.4 showed low cluster performance. Given the dataset, we observe from these results that a standard preprocessing procedure is not sufficient enough for the specific task of perspective discovery. We therefore introduce an iterative evaluation setup where we apply a number of preprocessing techniques to improve upon the preliminary results. We call the preliminary results that is based on the standard preprocessing procedure as the *baseline* and aim to find the best-performing preprocessing setup in order to answer **RQ2**. Moreover, understanding the best-performing preprocessing is important as the output generated from this setup is used to answer the subsequent research questions which employs a user study. Figure 6.1 illustrates the evaluation setting. The subsequent sections explain the different preprocessing functions being used and the evaluation metrics.

6.2.2. Data preprocessing and tokenization

When we analyzed the baseline results we identified two observations impacting the low performance.

First, there is a high number of irrelevant or neutral tokens in the output. For example, there are tokens with *"should"*, *"would"*, *"therefore"* which do not describe any perspective. The second observation is that documents are written too similar in which the distinctions based on tokens are hard to make. For example, *"this is wrong because the mother deserves to have a choice"* versus *"this is wrong because the child deserves to have a choice"* – both are about choice but one is for-abortion with the women's right to choose and the other against-abortion with the children's right to life.

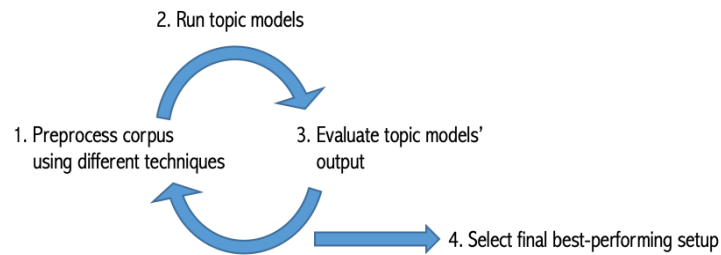


Figure 6.1: The iterative process of finding the most suitable tokenized data input to answer **RQ2**

The additional created preprocessing functions are based on these two identified observations and are in line with the literature in Chapter 2. These are: removing too frequent or too rare tokens (= outliers), removing happaxes, removing tokens not holding any sentiment and the implementation of different n-grams. Other techniques are based on valence shifters such as changing tokens into antonyms when needed.

6.2.3. Evaluation metrics

Topic Coherence

TC is a common metric for topic models. As explained in Chapter 2 the topic coherence score indicates how coherent the topic words are in each topic and how well the words within a topic support each other. Based on previous research, a model such as LAM may reach TC scores between 0.4 – 0.6. The assumption is that topic models with higher coherence score would be easier to understand for a user and are coherent to describe the main themes of a corpus. We use metric c_v as explained in Chapter 2 to calculate TC. With the use of TC, we want to evaluate the general quality of the topic models and whether the models reach the same scores on this dataset as in literature.

Adjust rand index

A topic model calculates a topic distribution for each document. Using this distribution we assign the most probable topic per document. These topic labels can be seen as a way to cluster the document set D . Here X is the total cluster set for D where each element X_i in X is a set of documents with the same most-probable topic k assignment. A cluster X has therefore X_1, \dots, X_r elements with r equals the total number of topics produced by the topic model. To evaluate this cluster we need a second cluster. Cluster Y is the cluster based on the ground truth labels. Each element Y_i in Y is a set of documents with the same ground truth perspective label. Cluster Y therefore equals Y_1, \dots, Y_s with s equals the total number of perspective labels in the final corpus.

We use aRI to measure the clustering which is an extension of Rand Index (RI) [53]. Because each document d has only one perspective label, we can use (a)RI to compute the agreement between the computed clusters and a ground truth [54]. RI can be calculated for the two clusters X and Y as follows, given a set S with n elements where n equals $r * s$:

$$RI = \frac{a + b}{a + b + c + d} = \frac{a + b}{\binom{n}{2}}$$

with

a = number of pair elements in S occurring in the same subset of X and Y

b = number of pair elements in S occurring in a different subset of X and Y

c = number of pair elements in S occurring in same subset of X but different subset in Y

d = number of pair elements in S occurring in different subset of Y but same subset in Y

aRI is an extension of RI incorporating correction-of-chance with values ranging from -1 to 1 and can be calculated as [55]:

$$aRI = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{n}{2}}{\frac{1}{2} [\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{n}{2}}$$

As derived from RI a high score may be obtained when a and b are high. These can also be seen as *True Positives* and *True Negatives* respectively. In this case True Negatives may be increased by introducing a high number of K clusters. This artificially increases the RI score as the probability of having documents not being in the same subset X and Y is now being increased. aRI prevents this by normalizing the values with a correction-of-chance and is therefore the preferred metric.

We use aRI to measure how well the five identified topic models are able to cluster the documents with similar perspectives together, given a ground truth. Because each JTM is based on LDA we would expect JTM performing better than LDA.

6.3. Finding the most optimal preprocessing techniques

We computed for the five topic models the TC and aRI scores. To improve the baseline results we also applied a number of preprocessing functions. These preprocessing functions were based on two observations: 1) a high number of irrelevant or neutral words in the final corpus and 2) many documents are written too similar in which distinctions between words are hard to make.

Removing irrelevant tokens: sentiment, outliers, and zips law

'baseline'-column in Table 6.1 illustrates TC values ranging between 0.30 – 0.50. This is somewhat low compared to the results found in literature where models such as LAM are able to reach 0.60. aRI scores are also close to 0 for most models. This means that most documents are randomly assigned without a clear reason for their clustering. A few preprocessing techniques are used to reduce the negative impact of the first observation, and we compare this with the baseline method.

Model	Metric	Baseline	Sentiment	Outliers	Zips law	BiUnigrams	TriBiUnigrams	Negation	Antonyms
LDA	aRI	0.014	0.015	0.003	0.007	0.003	0.004	0.015	0.003
	TC	0.366	0.332	0.310	0.342	0.627	0.834	0.376	0.345
JST	aRI	0.097	0.081	0.062	0.074	0.083	0.079	0.074	0.090
	TC	0.385	0.374	0.458	0.372	0.571	0.746	0.409	0.410
TAM	aRI	0.125	0.150	0.108	0.132	0.134	0.129	0.045	0.105
	TC	0.402	0.396	0.372	0.435	0.403	0.662	0.314	0.333
LAM	aRI	0.058	0.029	0.058	0.061	0.039	0.042	0.065	0.045
	TC	0.506	0.504	0.514	0.433	0.503	0.520	0.529	0.490
VODUM	aRI	0.018	0.022	0.0004	0.000	0.203	0.137	0.026	0.004
	TC	0.395	0.423	0.437	0.451	0.622	0.814	0.397	0.422

Table 6.1: Influence of the data input on the adjusted rand index and topic coherence scores per model

The first three columns after the baseline in Table 6.1 focus on the first observation meaning to reduce the number of irrelevant tokens as data input. As described in Chapter 2 not all tokens in a document are relevant to determine a sentiment or aspect. Removing words not holding any sentiment may improve the classification as general words holding a neutral stance are removed. We call this *non-sentiment removal* and its influence on the model output is shown in column 'sentiment' of Table 6.1.

Another method to remove irrelevant words is by removing the top common words and the top rare words. We call this *outlier removal* in which we empirically removed all tokens with frequency above and below five standard deviations away from the mean. Its effect is shown in column 'outliers'.

For LAM the authors of [8] removed the top 99% and lowest 65% of the tokens due to the Zip's law distribution. Similarly, when exploring the data in Chapter 4 there is a similar distribution in our corpus. This distribution is mostly due to the large number of happaxes. Happaxes are tokens that appear only once in a corpus and in our case almost 50% of the total number of unique tokens occurred only once. This is a form of irrelevancy and removing the happaxes results in the scores shown in column 'zips law'.

Based on these three techniques the performance did not increase by a lot. For some models it even decreases. This behaviour may be explained from the design of the topic models. All models aim

to reduce the irrelevancy of tokens either through sentiment (e.g., in JST and LAM), grammar structure (e.g., in VODUM and LAM) or granularity of words (e.g., in TAM). Removing tokens with no relevancy prior to running these models may confuse the models, because all tokens in the corpus would have a similar token frequency. This makes the distinction between tokens more difficult as the topic models are built on co-occurrences of tokens.

Increasing the variability of tokens: n-grams, negation, and antonyms

The second observation is to increase the set of unique tokens to create a better distinction between documents.

A possibility to increase variability is to take into account word order. This is done through n-grams in which n determines the number of subsequent tokens being glued together as one token. For example, a bigram would combine two subsequent tokens together with an underscore and trigrams combines three subsequent tokens together. When we apply bigrams, we would have tokens as *'child_choice'* and *'mother_choice'* instead of separating the tokens to *'child'* *'mother'* *'choice'*. Such a word order may help in better understanding the meaning of tokens as *'child_choice'* could be more in favor of the child and *'mother_choice'* could be more in favor of the mother. The results are shown in column *'BiUnigrams'* and *'TriBiUnigrams'* in which we created several n-grams of one corpus.

Another common mistake made by the model is not distinguishing *'murder'* and *'not murder'*. One could be about being against-abortion as abortion is seen as murder and the other being for-abortion as abortion is not seen as murder. However, the model does not recognize this because it is based on co-occurrences and both hold the token *'murder'*. To create such distinctions we applied valence shifters, often used in sentiment analysis. These valence shifters may transform a positive word into a negative word if it precedes a negative word. For this we employed two techniques. One is to create a bigram meaning that valence shifters are created by adding underscores between two tokens (e.g. "not murder" will be one token "not_murder"). The influence of this feature is shown in column *'negation'* of Table 6.1. However, this is not be recognized by all topic models such as JST and LAM because they use prior-knowledge through a subjectivity lexicon. Bigrams such as *'not_murder'* are not present in the lexicon so they are be ignored by the model. Another option is then to use antonyms. For example, *'not available'* would get as antonym *'unavailable'* whereas the bigram techniques changes it to *'not_available'*. These antonyms would be a better choice for models using prior-knowledge. Their influence is shown in column *'antonyms'*.

We learn from these techniques that n-grams lead to a performance increase. The strongest increase compared to the baseline is for VODUM. Here, aRI score increased from 0.018 to 0.203.

Combining preprocessing techniques

From Table 6.1 the second observation of increasing the token variability has more impact than the first observation. Although this holds when applying all techniques independently, there is a difference when techniques are combined. These results are shown in Table 6.2. Combining antonyms and sentiment with tri-bi-unigrams resulted in an overall higher score than only tri-bi-unigrams. When only applying tri-bi-unigrams the model's output tokens still has neutral words such as *'would'*, *'get'* and *'should'*. This effect is reduced by creating combinations of preprocessing techniques.

Model	Metric	Baseline	NegSentZip	BiZip	AntSentBiUni	AntBiUni	SentBiUni	AntSentBiUniOut	AntSentTriBiUni
LDA	aRI	0.014	0.000	0.005	0.008	0.007	0.006	0.003	0.015
	TC	0.366	0.396	0.455	0.606	0.628	0.640	0.412	0.809
JST	aRI	0.097	0.091	0.090	0.130	0.075	0.138	0.087	0.108
	TC	0.385	0.423	0.447	0.600	0.592	0.526	0.438	0.727
TAM	aRI	0.125	0.112	0.134	0.194	0.139	0.060	0.035	0.182
	TC	0.402	0.516	0.379	0.595	0.574	0.403	0.321	0.719
LAM	aRI	0.058	0.060	0.045	0.035	0.056	0.059	0.040	0.061
	TC	0.506	0.498	0.466	0.516	0.527	0.526	0.524	0.500
VODUM	aRI	0.018	0.003	0.012	0.174	0.146	0.163	0.005	0.178
	TC	0.395	0.455	0.443	0.583	0.615	0.623	0.470	0.775

Table 6.2: Influence of combining preprocessing techniques on the adjusted rand index and topic coherence scores per model

To conclude, the models using a combination of antonyms, non-sentiment removal and tri-bi-unigrams performed best. Except for LAM, TC is above 0.7 for all models. LDA has the highest TC

score and LAM the lowest. Regardless of TC score, all models have aRI scores close to 0. There are some nuances with JTM's performing better than LDA with higher aRI scores. This shows that each added contribution of a JTM does improve the performance. When comparing each JTM, the results show that TAM performed best followed by VODUM, JST and LAM.

6.4. Qualitative analysis of the final topic model setup

After several evaluations of the topic models with different preprocessing functions we have a final topic model setup that results in the overall highest aRI score. This is one where the corpus is tokenized using the baseline techniques with the additional techniques: antonyms, non-sentiment removal and the inclusion of tri-bi-unigrams. With this setup we computed the output of the topic models for a qualitative analysis. Each topic model creates six topics with each topic describing one perspective. These perspectives are described by the top 10 tokens that are drawn from the per-topic words distribution. The six topics of each topic model are shown in Figure 6.2

LDA	JST
1 [life, not, baby, child, woman, murder, right, want, people, mother]	1 [not, woman, people, choice, body, care, well, option, kid, responsibility]
2 [not, baby, child, woman, life, fetus, want, god, right, human]	2 [not, person, wrong, kill, no, yet, week, pro, point, alive]
3 [not, child, woman, baby, life, people, human, want, fetus, think]	3 [fetus, human, murder, killing, cell, another, living, still, really, mean]
4 [not, child, life, baby, people, want, woman, fetus, right, murder]	4 [child, mother, life, give, pregnancy, parent, adoption, time, choose, reason]
5 [not, baby, child, life, woman, want, right, fetus, people, human]	5 [life, right, baby, god, believe, feel, people, innocent, odd, even]
6 [not, child, baby, life, woman, right, think, mother, people, want]	6 [baby, want, get, think, pregnant, sex, not_want, no, way, rape]
TAM	VODUM
1 [woman, choice, body, fetus, control, pregnant, birth, baby, foetus, sex]	1 [want, pregnant, right, think, get, child, woman, life, people, baby]
2 [fetus, human, brain, person, fetus_not, cell, murder, alive, killing, egg]	2 [human, right, wrong, even, dead, life, fetus, baby, murder, woman]
3 [sex, woman, pregnant, parent, forced, child, want, child_not, option, unwanted]	3 [human, think, even, not_want, give, pregnant, child, baby, woman, life]
4 [god, life, wrong, child, womb, baby, murder, killing, kill, god]	4 [want, think, not_want, give, pregnant, child, baby, woman, life, mother]
5 [want, woman, sex, not, responsibility, child, get, not_want, pregnant, choice]	5 [killing, think, want, pregnant, kill, baby, child, life, woman, people]
6 [life, god, begin, baby, life_begin, choice, choose, use, protection, responsibility]	6 [right, kill, want, wrong, human, life, god, baby, child, murder]
LAM	
1 [prevent, die, nature, organ, infant, bad, not_really, people_want, sick, life_child]	
2 [murder_baby, god_creation, baby_fault, creature, poverty, deserves, bad, emotional, immoral, another_human]	
3 [body_not, innocent_baby, life_begin, life_support, line, well, carry, good, not_care, love]	
4 [use_protection, call, order, homicide, pas, good, sexual, result, okay, love]	
5 [child_right, woman_baby, foster_care, terminate_pregnancy, opportunity, pro, beating, best, yes, honestly]	
6 [sex_not, foster, pleasure, someone_baby, face_consequence, support, good, woman_want, pro, best]	

Figure 6.2: Top 10 tokens per topic per topic model

The top 10 tokens for each topic in LDA show high similarity with a few nuances. For example, there are some tokens not present in all the topics such 'murder' and 'god', but also many tokens appearing in all topics such as 'not' and 'child'. In that respect, the four JTM's show more variety in the tokens per topic and indicate that JTM's are able to distinguish more diverse tokens than LDA.

Recall that the dataset about abortion is highly homogeneous. Words as 'murder' is used in almost any document and could confuse a topic model to describe a specific perspective. For example, in Figure 6.2 it is not always clear whether LDA topic 1 or topic 3 is describing perspective p_4 in P . This shows the difficulty to describe a perspective with merely a set of tokens and that tokens may be interpreted in different ways. It further raises the need for a user study to investigate how each topic is being understood by humans and how strong these differences are between topic models.

6.5. Summary

This chapter focused on **RQ2**. The clustering ability of topics models were analyzed by comparing the most-probable topic per document with its ground truth label. We compared this by using the aRI metric. During a preliminary study, two observations have been identified that may cause the low performance for the baseline tokenization technique. We therefore have an iterative evaluation setup where we applied different preprocessing techniques to improve the overall model's performance. From there we conclude that the best-performing technique is one that includes antonyms, sentiment removal and n -grams on top of the baseline tokenization. With this preprocessing method, TAM has the highest aRI score followed by VODUM, JST, LAM and LDA. Although topic coherence is a widely used metric to

evaluate topic models, the aRI values show that such a metric is not sufficient for the specific task of perspective discovery. For example, LDA has the highest TC score, but has the lowest aRI score.

To evaluate how the topics describe perspectives with the above setup, we also drew for each topic the top 10 tokens from their per-topic words distributions. There LDA computed more similar tokens per topic than JTMs. JTMs computed more diverse tokens per topic which makes the computed topics more distinct and descriptive. Furthermore, each topic has been described differently but may have the same perspective-label. This shows that documents can be interpreted in different ways and each topic model gives their own interpretation on each of the six perspectives. A user study would provide more insights as to how the topics are being understood by users and if they are explanatory enough.

To answer **RQ2**: joint topic models perform better than LDA according to the aRI score. When analyzing the joint topic models with each other we conclude that TAM is most suitable to correctly cluster documents. This is followed by VODUM, JST and LAM.

7

Human understandability of topic models

7.1. Problem formulation

In Chapter 5 we explained the identified joint topic models and the final corpus. This final corpus is a balanced set of 600 documents with each document having one of the 6 perspectives in set P as label. The subsequent Chapter 6 centered around the clustering ability of the five identified topic models regardless of how the model output looks like. It questioned whether a topic model can correctly cluster documents based on a perspective. From these results we have derived a final topic model setup that is used for this experiment.

This experiment focuses on answering **RQ3** and **RQ4**. **RQ3** is *What topic model is most informative to correctly explain perspectives to humans?* **RQ4** is *Do users choose more perspectives in line with their own pre-existing stance?* In this chapter we investigate how well a topic k from the final topic model setup can describe perspective p such that a human is able to recognize and understand p . For this we conduct a user study. Moreover, we investigate whether users tend to choose more perspectives in line with their own pre-existing stance when judging a topic model and answers **RQ4**. Here we introduce the concept of *false consensus effect* [47] as explained in Chapter 3. Whereas **RQ2** and **RQ3** questions the model's performance, **RQ4** concentrates on how topic models are perceived based on human characteristics.

7.2. Experimental setting

We conducted a between-subjects user study to answer research questions **RQ3** and **RQ4**. For this between-subjects user study we have six models as the independent variable. The dependent variable indicates how many of the six correct perspectives in P users have found. Users would be able to find and understand all the six correct perspectives in case of a perfect model.

The six models in this user study were the four JTMs explained in Chapter 5, LDA and a random model. We view the latter two models as two baseline models. Recall that each JTM is based on LDA in which the results from literature showed that JTMs are an improvement on LDA, computing topics that are more coherent. We therefore use LDA to evaluate whether JTMs also perform better than LDA in case of perspective discovery. The random model is a second baseline model and is a control condition to analyze how many perspectives users get correct if a topic model would not help them at all. This second baseline is based on Term Frequency–Inverse Document Frequency (TF-IDF) and is another method used within information retrieval. We create this baseline model by creating n buckets (= topics). Each bucket contains ten words that are randomly chosen from the top $n * 10$ words computed by TF-IDF. To prevent confusion we call this random model as a TF-IDF model throughout this thesis.

For **RQ3** we performed a statistical analysis to analyze how many of the six correct perspectives are found between the six models and if there is significant difference between each of them. In this case we expect JTMs to perform better than the two baselines TF-IDF and LDA. For **RQ4** we also used the collected participant's data. This is used to analyze the user's own stance towards abortion on the

chosen perspectives.

The user study used *Qualtrics*¹ to create the study and the online platform *Prolific*² to distribute the user study. Because the user study used actual users, we made sure we first get an approval from the ethics committee and data steward. This also means that each user needed to read and agree to the provided consent form in the user study. Moreover, all data are anonymized.

7.2.1. User task

The task of a participant was to evaluate one of the six models. Each model gave an output of K topics with k described by their top 10 words. A user sees in total eight topics with six being the topics computed by the model and two being honeypot checks. The honeypot checks are often referred to as hidden quality control tests to reject unconscientious participants who may perform the user study not according to the instructions [56]. For this user study the two honeypot checks were two topics that are an exact copy of a randomly chosen perspective from the *Procon* list. These two perspectives were the pro-perspective “*Abortion is justified as a means of population control*” and con-perspective “*Abortion eliminates the potential societal contribution of a future human being*”. To illustrate, when users see a honeypot’s topic: {‘abortion’ ‘is’ ‘justified’ ‘as’ ‘a’ ‘means’ ‘of’ ‘population’ ‘control’}, they needed to match this with the perspective: *abortion is justified as a means of population control*. We would therefore want to accept only the participants who have properly read the eight given topics.

Given a list V of 16 perspective options, a user thus evaluated eight topics where six of them should be one of the correct perspectives p of P . The list V equals twice the number of topics n a user sees. This gave us 16 perspective options for V as shown in Table 7.1. Eight of them corresponded to the eight given topics (which was set P and the two honeypot checks) and the other eight was a randomly chosen subset from the 30 perspectives given by *ProCon*. Moreover, V was a list of 16 perspectives with an even number of pro- and con perspectives³. Choosing the number of perspectives in V larger than n allows us to evaluate whether a topic model is descriptive enough to let a human identify the six correct perspectives P that are present in the corpus. These six should have higher probability of being chosen than any other random perspective in V with probability $1/16$ ($= 0.0625$).

Index	Perspective label
p_1	Reproductive choice empowers women by giving them control over their own bodies.
p_2	Personhood begins after a fetus becomes ‘viable’ (able to survive outside the womb) or after birth, not at conception.
p_3	A baby should not come into the world unwanted.
p_4	Abortion is murder, because unborn babies are human beings with a right to life.
p_5	Abortion is the killing of a human being, which defies the word of God.
p_6	If women become pregnant, they should accept the responsibility that comes with producing a child.
$p_{honeypot1}$	Abortion is justified as a means of population control.
$p_{honeypot2}$	Abortion eliminates the potential societal contribution of a future human being.
p_{7other}	The US Supreme Court has declared abortion to be a fundamental right guaranteed by the US Constitution.
p_{8other}	Fetuses are incapable of feeling pain when most abortions are performed.
p_{9other}	Women who receive abortions are less likely to suffer mental health problems than women denied abortions.
$p_{10other}$	Abortion gives pregnant women the option to choose not to bring fetuses with profound abnormalities to full term.
$p_{11other}$	Fetuses feel pain during the abortion procedure.
$p_{12other}$	Abortion reduces the number of adoptable babies.
$p_{13other}$	Select abortion based on genetic abnormalities (eugenic termination) is overt discrimination.
$p_{14other}$	Women should not be able to use abortion as a form of contraception.

Table 7.1: The total list of perspectives V shown to a participant. p_1 to p_6 are the correct perspectives that are present in the corpus. $p_{honeypot1}$ and $p_{honeypot2}$ are the honeypot checks and the other eight perspectives are the randomly chosen perspectives from the *Procon* list. List V has an even number of pro- and con-perspectives.

¹www.qualtrics.com

²www.prolific.co

³Due to high cognitive load we did not give all the 30 perspective options and reduced the size based on the number of topics a user sees.

7.2.2. Pilot study

Prior to the actual user study we recruited 20 participants to conduct a final pilot study. These participants were recruited by directly contacting people in our personal networks. We used the pilot study to evaluate the average duration of the survey, number of correct honeypot answers and to gain an initial idea of the results. Based on these results we adjusted the models and have refined the user study design to make the study more understandable. We also learned that around 20% of the participants answered one or two of the honeypot checks wrong. Half of them answered at most one correctly. At least one of the honeypot checks needs to be answered correctly before we accept the user's data input.

7.2.3. Participants

A power analysis was used to determine the total number of required participants ($f = 0.3, \alpha = 0.05, 1 - \beta = 0.8$). This resulted in at least 150 participants needed to evaluate the six models. We used software *G*power* to perform the power analysis [57]. Based on this analysis and the pilot study we recruited 170 users to evaluate a randomly assigned model with 158 having passed the honeypot criteria which is above the needed 150 participants. Furthermore, we asked native-English participants to perform the study. The participant's age varied between 18 – 64 years with 49.4% male and 50.6% female. We also implemented a prescreening filter on stance to create a balanced pool of participants. The platform *Prolific* provides this filter option to choose the number of pro-choice and pro-life needed for the user study. Although we have applied this option the eventual participants are more pro-choice than pro-life as seen in Table 7.2. Lastly, most participants believe they are highly familiar with the topic with 51.3% 'strongly agree' and 16.5% 'somewhat agree' on a 5-point likert scale, with 0% of the users choosing the extra option 'I don't know'.

stance value	label	Frequency	Percent
1	Strongly disagree	16	10.1
2	Somewhat disagree	19	12.0
3	Neutral	16	10.127
4	Somewhat agree	26	16.5
5	Strongly agree	81	51.3
6	I don't know	0	0
Total		158	100.000

Table 7.2: Frequencies based on stance

7.2.4. Procedure

Question	Question type	Possible answers
Step 1		
(1) What is your age?	Slider	1 – 100
(2) What is your gender?	Single answer	Male, Female, Prefer not to say
(3) In my opinion abortion should be legal	Likert-scale	Scale from Strongly disagree to Strongly agree, with extra option 'I don't know'
(4) I have good knowledge about the abortion debate	Likert-scale	Scale from Strongly disagree to Strongly agree with extra option 'I don't know'
Step 2		
(5) Use the dropdown to select the viewpoint for a word group	Drop-down menu	Values in list <i>V</i> , see Table 7.1
Step 3		
(6) A model that can automatically show all viewpoints is useful to quickly understand a debate	Likert-scale	Scale from Strongly disagree to Strongly agree
(7) I'm now better aware of the possible viewpoints than before	Likert-scale	Scale from Strongly disagree to Strongly agree
(8) I'm confident that I've correctly assigned the viewpoints to the word groups	Likert-scale	Scale from Strongly disagree to Strongly agree
(9) If you have any feedback, please let us know	Open question	-

Table 7.3: An overview of the questions being asked in the user study

After participants read and agree to the conditions stated in the informed consent, they started

with the user study. The user study consisted of three steps. Table 7.3 provides an overview of the questions that were asked. Appendix B shows the screenshots of the complete user study interface.

Step 1 The first step is to understand the user pool participating in the user study. Here we asked them their age and gender. We also asked for their own stance towards abortion and how well they know the topic abortion. For these last two questions we used a 5-point likert scale from *Strongly disagree* to *Strongly agree* and an extra option 'I don't know'. The latter option was added because we cannot assume that everyone is familiar with the abortion topic.

Step 2 The second step is the actual main task as described in Section 7.2.1. This is illustrated in Figure 7.1. We used this step to evaluate how many unique perspective labels users on average chose per topic per topic model as well as the average number of correct perspectives found per topic model. As observed from the results in Chapter 6 the computed topics are different for each model. For example, when comparing JTM with LDA we derive that LDA computes six highly similar topics where the distinctions between them may be difficult to make for a user. In the end, **RQ3** aims to answer which topic model is most informative to let users correctly identify a perspective. A topic model that computes too many ambiguous topics would not lead to a clear informative perspective description, leaving users to make their own interpretations on the given topic. In this case we would expect users to choose less unique perspective labels per topic for JTMs as well as choosing more of the six correct perspectives per JTM. In the second step, users therefore had to read the instructions and then needed to evaluate the eight topics as described in Section 7.2.1. Topics 1 and 7 were the honeypot checks. The other six topics were the topics computed by the model. By using the drop-down menu they assigned a corresponding perspective from list V . List V are the perspectives as seen in Table 7.1. Participants could only choose one of the 16 perspectives in V per topic and no topic could have the same perspective. They got an error when these constraints were not met.

Step 3 In the third step we gave the participant the opportunity to review their experience with the models, because we also want to understand the possible potential a topic model has for a user. We asked three questions on a 5-point likert scale as shown in Table 7.3. The first question is posed to investigate whether a user, who has evaluated a topic model, thinks a model that can automatically show all perspectives can be a useful tool to understand a debate. The next question asked how much their view on abortion has broadened after conducting the user task. Here we can compare the results of this question with the fourth question in step 1 of the user study. If they are aware of more perspectives after conducting the user study then it can highlight the potential usefulness of topic models to increase the understanding of complex debates. The last question asked users their confidence level when evaluating the models. Whereas the first question in step 3 investigates whether users see the possible potential of a topic model, this last question focuses on whether topic models are currently good enough to perform perspective discovery. For example, it is possible that users believe a topic model can be useful, but they may not be perceived well enough to make them useful. In this case their own confidence could be low when judging topic models. Lastly, if needed they were also given the option to provide feedback on the user study.

7.2.5. Statistical analyses

One-way ANOVA

With the user study results we performed a one-way ANOVA with significance level of 0.05. We used the software *JASP* to perform the analysis [58]. We performed this statistical analysis to evaluate the mean differences of correctly identifying the six perspectives between topic models. In this case the null hypothesis is that all topic models, including the TF-IDF model, have the same average number of correctly identifying perspectives by users. The null hypothesis is rejected if there is a significant difference in terms of how many perspectives participants could correctly identify. If the one-way ANOVA gives a significant difference, we perform post-hoc tests with Bonferroni correction for p-values to evaluate where the significant differences lie. This means that with the testing of multiple hypotheses (i.e., $\binom{6}{2} = 15$), we only regard p -values below $\frac{0.05}{15} = 0.003$ as significant.

In order to perform this statistical analysis we also analyzed whether the three required criteria for such an analysis is met [59]. The three criteria are: 1) the data needs to be normally distributed 2)

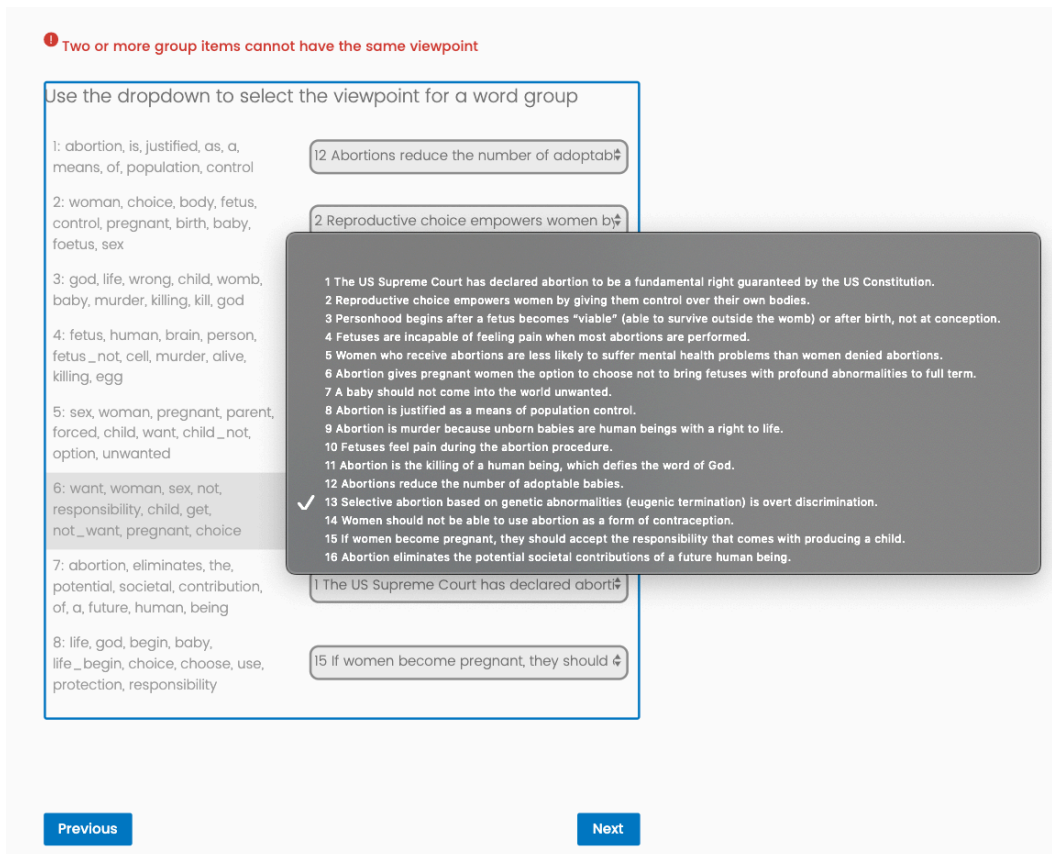


Figure 7.1: Snippet of the second page of the user study. Participants had to assign a perspective to the correct topic. An error was given if the criteria stated in the instruction has not been met. Here the viewpoints refer to perspectives.

the data is homogeneous and 3) the variables are independent.

Spearman correlation

To understand whether there is correlation between the variables stance and perspectives, we calculated the Spearman correlation. We used the Spearman correlation which is a non-parametric test for the correlation between two variables [60]. The two variables in this test are stance and perspectives. For the first variable we use question 3 of the user study which is the user's stance on abortion. Here, the users' answers fell on a 5-point likert scale with 1 being *Strongly disagree* and 5 being *Strongly agree*. The distribution of this scale is shown in Table 7.2. The second variable stemmed from question 5 where we took the number of chosen pro-perspectives for each user⁴.

As the correlation value may range between -1 to 1 , the correlation should be 0 if users ignored their own stance to judge a topic model and choose three pro-perspectives and three con-perspectives. If the correlation value is below 0 with a significant p-value then the number of pro-perspectives would decrease when the stance value increases. This means that users choose less perspectives in line with their own stance. However, when the correlation value is above 0 with a significant p-value then the number of pro-perspectives would increase when the stance values increases. In this case users choose more perspectives in line with their own stance.

Hypotheses

For this user study we pose two hypotheses H_1 and H_2 where we make use of one-way ANOVA and Spearman correlation respectively.

- Hypothesis H_1 for **RQ3** is: *Users identify more correct perspectives for joint topic models than the two baseline models LDA and the random model that is based on TF-IDF.*

⁴Due to symmetry between pro- and con perspectives, we do not calculate the Spearman correlation between stance and number of con-perspectives.

- Hypothesis H_2 for **RQ4** is: *Users choose more perspectives in line with their own stance when judging a topic model.*

7.3. Results

Descriptive analysis of topic model performance

The main task for users is to assign per topic k a perspective label p . Figure 7.2 is the distribution of chosen perspectives by participants across all the models. The x-axis illustrates the possible perspectives V , excluding the two honeypot checks. The perspective labels ranging from p_1 to p_6 correspond to the six correct perspectives P and the labels from p_{7other} to $p_{14other}$ are the randomly chosen perspectives. The y-axis is the total frequency of a perspective being chosen. The five most chosen

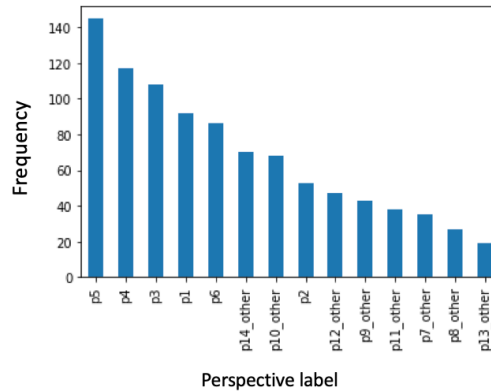


Figure 7.2: Distribution of chosen perspectives by the users across all the models. Perspective labels are on the x-axis and frequency on the y-axis.

perspectives illustrated in Figure 7.2 are also five of the six correct perspectives in P . Although this may seem high, it is uncertain whether this behaviour occurs because of the model's performance or because these five perspectives are the most well-known perspectives for the topic abortion. In the latter case users may have chosen these perspectives because they are most familiar with them. We therefore have to analyze the results between different models as explained in Section 7.2.1. Recall that the TF-IDF model is used to evaluate the usefulness of topic models compared to a user's own knowledge and LDA to analyze the performance of JTMs.

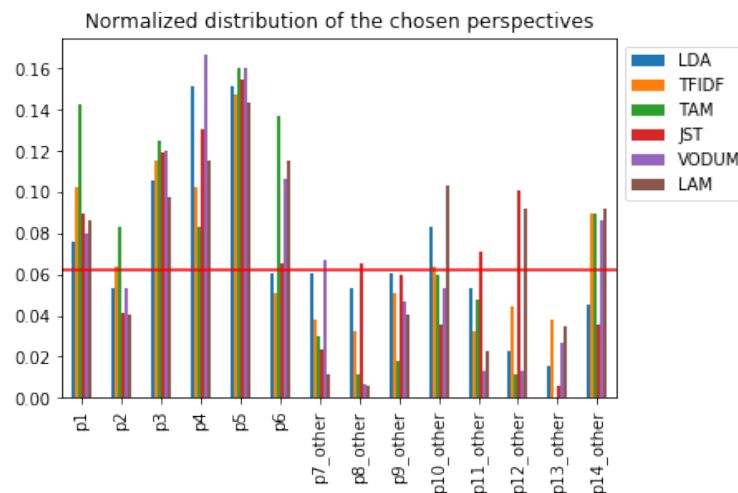


Figure 7.3: Normalized distribution of how often each perspective in V (excluding honeypot checks) has been chosen. The true perspective labels can be found in Table 7.1. The red line equals 0.0625 ($= 1/16$) and is the probability of an arbitrary chosen perspective in list V . For a perfect model, the six correct perspectives should be chosen more often than any other random perspective with probability 0.0625.

Figure 7.3 illustrates the normalized distribution of the chosen perspectives per topic model. Section 7.2.1 explained that each perspective in P should have a higher probability of getting chosen than any other perspective in V . In case of a random judgement, each perspective would have $1/16$ probability of getting chosen, which equals 0.0625 and is indicated by the red horizontal line in Figure 7.3. For all the six models, four out of the six correct perspectives are above the red line and are p_1, p_3, p_4, p_5 . A question raised in Section 4.3 was whether the topic models may find all the six correct perspectives, the three overarching perspective topics or none. From Figure 7.3 we can derive that at least the three overarching perspective topics are identified for all topic models. Recall that the three main topics are (p_1, p_5) , (p_2, p_4) and (p_3, p_6) as mentioned in Chapter 4 and at least one perspective of each pair is above the red line. We also notice differences in perspective distribution between the models where users chose almost equally often perspective p_5 than in the case of perspective p_4 or p_1 . Random perspectives indicated with ‘_other’ are for most models also below the red line with a few above the red line such as p_{10_other} , p_{12_other} and p_{14_other} . Another observation is that less users chose perspectives p_2 and p_6 for models such as LDA and TF-IDF and could suggest that they are the most difficult perspectives to identify out of the total six correct perspectives. Moreover, users who have evaluated TAM chose more of the six correct perspectives than the random perspectives, whereas users chose more the random perspectives than the six correct ones when evaluating the other topic models. For example, p_{8_other} and p_{13_other} have been chosen little when evaluating TAM compared to the other models. To evaluate what other perspectives have been chosen instead, we analyze the unique chosen perspectives per topic, per topic model. Figure 7.4 is an example of what participants chose for a topic for two different topic models.

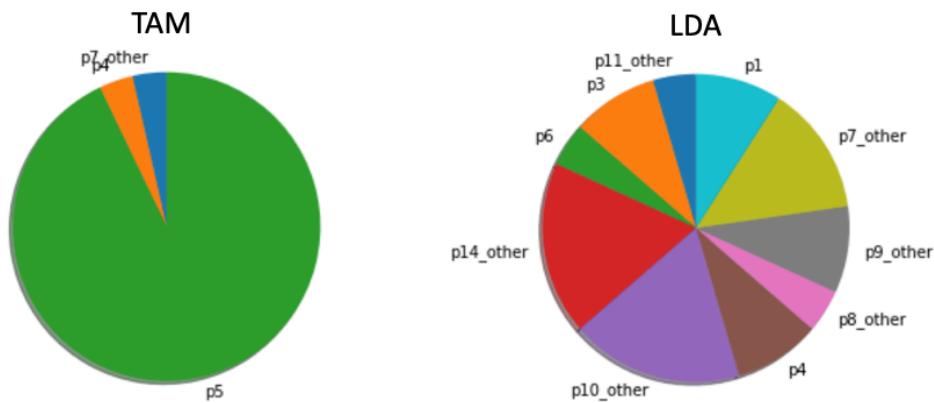


Figure 7.4: The difference between LDA and TAM in the user’s chosen labels on a given topic. Users gave 10 different answers for LDA and 3 for TAM.

We find that participants chose for LDA’s topic k 10 unique perspective labels whereas participants evaluating TAM gave 3 different answers. Moreover, for LDA the majority of participants chose perspectives that are not even one of the six correct labels. Column ‘ $\mu_{perspectives}$ ’ in Table 7.4 summarizes for all the six models the number of chosen unique perspectives per topic. For LDA and TF-IDF, users chose more than 9 unique perspectives for one topic.

Another analysis is how many of the six correct perspectives can be identified by a user per topic model. This is illustrated in Table 7.4 with the average number of correct perspectives $\mu_{correct}$ with standard error (= SE) and number of participants $N_{participants}$.

From the TF-IDF results in Table 7.4 and Figure 7.5 participants find three to four of the six correct perspectives using their own knowledge. In this case topic models LDA, LAM and JST performed almost similar compared to TF-IDF model. TAM and VODUM performed better than these three models where users found more correct perspectives as well as interpreted less unique perspectives from $\mu_{perspectives}$ in Table 7.4. To evaluate whether there is a significant difference between the topic models, we computed a one-way ANOVA.

model	$N_{participants}$	$\mu_{perspectives}$	$\mu_{correct}$	SE
tfidf	26	10.000	3.500	0.178
lda	22	9.667	3.591	0.193
tam	28	5.667	4.393	0.171
jst	28	8.333	3.607	0.171
vodum	25	7.667	4.120	0.181
lam	29	7.833	3.586	0.168

Table 7.4: Descriptive table of the topic models. $N_{participants}$ = number of participants, $\mu_{perspectives}$ = average number of perspectives chosen per topic, $\mu_{correct}$ = average number of total correct perspectives found per topic model, SE = standard error of $\mu_{correct}$.

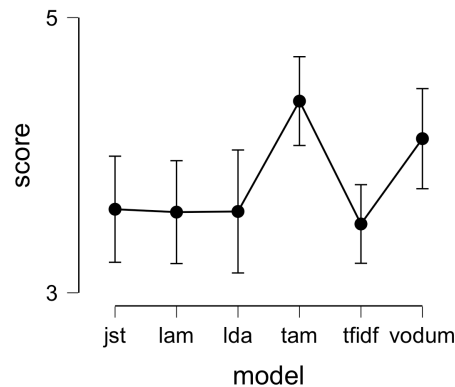


Figure 7.5: Average number of correct perspectives found by users, with standard error

Correctness of topic models using one-way ANOVA

The statistical analysis is the effect of topic models on the correct number of perspectives by conducting a one-way ANOVA. From the one-way ANOVA we find a significant difference between the models ($F = 4.399, df = 5, p < 0.001, \eta^2 = 0.126$). This suggests an impact of a topic model on the number of correct perspectives identified by users. We assessed the assumption of normality and heterogeneity of variances by using Shapiro-Wilk test and Levene's test respectively for the one-way ANOVA. Levene's test shows no significance ($F = 0.768, df = 5, p = 0.574$), the Shapiro-Wilk test does ($W = 0.905, p < 0.001$). Because the Shapiro-Wilk test gives a significant test result we conduct a non-parametric alternative to confirm the one-way ANOVA results. This is the Kruskal-Wallis test and confirms the results in the one-way ANOVA ($X^2 = 20.611, df = 5, p < 0.001$). We conduct a series of Mann-Whitney U tests as non-parametric post-hoc tests. From the post-hoc tests, TAM leads to significantly more correctly identified perspectives compared to TF-IDF and LAM. Otherwise the post-hoc tests show no significant differences. Table 7.5 shows the one-way ANOVA result with Table 7.6 the results of the post-hoc tests.

Cases	Sum of Squares	df	Mean Square	F	p
model	18.068	5.000	3.614	4.399	< .001
Residual	124.850	152.000	0.821		

Table 7.5: ANOVA score - The impact of topic models on the number of correct perspectives

Influence of personal stance on model judgement

The second experiment on the user study aims to answer **RQ4**. The assumption is that people would perceive more topics to be in line with their own stance. Recall that the given answers on stance fell on a 5-point likert scale with 1 being 'strongly disagree' and 5 'strongly agree'. We have more users strongly agreeing to abortion than strongly disagreeing as shown in Table 7.2. Table 7.7 shows a Spearman's correlation with its p-value. This correlation is between stance and number of chosen

	model	comparison	W	p
1	tfidf	lda	249.50	0.423
2	tfidf	jst	325.50	0.481
3	tfidf	vodum	190.50	0.007
4	tfidf	tam	162.00	< 0.001*
5	tfidf	lam	365.50	0.842
6	lda	jst	311.00	0.959
7	lda	vodum	198.50	0.082
8	lda	tam	176.00	0.006
9	lda	lam	340.50	0.674
10	jst	vodum	251.50	0.065
11	jst	tam	225.00	0.004
12	jst	lam	428.50	0.712
13	vodum	tam	292.50	0.278
14	vodum	lam	482.00	0.030
15	tam	lam	593.50	0.002*

* = < 0.003

Table 7.6: Post-hoc comparison - The impact of topic models on the number of correct perspectives

pro-perspectives.

	Spearman's ρ	p
stance - pro_n	-0.122	0.163

Table 7.7: Correlation table between stance and number of chosen pro-perspectives (pro_n)

We did not find a significant correlation between stance and pro-perspectives ($\rho = -0.122$, $p = 0.163$). Based on these results we cannot reject the null hypothesis that these two variables do not correlate. Our data thus suggests that users are not more likely to interpret the output of topic models in line with their own pre-existing stance.

Metadata: user feedback

The last step of the user study was a review of the topic models. Most participants believe a topic model is useful to automatically filter documents with 27.2% strongly agree and 44.9% somewhat agree on a 5-point likert scale. However, there are more users that are less confident than confident in assigning the correct perspective to a topic. 43.6% is not confident on a 5-point likert scale, 32.2% is confident and the remainder is neutral. Moreover, 18.4% strongly agree and 37.3% somewhat agree to being aware of more perspectives after they have conducted the user study. There does seem to be slight differences when data are sorted on the type of model as explained in Section 7.2.4. Table 7.8 shows the mean and standard deviation of questions Q6, Q7 and Q8. Q6 was: *A model that can automatically show all viewpoints is useful to quickly understand a debate.* Q7 was: *I'm now better aware of the possible viewpoints than before.* Q8 was: *I'm confident that I've correctly assigned the viewpoints to the word groups.* For all three questions the answers are on a 5-point likert scale which means that 1 is 'Strongly disagree' and 5 'Strongly agree'.

Row 'Usefulness' in Table 7.8 corresponds to Q6 where most users are either neutral or agree that a model for perspective discovery can be useful. Users seem to be more positive when evaluating VODUM or JST than for the other models. Row 'Perspective awareness' also shows that participants are either neutral or agree to being aware of more perspectives after the user study's main task. Users seem to be aware of less new perspectives when evaluating LDA than for example VODUM or JST. There is also a slight difference between models when participants were asked for their own confidence in evaluating the models, as found from Row 'Confidence' Table 7.8. Although this is not in line with the one-way ANOVA results, we find that users are more confident when evaluating TAM and JST than for the other four models.

Question		TFIDF	LDA	TAM	VODUM	JST	LAM
Q6 Usefulness	Mean	3.62	3.60	3.64	4.20	4.04	3.79
	Std	1.27	1.0	1.16	0.76	1.00	1.01
Q7 Perspective awareness	Mean	3.38	3.32	3.50	3.80	3.68	3.17
	Std	1.27	1.09	1.11	1.00	0.94	1.31
Q8 Confidence	Mean	2.46	2.68	3.18	2.72	3.25	2.62
	Std	1.24	1.17	1.12	0.94	1.08	1.12

Table 7.8: Mean and standard deviation of the answers given on Q6, Q7 and Q8 of the user study. The answers are a 5-point likert scale with 1 being strongly disagree and 5 strongly agree.

We can further analyze the difference between the participant’s own knowledge level on the abortion debate and Q7. In the first step of the study, 22.8% of the total participants answered ‘Strongly agree’ when asked if they know a lot about the topic abortion. In the third step of the study, 30.6% of this group answered ‘Strongly agree’ and 36.1% ‘Somewhat agree’ on the statement ‘*I’m now better aware of the possible viewpoints than before*’. This contradicts their answers in the first step of the study in which they stated to have high knowledge on the topic of abortion.

Lastly, from the open question’s user feedback, a bag-of-words representation has not always been understandable enough to identify a perspective and singular words may therefore not be meaningful enough. For example, a participant answered “*The similarities in the word groups was really difficult to select a viewpoint*”. Table B.1 in Appendix B shows all the feedback given on the user study. In total 27 out of 158 participants gave feedback on the user study. Most users who gave feedback showed a negative sentiment and did not evaluate TAM. This may give indication that users are more positive when evaluating TAM and compute more descriptive perspectives than the other models. Moreover, their feedback suggests that post-processing of the topic models could be a direction to improve the representation and understandability of the model’s output.

7.4. Summary

The user study focused on hypotheses H_1 and H_2 . We compared four joint topic models with two baseline models LDA and a random model based on TF-IDF. To compare these six models we have performed a descriptive analysis suggesting that users who have evaluated TAM chose less unique perspectives per topic and chose more often the six correct perspectives P . Moreover, the two perspectives in P that are most difficult to identify for most models are p_2 and p_6 . At least three of the correct perspectives - p_3, p_4, p_5 - have been chosen more often than any other random perspective in set V . To further investigate the differences between the topic models we also performed a statistical analysis through a one-way ANOVA. This analysis showed a significant difference between the models and led to the computation of post-hoc tests. To answer **RQ3** and hypothesis H_1 : TAM performed best where the results of the post-hoc tests show that TAM performed significantly different from TF-IDF and LAM.

We also analyzed what perspectives were chosen based on someone’s own stance on abortion, focusing on hypothesis H_2 . For this we calculated a Spearman correlation between the two variables stance and the number of chosen pro-perspectives. To answer **RQ4** and hypothesis H_2 : the p-value of the Pearson correlation is not significant to reject the null-hypothesis. This means that the correlation is not strong enough to state that users chose more perspectives in line with their own stance. This suggests that users do not show signs of the false consensus effect when evaluating a topic model.

We also gained insights in the usefulness of topic models where users see the potential of topic models, but depending on the type of model have variety in their own confidence level when evaluating the model. Moreover, they were aware of more perspectives after conducting the main task than before. However, as found in the user’s feedback: to increase the usefulness for end-users we would have to improve the topic model’s performance but also the representation of this output.

8

Discussion

8.1. Introduction

In this research we aimed to investigate the clustering ability of topic models and how well their computed topics describe understandable perspectives for users. We performed both an offline evaluation and online evaluation. The metrics TC and aRI with ground truth labels were used for the first type of evaluation and a user study for the latter. From these results we saw a relationship between the offline quantitative metrics and the online user study results. In both studies TAM performed best where the model has the highest aRI score and is more understandable for humans. Based on the user study results there is indication that a topic model is useful compared to a random model that is based on TF-IDF. However, this does not apply for all topic models. When comparing the different models with each other it becomes apparent that TAM is significantly better than LAM and TF-IDF.

8.2. Dataset

A topic model had to describe six perspectives by computing six topics. The final corpus to compute these topics is a balanced set of 600 documents where each of the six perspectives are represented by 100 documents. Moreover, the six labels are three pro- and three con-perspective labels. In Section 4.3 we analyzed the complete dataset. From these results we conclude that the dataset is highly homogeneous. Because the data are about one specific topic, almost all documents contain similar words with few nuances to distinguish them. For example, when computing the distribution of the frequency of unique tokens in the documents, we noticed a Zip's law where 50% of the unique tokens appear only once in the corpus. Moreover, when comparing the tokens split on stance there was little difference in word use: regardless of a person's stance, many people use the same words to explain their own perspective. As we also identified three overarching topics where each perspective has a counter-perspective, it challenged the question whether topic models could find the six perspectives that are present in the final corpus, the three identified overarching topics or none of the perspectives. To evaluate this we used a user study.

Given six topics computed by a topic model (excluding the honeypot topics), participants had to identify the six correct perspectives that are present in the corpus. We learned that for all topic models users could find at least three of the six perspectives and were chosen more often than an arbitrary perspective with probability 0.0625 in the list of perspective options V . By looking at the definitions of these three perspective labels, topic models are able to discover the three overarching perspectives but not all six. The main perspective that users cannot identify - but can when evaluating TAM - is p_2 and has been shown in Figure 7.3. Recall that p_2 is perspective *Personhood begins after a fetus becomes 'viable' (able to survive outside the womb) or after birth, not at conception*. When analyzing this perspective with the topic models' output almost none of the topics have tokens that exactly correspond to p_2 . To illustrate, if we would apply a simple baseline tokenization on each perspective (as shown in Figure 8.1) all perspectives except for p_2 have strong explicit keywords that distinguishes them from other perspectives. For example, p_1 has keywords such as 'choice', 'empower', 'control' and p_5 'kill', 'defy', 'God', whereas p_2 has words as 'after', 'outside', 'fetus'. The latter set of bag-of-words may be viewed as more ambiguous when represented to a user. We also see that documents with label

PRO		CON	
P1	<p>Reproductive choice empowers women by giving them control over their own bodies.</p> <p><i>reproductive, choice, empower, woman, give, control, body</i></p>	P4	<p>Abortion is murder because unborn babies are human beings with a right to life.</p> <p><i>murder, unborn, baby, human, being, right, life</i></p>
P2	<p>Personhood begins after a fetus becomes “viable” (able to survive outside the womb) or after birth, not at conception.</p> <p><i>personhood, begin, after, fetus, become, viable, survive, outside, womb, after, birth, not, conception</i></p>	P5	<p>Abortion is the killing of a human being, which defies the word of God.</p> <p><i>kill, human, being, defy, word, God</i></p>
P3	<p>A baby should not come into the world unwanted.</p> <p><i>baby, not, come, world, unwanted</i></p>	P6	<p>If women become pregnant, they should accept the responsibility that comes with producing a child.</p> <p><i>woman, become, pregnant, accept, responsibility, come, produce, child</i></p>

Figure 8.1: The six correct perspectives that are present in the final corpus, together with their baseline tokenization.

p_2 in the final corpus are explained in various implicit ways. Documents with label p_2 may be explained as, *A fetus has no thought or consciousness and does not feel pain. Up to a certain point all a fetus is is a clump of cells, or It’s no different to removing any other sort of unwanted cellular growth.* These documents do not explicitly address the perspective that a fetus is only a person after birth, but rather explain *why* this perspective holds. Because of this implicit language, documents with label p_2 match less with the exact perspective sentence p_2 . As a result, topic models may compute ambiguous topics to describe such perspectives, leaving users to give their own interpretation. This may suggest that 1) users match a perspective for a topic based on explicit keywords that they can identify and 2) topic models compute less ambiguous topics when authors use more explicit language. The first observation may also be in line with their feedback where participants answered *“Really found the word matching confusing”*, or *“very difficult to give categories for most of these”*. The latter observation may also correspond to the observed limitations of topic models from literature as explained in Section 2.4.5

8.3. Topic model performance

From the results in Chapter 6 and Chapter 7 a topic model can be used for perspective discovery, but not all topic models perform well enough for this task. The results of the post-hoc tests suggest that TAM performed significantly better than TF-IDF or LAM, and otherwise did not show significant differences. Although, a potential effect between TAM and the other models could also be found if the sample size was higher (e.g., the p -values between TAM-JST, TAM-LDA and VODUM-TF-IDF are quite low and close to 0.003, but not significant).

Section 8.2 suggests that users can better match perspectives with topics when explicit keywords in the sentence of the perspective label are also present in the computed topics. Comparing LDA with the other JTMs, LDA’s topics seem to be more ambiguous. Participants find it harder to distinguish the six topics and find it harder to identify the six correct perspectives. This was apparent in 1) the LDA’s output that computed six highly similar topics, 2) the higher number of uniquely chosen perspective labels per topic compared to the other topic models and 3) the user’s feedback. In that respect, JTMs compute more variation between the topics. If we compare the computed topics of the JTMs with each other (as seen in Figure 6.2 of Section 6.4) then all JTMs have at least one topic with tokens *‘unwanted’, ‘want’, ‘murder’, ‘god’, ‘kill’*. These tokens are also present in the perspective sentences p_3 , p_4 and p_5 respectively. As seen in Figure 7.2 and Figure 7.3 these three perspectives turn out to be the most chosen perspectives for all topic models. This further suggests that users match perspectives with topics based on the explicit keywords they can find. It also explains why users can find more correct perspectives for TAM than other topic models. For example, tokens as *‘fetus’, ‘womb’, ‘alive’* and *‘responsibility’* is hardly present in the other JTMs, but can be found in TAM. These tokens may correspond to p_2 and p_6 respectively which are also out of the six correct perspectives the two most difficult ones to identify. Given the homogeneous and implicit nature of the data, TAM may perform better than the other topic models because it computes more tokens that correspond to a perspective. The reason may be explained from its design.

Recall that for TAM the per-topic words distributions are based on whether words are background words covering the complete corpus (e.g. *‘abortion’, ‘mother’, ‘baby’*) or are perspective-dependent

(e.g. *'god'*, *'religion'*, *'responsibility'*). These are distributions that are not drawn by the other topic models. With our domain-specific dataset this is an important aspect as the dataset is one with documents containing many similar words about abortion as a whole and has only a few perspective-focused words (or as we have described them, the keywords). A model that makes such distinctions more profound creates better nuances between tokens and therefore lead to better descriptive topics. As all documents are opinions, a mere distinction based on sentiment or POS-tags as apparent in the other JTM's may therefore not be sufficient. Note that users did find almost the same number of correct perspectives for VODUM as for TAM. VODUM may therefore show a significant difference when increasing the sample size of the user study.

Another reason for TAM's performance (and potentially VODUM) is because it does not use prior knowledge, unlike JST and LAM. As mentioned in Chapter 2, the authors of VODUM did not want to use prior knowledge because they aimed for a flexible model that can be used in various domains. However, both JST and LAM use prior knowledge by using a subjectivity lexicon to compute per-topic negative words and per-topic positive words. As mentioned in Section 8.2 a stance separation on the dataset did not show more positive words when being for-abortion or more negative words when being against-abortion. This could make it difficult to use sentiment in order to distinguish perspectives between and within stances. Additionally, the perspectives that TAM can find better than the other models have topics with tokens *'responsibility'*, *'brain'*, *'alive'*, *'fetus'* and *'cell'*. Notice that these tokens do not hold a sentiment, so computing per-topic sentiment words distributions through a subjectivity lexicon could make JST and LAM less useful.

Moreover, we elevated TAM's and VODUM's performance by incorporating preprocessing techniques that can have less impact for JST and LAM. Recall that the most optimal setup is one where the final corpus has been tokenized by using antonyms, removal of non-sentiment words and n-grams. Because JST and LAM use a subjectivity lexicon it would not make optimal use of n-grams. As seen in Figure 6.2 JST mostly ignores the n-grams. It may not know how to incorporate the n-grams because it does not appear in the subjectivity lexicon. LAM does incorporate them which can partly be due to the POS-tags that LAM uses but is confused as to how it can produce matching n-gram tokens with each other, creating less coherent topics. Because TAM does not depend on prior knowledge or POS-tags it can make better use of the n-grams. For example, TAM computed topics with tokens not present in other topic models such as *fetus_not*, *child_not* and *life_begin*. Again, these are tokens that can be matched with p_2 and is one of the perspectives that are most difficult to identify for most topic models, creating the difference between TAM and the other topic models.

8.4. User experience

We explored the relationship between a user's own stance and on how a topic model is being perceived. We computed the Spearman correlation that showed no significant correlation between stance and the type of chosen perspectives. We can therefore not conclude that stance had an effect on the interpretation of the topic model output. This opens up the potential usefulness of a topic model as the model does not seem to be subject to a user's own assumptions with no sign of the false consensus effect.

This potential of topic models may also be derived from the metadata. There is a portion of participants strongly agreeing to have high knowledge about the topic abortion. However the majority of them also answered that they were aware of more perspectives after performing the user study's main task. People may believe to have a good understanding about the topic but in reality their knowledge may be confined in a limited space. If a topic model performs accordingly it may prevent such behaviour and also improve the reading experience for humans. Topic models can provide a structured overview where people read the main arguments of a discussion and it excludes the necessity of reading thousands of documents to understand a debate. It can broaden their own view and also reduce their cognitive load to prevent confirmation bias as explained in Chapter 1. Another possibility is to incorporate the topic model as functionality in existing tools to represent labels of documents and to automatically classify documents. This could be used as triggers to nudge people in reading diverse content.

Another observation is that the average confidence level of users slightly varies between the models. For example, users feel more confident in their answers when evaluating JST than LAM where users had an average confidence level of 3.25 and 2.62 on a 5-point likert scale respectively. This suggests

that the perceived correctness of a user is different from the objective correctness. Users are more confident when evaluating JST, but this model did not perform significantly best and users did not find most of the six correct perspectives for this model. Users identified for example more correct perspectives when evaluating VODUM than for JST with average scores of 4.120 and 3.607 respectively (= objective correctness). However, users have a lower confidence in their answers when evaluating VODUM than JST with scores of 2.72 and 3.25 respectively (= perceived correctness). This suggests a possible difference in objective correctness and perceived correctness.

8.5. Limitations

We have performed an extensive analysis on the topic models and have evaluated their performance on the basis of several factors. Although these results give an indication of how the models perform, it should be noted that these models have been evaluated with only one dataset. This research introduces a new type of evaluative framework in which different topic models are being compared. This means that we have focused on one particular controversial issue, because we want to establish a working evaluative pipeline before curating a new dataset. We now know the performance on this particular dataset, but we are not certain how generic the observations are for other datasets. Conducting the same experiments with multiple datasets aids in understanding whether the topic models would behave similar. For example, we can better conclude that TAM performs equally best given a different controversial issue. Moreover, the dataset on abortion is highly controversial where people have strong opinions on the topic. It would be interesting to know whether the topic models perform differently when the controversial issue is less polarized than abortion.

Another limitation is that the models have an evenly distributed data input with fixed perspective values and excludes any optimization through hyperparameter tuning. If we chose an uneven dataset and the topic model results would show a bad performance, then understanding the actual cause would be uncertain. It raises the question of whether a topic model performs poorly because of the topic model itself or because of the data being imbalanced. The objective of this research is to investigate what topic model performs best and incorporating extra factors into the experiments would not make clear whether the performance is truly because of the topic model's design. This also means that we did not perform any hyperparameter tuning as this also affects the topic model's performance which makes the comparison between them unequal. The caveat of these choices is that the results do not completely represent real-life scenarios. Documents are normally unevenly distributed without labels and is often optimized. To cater for this limitation we could further investigate the preprocessing process to better match a real-life scenario.

Another point is that the participants in the user study are not evenly distributed according to their stance. Although we added prescreening filters to aim for an even number of pro-life and pro-choice participants, we have more pro-choice participants than pro-life through platform *Prolific*. This uneven distribution may affect the user study experiment and we could aim for more users or more strict selection in future experiments.

Lastly, participants of the user study encountered difficulty in judging the computed topics because the top 10 tokens per topic are not always descriptive enough. Because we want to evaluate topic models through their immediate computed results, we did not perform any post-processing to better represent the topics. In the future, we could expand on this and focus on how topics can be better represented to make topic models more understandable and useful for end-users.

8.6. Summary

By analyzing the property of the dataset and the user study results, we derive that users mostly match a perspective to a topic based on the presence of explicit keywords. This made it more likely to match topics with tokens 'God', 'kill' to perspective p_5 *Abortion is the killing of a human being, which defies the word of God*. As seen, these tokens match with the exact words present in p_5 . Moreover, most users were able to identify p_5 for all topic models because each topic model have computed these tokens for at least one topic. Based on these observations users identified the perspectives p_2 or p_6 less because not all topic models computed topics with explicit tokens that correspond to these two perspectives. The question why certain topic models discover more correct perspectives than others lies therefore on which topic model can discover the explicit keywords best.

In the end, users chose less unique perspective labels for each topic and found more of the six

correct perspectives for topic model TAM. The performance of TAM can be explained by how the per-topic words distributions are designed and the exclusion of prior knowledge. Unlike the other joint topic models and LDA, TAM is built to separate the tokens based on whether it is a background-token that describes a corpus as a whole or a perspective-dependent token that describes a specific perspective. This led to per-topic token distributions that can find the nuanced explicit keywords better and is necessary to describe distinct perspectives due to the homogeneous and implicit nature of the corpus. Moreover, the joint topic models LAM and JST did not perform better than the two baseline models. Both use sentiment through a subjectivity lexicon to separate the tokens. This shows to be less sufficient for perspective discovery as the dataset is homogeneous but also the keywords that needs to be identified cannot always be categorized based on sentiment. This makes such models less useful for our dataset.

We also explored how users experienced the topic models. From there we conclude a potential usefulness of topic models where users are positive towards topic models for perspective discovery, but are not confident enough to use such a model. The models in this case do not perform well enough to distinguish all the perspectives, but given the response their performance is worthwhile to improve.

Lastly, we posed a set of limitations which makes current results not representative for real-life situations. This may be overcome in future work.

9

Conclusion

9.1. Conclusion

Given a controversial debate people may take a stance towards this debate. Underlying this stance is a person's perspective, which is the reason for taking on that stance. Highly controversial debates could have highly unstructured discussions with polarizing opinions, making it hard for a person to understand the discussion. In Chapter 1 we raised the question of whether unstructured opinionated documents on controversial debates could automatically be transformed into a structured overview of main perspectives. Such an overview can prevent the high cognitive load and may prevent the selective reading according to someone's ideological view. The process of automatically finding and extracting a structured overview of perspectives from unstructured text is defined as perspective discovery. To perform perspective discovery we explored the use of topic models and formed the basis of the main research question:

What topic model is able to discover human understandable perspectives on controversial issues?

We posed four sub research questions to answer the main research question. Four potential joint topic models for perspective discovery have been identified: TAM, JST, VODUM and LAM. Since all four are not compared with each other and use different evaluative frameworks, it is uncertain which model performs better over the other and why. We therefore created a new evaluative framework with the purpose of comparing various joint topic models with each other. All models used the same labelled dataset, evaluation metrics and are compared with their predecessor LDA as a baseline model. We also used a second baseline model to evaluate a topic model's performance against a user's own knowledge. This model is a random model based on TF-IDF. More specifically, we compared six models by their ability to find correct, human understandable perspectives. We introduced the aRI metric to quantify the clustering ability of topic models and performed a user study to investigate their level of understandability such that users understand the set of topics and match this with the correct set of perspectives.

To answer the main research question stated above: there exist topic models to discover perspectives from text but not all topic models perform equally well. When we compared the models on their clustering ability we noticed that TAM have a higher clustering score followed by VODUM, JST, LAM and LDA. In this case TAM is able to better assign the most-probable topic as perspective label to a document, that is similar to the corresponding ground truth label. Moreover, the user study suggests that TAM computes the most human understandable topics to describe perspectives.

In the user study we investigated the number of correct perspectives found by users per model. We noticed that all the participants found at least three of the six correct perspectives regardless of evaluating a topic model or a random model based on TF-IDF. If we compare the baseline model LDA with the other four joint topic models then LDA's topics seem to be more ambiguous. Participants find it harder to distinguish the six topics and find it harder to identify the six correct perspectives. In that respect, joint topic models compute more variation between the topics than LDA. Upon analyzing the differences between the topic models, we noticed that users match a topic with perspective by

looking at the explicit word matches that appear in both the topic and the perspective. In other words, users try to find keywords in topics that explicitly appear in the perspective in order to match them together. In that respect, users found more correct perspectives in TAM because it computes more of the keywords that do not appear in other topic models.

Several reasons were given for why TAM is the most suitable model for perspective discovery on the topic abortion. TAM is most suitable compared to the other four topic models because its design best addresses the homogeneous property of the dataset and does not use domain-dependent prior knowledge. Because the dataset is focused on one particular topic, only a few tokens are about one specific perspective. We created these nuances more profound by preprocessing the data, but this does not eliminate the overall limitation. TAM performed best because it is designed to separate words that appear in all documents (= background-words) from words that are only about specific perspectives (= perspective-specific words). This design principle is in line with how the homogeneous data looks like, leading to better descriptive topics than the other joint topic models. For example, some of the keywords that TAM computed and not the other topic models were *'responsibility'*, *'alive'*, *'brain'*, *'cell'*. These tokens are closely related to two perspectives that were harder to identify for the other topic models. These two perspectives are harder to describe, because only a few documents in the corpus have these words and the other four topic models do not focus on separating tokens based on such nuances. Another reason is that the keywords in the example do not hold any sentiment. Similarly, when we analysed the dataset there were no more positive and negative documents when splitting the documents based on pro-abortion and con-abortion respectively. Being dependent on prior knowledge such as a subjectivity lexicon could therefore hurt the performance as the needed keywords are not always sentiment-based. This makes the task of perspective discovery between and within stances for both LAM and JST more difficult.

Moreover, there is a potential link between the aRI metric and user study results. The order of the topic models from performing best to worse based on aRI is comparable to the user study. In both cases, TAM came out as best. This introduces the relationship between clustering ability and human understandability of topic models. It suggests that the better a topic model can assign a topic to a document that is in line with the ground truth label, the better it can describe a perspective that is understandable for a user to interpret the correct perspective.

As final words, in this thesis we showed that joint topic models have the potential to help users distill perspectives from large sets of documents. This opens up a potential research area on controversial debates in which perspectives between and within stances can automatically be discovered. Moreover, if we would carry out the research on the improvement of topic models, we can research the potential usefulness of topic models to reduce the cognitive load of people. A topic model could then be a tool to understand a main discussion quicker as opposed to overwhelming a reader with large sets of unstructured text.

9.2. Future work

Considering the obtained results the majority of users see potential in the usefulness of topic models. Moreover, all topic models were able to discover at least half of the correct perspectives, highlighting the potential of computational models for perspective discovery. To further improve upon current research there are a few research directions to explore for better usability of the models.

1. *Representation of topics*: Currently the topic models computes a set of word groups to describe a topic. This has shown to be difficult to understand for humans. Users have misunderstood topics and they are less user-friendly to represent to users. A possible direction is to explore various ways to represent a topic model output such that the models can be more understandable and useful for end-users. We may find better visual representations of the model's output or incorporate techniques to transform word groups into full sentences for better interpretation.
2. *Improving topic models*: Although each JTM shows a slight improvement compared to LDA there is no significant difference among most topic models. We still encounter topic models being purely based on token frequency which is highly focused on the statistical relationship of words. This however loses the semantic relationship between them. A possible direction is to combine topic models with word embeddings. Whereas topic models focus on the statistical properties of a corpus, word embeddings focus on the semantic relationship. Words that may not occur often

together could still hold a similar meaning which would be ignored by a topic model. For example, word embeddings could put documents together about 'death of a child', 'murder of a fetus' or 'killing an innocent'. All hold a semantic similarity, but a topic model would not group them into one as long as they do not appear statistically together. Combining both could strengthen the topic model's capabilities.

Another point is that topic models have mostly been used on explicit language use. Given our dataset it could be beneficial to explore the possibilities to also discover perspectives that takes implicit language into account.

3. *Introducing other datasets:* As mentioned in Chapter 8 we only worked on one dataset. To see if the same observations made in this research applies for other domains, we can run other datasets on the same pipeline. For this we would create labelled datasets focused on one controversial topic to keep the homogeneous property of the data. Moreover, the final corpus should be an evenly distributed set of documents with a balanced number of pro- and con perspective labels, so that we can evaluate whether the same topic models can distinguish perspectives between and within stances. Each perspective should also be represented by the same number of documents such that each perspective has an equal chance of being discovered. The topic model's performance can then be better explained through its design and not whether it is due to the imbalance of the data, as highlighted in Section 8.5.



Data and libraries

A.1. List of perspectives

Figure A.1 is a list of the 30 perspectives from *ProCon* to annotate the documents retrieved from *Debate.org*. Perspective 16 is a combination of two perspectives as explained in Section 4.2.3

Pro 1	The US Supreme Court has declared abortion to be a fundamental right guaranteed by the US Constitution.	Con 16	Abortion is murder because unborn babies are human beings with a right to life.
Pro 2	Reproductive choice empowers women by giving them control over their own bodies.	Con 17	Fetuses feel pain during the abortion procedure.
Pro 3	Personhood begins after a fetus becomes "viable" (able to survive outside the womb) or after birth, not at conception.	Con 18	Abortion is the killing of a human being, which defies the word of God.
Pro 4	Fetuses are incapable of feeling pain when most abortions are performed.	Con 19	The decision in <i>Roe v. Wade</i> was wrong and should be overturned.
Pro 5	Access to legal, professionally-performed abortions reduces maternal injury and death caused by unsafe, illegal abortions.	Con 20	Abortions cause psychological damage.
Pro 6	Modern abortion procedures are safe and do not cause lasting health issues such as cancer and infertility.	Con 21	Abortions reduce the number of adoptable babies.
Pro 7	Women who receive abortions are less likely to suffer mental health problems than women denied abortions.	Con 22	Selective abortion based on genetic abnormalities (eugenic termination) is overt discrimination.
Pro 8	Abortion gives pregnant women the option to choose not to bring fetuses with profound abnormalities to full term.	Con 23	Women should not be able to use abortion as a form of contraception.
Pro 9	Women who are denied abortions are more likely to become unemployed, to be on public welfare, to be below the poverty line, and to become victims of domestic violence.	Con 24	If women become pregnant, they should accept the responsibility that comes with producing a child.
Pro 10	Reproductive choice protects women from financial disadvantage.	Con 25	The original text of the Hippocratic Oath, traditionally taken by doctors when swearing to practice medicine ethically, forbids abortion.
Pro 11	A baby should not come into the world unwanted.	Con 26	Abortion promotes a culture in which human life is disposable.
Pro 12	Abortion reduces welfare costs to taxpayers.	Con 27	Allowing abortion conflicts with the unalienable right to life recognized by the Founding Fathers of the United States.
Pro 13	Abortion reduces crime.	Con 28	Abortion disproportionately affects African American babies.
Pro 14	Abortion is justified as a means of population control.	Con 29	Abortion eliminates the potential societal contributions of a future human being.
Pro 15	Many religious organizations and people of faith support women's reproductive choice.	Con 30	Abortion may lead to future medical problems for the mother.

Figure A.1: List of perspectives to annotate the documents, according to abortion.procon.org

A.2. Preprocessing of the data

The libraries used for this research are shown in Table A.1. Python 3.7 is used to run all the code. For the complete code see <https://osf.io/uns63/>

Library	Version	Purpose
Pandas	0.24.2	Data analysis
Numpy	1.16.2	Data analysis
NLTK	3.4	Preprocessing
Spacy	2.2.1	Preprocessing
Contractions	0.0.24	Preprocessing
Spellchecker	0.5.4	Preprocessing
Gensim	3.7.1	Model creation of LDA and computation of topic coherence
TAM	0.16	Model creation of TAM
JST	0	Model creation of JST
VODUM	0	Model creation of VODUM
LAM	0	Model creation of LAM
Matplotlib	3.0.3	Data visualization
WordCloud	1.6.0	Data visualization
Sklearn	0.20.3	Computation of aRI score
Statsmodels	0.9.0	Statistical analysis of the data

Table A.1: The libraries used for this research

B

User study

B.1. Results

Table B.1 shows the feedback on the user study given by the participants.

Model	Feedback
TFIDF	The survey went seamlessly. Thank you!
TFIDF	So confused!!
TFIDF	Somewhat confusing!
LDA	none
LDA	this study made absolutely no sense.
LDA	very difficult to give categories for most of these
LDA	That was tricky!
TAM	If this is used for good then this modal can be used to assist governments in making decisions based on a majority within the population. This could be good if used impartially.
TAM	I know one was definitely wrong but I couldn't work it out
TAM	some double negatives were difficult to understand
TAM	none, great survey:)
VODUM	It certainly gives one food for thought on the right to an abortion
VODUM	The similarities in the word groups was really difficult to select a viewpoint
VODUM	Really found the word matching confusing
VODUM	Bad English grammar, makes it hard to understand some things. Also, some of the option presented do not match any statement correctly, thus forcing the participant to selecte randomly.
VODUM	Some were very confusing to differentiate
JST	not all the options seem covered, i.e. the answers were prescriptive
JST	This Survey is somewhat confusing
JST	Hope I did well!
JST	A GOOD TOPIC AND AWESOME SURVEY.
JST	No thank you
JST	none
LAM	A foetus will have feelings and i believe it is wrong to kill a baby within a mother.
LAM	None
LAM	A little more context could make this simpler but pretty easy to understand for the most part
LAM	involves lots of deep thinking
LAM	Struggled to understand task

Table B.1: Open feedback comments from participants on the user study

B.2. Interface design

The images below are screenshots of the user study. Figure B.1 is the informed consent that participants need to read and agree to before they can proceed with the user study. The next page is the first step of the user study as illustrated in Figure B.2. These are metadata to gain insights in the user pool. The

second step of the study is the main task. The last page is a review of the topic models, illustrated in Figure [B.4](#).

Informed consent

The task that you are about to participate in is part of a research project at the Technical University of Delft in The Netherlands. The aim of the research project is to evaluate if a computational model can be used to summarize a controversial debate. For example, a controversial debate about abortion may have diversifying opinions for being either 'for' or 'against' abortion. The question is whether a model can automatically detect these opinions.

The task is part of the research project described above. In this task there is no right or wrong answer and it is solely your own interpretation. The task consists of three main parts. In the first part we ask your opinion on abortion to understand your familiarity with this topic. This is followed by an evaluation of the model and how interpretable the model is for you. The last part focuses on your experience while evaluating the models.

Your participation is entirely voluntary and you can withdraw at any time. No data aside from the above description and standard Prolific profile information are collected and stored. All your data are kept confidential and are stored in a password protected electronic format. The gathered data might be published to an online research repository, and this data should and will only be used for research purposes. Your data will be anonymized prior to any publication. You have the right to have your data removed within one month after your submission.

By selecting 'I agree' and clicking 'next' at the bottom of this page, you confirm that you have read, understood and consent to the above information.

I read, understood and consent the above information

I Agree


Next

Figure B.1: Informed consent

What is your age? (Drag left or right)

1 11 21 31 41 51 60 70 80 90 100

Age



What is your gender?

Male

Female

Prefer not to say

Your familiarity with the topic

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree	I don't know
In my opinion abortion should be legal	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have good knowledge about the abortion debate	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[Previous](#) [Next](#)

Figure B.2: First page of the user study after the informed consent

Below you see 16 different viewpoint statements regarding the claim: "abortion should be legal". Whereas half of the statements support abortion, the other half oppose to abortion. Underneath these viewpoint statements, you see 8 word groups. For example, a word group could look like this:

[decision, choose, abortion, legal, wrong, child, mother, right, law, choice]

Each of the 8 word groups belong to one of the 16 viewpoint statements. **Your task is to find the correct viewpoint statement for each word group. Do this by selecting one of the 16 viewpoint statements from the drop-down menu for each word group.** Whereas some word groups match a specific viewpoint statement very clearly, other matches might be harder to identify. Do not worry if a word group is too difficult to match, choose them to the best of your abilities.

Guidelines

- Per word group only one viewpoint is possible.
- A viewpoint cannot be selected for multiple word groups.
- You will not be judged on how many correct viewpoints you have selected. Choose them to the best of your abilities.

Pro 1	The US Supreme Court has declared abortion to be a fundamental right guaranteed by the US Constitution.	Con 9	Abortion is murder because unborn babies are human beings with a right to life.
Pro 2	Reproductive choice empowers women by giving them control over their own bodies.	Con 10	Fetuses feel pain during the abortion procedure.
Pro 3	Personhood begins after a fetus becomes "viable" (able to survive outside the womb) or after birth, not at conception.	Con 11	Abortion is the killing of a human being, which defies the word of God.
Pro 4	Fetuses are incapable of feeling pain when most abortions are performed.	Con 12	Abortions reduce the number of adoptable babies.
Pro 5	Women who receive abortions are less likely to suffer mental health problems than women denied abortions.	Con 13	Selective abortion based on genetic abnormalities (eugenic termination) is overt discrimination.
Pro 6	Abortion gives pregnant women the option to choose not to bring fetuses with profound abnormalities to full term.	Con 14	Women should not be able to use abortion as a form of contraception.
Pro 7	A baby should not come into the world unwanted.	Con 15	If women become pregnant, they should accept the responsibility that comes with producing a child.
Pro 8	Abortion is justified as a means of population control.	Con 16	Abortion eliminates the potential societal contributions of a future human being.

Use the dropdown to select the viewpoint for a word group

1: abortion, is, justified, as, a, means, of, population, control

2: woman, choice, body, fetus, control, pregnant, birth, baby, fetus, sex

Figure B.3: Snippet of the second page of the user study. The participant first needs to read the instructions.

Review of the model

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
A model that can automatically show all viewpoints is useful to quickly understand a debate	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I'm now better aware of the possible viewpoints than before	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I'm confident that I've correctly assigned the viewpoints to the word groups	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

If you have any feedback, please let us know

Thank you for participating in this survey! By clicking 'next' you submit your answers.

[Previous](#) [Next](#)

Figure B.4: Final part of the user study where the participant reviews the topic models.

Bibliography

- [1] Debate.org, *Do women have the right to abortion?* .
- [2] B. Pang and L. Lee, *Opinion mining and sentiment analysis*, *Foundations and Trends in Information Retrieval* **2**, 1 (2008).
- [3] S. Mohammad, P. Sobhani, and S. Kiritchenko, *Stance and sentiment in tweets*, *ACM Transactions on Internet Technology* **17** (2016), 10.1145/3003433.
- [4] T. Thonet, G. Cabanac, M. Boughanem, and K. Pinel-Sauvagnat, *Users are known by the company they keep: Topic models for viewpoint discovery in social networks*, (2017) pp. 87–96.
- [5] K. S. Hasan and V. Ng, *Stance classification of ideological debates: Data, models, features, and constraints*, in *Proceedings of the Sixth International Joint Conference on Natural Language Processing* (Asian Federation of Natural Language Processing, Nagoya, Japan, 2013) pp. 1348–1356.
- [6] Y. Lu, H. Wang, C. Zhai, and D. Roth, *Unsupervised discovery of opposing opinion networks from forum discussions*, (2012) pp. 1642–1646.
- [7] L. Akoglu, *Quantifying political polarity based on bipartite opinion networks*, *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014* , 2 (2014).
- [8] D. Vilares and Y. He, *Detecting perspectives in political debates*, in *EMNLP* (Association for Computational Linguistics, 2017) pp. 1573–1582.
- [9] D. Bountouridis, E. Sullivan, J. Harambam, N. Tintarev, C. Hauff, and M. Makhortykh, *Annotating credibility: Identifying and mitigating bias in credibility datasets*, (2019).
- [10] A. Konjengbam, S. Ghosh, and N. Kumar, *Debate stance classification using word embeddings: 20th international conference, dawak 2018, regensburg, germany, september 3–6, 2018, proceedings*, (2018) pp. 382–395.
- [11] M. Quraishi, P. Fafalios, and E. Herder, *Viewpoint discovery and understanding in social networks*, in *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18 (ACM, New York, NY, USA, 2018) pp. 47–56.
- [12] R. Nickerson, *Confirmation bias: A ubiquitous phenomenon in many guises*, *Review of General Psychology* **2**, 175 (1998).
- [13] M. Quraishi, P. Fafalios, and E. Herder, *Viewpoint discovery and understanding in social networks*, (2018) pp. 47–56.
- [14] D. Callahan, *The abortion debate: Can this chronic public illness be cured?* *Clinical obstetrics and gynecology* **35**, 783 (1992).
- [15] T. B. Edsall, *Why the fight over abortion is unrelenting*, (2019).
- [16] R. Kennedy, *In-class debates: Fertile ground for active learning and the cultivation of critical thinking and oral communication skills*, *Int J Teach Learn Higher Educ* **19** (2006).
- [17] S. Sarawagi *et al.*, *Information extraction*, *Foundations and Trends® in Databases* **1**, 261 (2008).
- [18] L. Zhang and B. Liu, *Aspect and entity extraction for opinion mining*, (2014) pp. 1–40.
- [19] E. Riloff, S. Patwardhan, and J. Wiebe, *Feature subsumption for opinion analysis*, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Sydney, Australia, 2006) pp. 440–448.

- [20] Y. Choi, Y. Jung, and S.-H. Myaeng, *Identifying controversial issues and their sub-topics in news articles*, (2010) pp. 140–153.
- [21] E. Barker and R. Gaizauskas, *Summarizing multi-party argumentative conversations in reader comment on news*, in *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)* (Association for Computational Linguistics, Berlin, Germany, 2016) pp. 12–20.
- [22] S. Somasundaran and J. Wiebe, *Recognizing stances in ideological on-line debates*, *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 116 (2010).
- [23] B. Liu and L. Zhang, *A survey of opinion mining and sentiment analysis*, in *Mining text data* (Springer, 2012) pp. 415–463.
- [24] X. Ding, B. Liu, and P. S. Yu, *A holistic lexicon-based approach to opinion mining*, in *Proceedings of the 2008 international conference on web search and data mining* (ACM, 2008) pp. 231–240.
- [25] A. Alaei, S. Becken, and B. Stantic, *Sentiment analysis in tourism: Capitalizing on big data*, *Journal of Travel Research* **58**, 175 (2019).
- [26] S. S. Tan and J.-C. Na, *Mining semantic patterns for sentiment analysis of product reviews*, (2017) pp. 382–393.
- [27] N. Godbole, M. Srinivasaiah, and S. Skiena, *Large-scale sentiment analysis for news and blogs*, (2007).
- [28] B. Liu, *Sentiment analysis and opinion mining*, *Synthesis lectures on human language technologies* **5**, 1 (2012).
- [29] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, *Lexicon-based methods for sentiment analysis*, *Computational Linguistics* **37**, 267 (2011).
- [30] T. Thet, J.-C. Na, and C. Khoo, *Aspect-based sentiment analysis of movie reviews on discussion boards*, *J. Information Science* **36**, 823 (2010).
- [31] S. Baccianella, A. Esuli, and F. Sebastiani, *Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining*. in *LREC*, Vol. 10 (2010).
- [32] N. M. Shelke, S. Deshpande, and V. Thakre, *Survey of techniques for opinion mining*, *International Journal of Computer Applications* **57**, 0975 (2012).
- [33] M. Walker, P. Anand, R. Abbott, J. Fox Tree, C. Martell, and J. King, *That is your evidence?: Classifying stance in online political debate*, *Decision Support Systems* **53**, 719–729 (2012).
- [34] H.-T. Zheng, B.-Y. Kang, and H.-G. Kim, *Exploiting noun phrases and semantic relationships for text document clustering*, *Information Sciences* **179**, 2249 (2009).
- [35] K. Khan and W. Khan, *Sentence level domain independent opinion and targets identification in unstructured reviews*, *Computers* **7**, 70 (2018).
- [36] D. Blei, A. Ng, and M. Jordan, *Latent dirichlet allocation*, *Journal of Machine Learning Research* **3**, 993 (2003).
- [37] S. Bunk and R. Krestel, *Welda: Enhancing topic models by incorporating local word context*, (2018) pp. 293–302.
- [38] M. Shams and A. Baraani-Dastjerdi, *Enriched lda (elda): Combination of latent dirichlet allocation with word co-occurrence analysis for aspect extraction*, *Expert Systems with Applications* **80** (2017), 10.1016/j.eswa.2017.02.038.
- [39] M. Paul and R. Girju, *A two-dimensional topic-aspect model for discovering multi-faceted topics*. in *AAAI*, Vol. 1 (2010).

- [40] C. Lin and Y. He, *Joint sentiment/topic model for sentiment analysis*, in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09 (Association for Computing Machinery, New York, NY, USA, 2009) p. 375–384.
- [41] T. Thonet, G. Cabanac, M. Boughanem, and K. Pinel-Sauvagnat, *Vodum: A topic model unifying viewpoint, topic and opinion discovery*, in *ECIR*, Vol. 9626 (Springer, Toulouse, France, 2016) pp. 533–545.
- [42] V. R. Sathi and J. S. Ramanujapura, *A quality criteria based evaluation of topic models*, (2016).
- [43] M. Röder, A. Both, and A. Hinneburg, *Exploring the space of topic coherence measures*, in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15 (ACM, New York, NY, USA, 2015) pp. 399–408.
- [44] J. H. Lau, T. Baldwin, and D. Newman, *On collocations and topic models*, *ACM Trans. Speech Lang. Process.* **10** (2013), 10.1145/2483969.2483972.
- [45] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei, *Reading tea leaves: How humans interpret topic models*, (2009) pp. 288–296.
- [46] D. Newman, Y. Noh, E. Talley, S. Karimi, and T. Baldwin, *Evaluating topic models for digital libraries*, in *Proceedings of the 10th annual joint conference on Digital libraries* (2010) pp. 215–224.
- [47] L. Ross, D. Greene, and P. House, *The "false consensus effect": An egocentric bias in social perception and attribution processes*, *Journal of experimental social psychology* **13**, 279 (1977).
- [48] k. krippendorff, *Agreement and information in the reliability of coding*, *Communication Methods and Measures* **5**, 93 (2011).
- [49] N. Mouter and D. Vonk Noordegraaf, *Intercoder reliability for qualitative research: You win some, but do you lose some as well?* in *Proceedings of the 12th TRAIL congress, 30-31 October 2012, Rotterdam, Nederland* (TRAIL Research School, 2012).
- [50] A. Zapf, S. Castell, L. Morawietz, and A. Karch, *Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate?* *BMC Medical Research Methodology* **16** (2016), 10.1186/s12874-016-0200-9.
- [51] T. L. Griffiths and M. Steyvers, *Finding scientific topics*, *Proceedings of the National academy of Sciences* **101**, 5228 (2004).
- [52] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu, *Short text topic modeling techniques, applications, and performance: a survey*, *IEEE Transactions on Knowledge and Data Engineering* (2020).
- [53] W. M. Rand, *Objective criteria for the evaluation of clustering methods*, *Journal of the American Statistical association* **66**, 846 (1971).
- [54] K. Y. Yeung and W. L. Ruzzo, *Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data*, *Bioinformatics* **17**, 763 (2001).
- [55] L. Hubert and P. Arabie, *Comparing partitions*, *Journal of Classification* **2**, 193 (1985).
- [56] M. L. Mortensen, G. P. Adam, T. A. Trikalinos, T. Kraska, and B. C. Wallace, *An exploration of crowdsourcing citation screening for systematic reviews*, *Research synthesis methods* **8**, 366 (2017).
- [57] J. Bruin, *newtest: command to compute new test @ONLINE*, (2011).
- [58] JASP Team, *JASP (Version 0.11.1.0)[Computer software]*, (2020).
- [59] J. Freeman, D. Anderson, D. Sweeney, T. Williams, and E. Shoemith, *Statistics For Business and Economics.*, 4th ed. (Cengage, United States, 2017).
- [60] *Spearman rank correlation coefficient*, in *The Concise Encyclopedia of Statistics* (Springer New York, New York, NY, 2008) pp. 502–505.