Unravelling Twitter chaos during a policy crisis

APPLYING SENTIMENT ANALYSIS AND TOPIC MODELLING TO TWEETS ABOUT THE DUTCH NITROGEN CRISIS

Thesis submitted in fulfilment of the requirements for the degree of

MASTER OF SCIENCE

In Industrial Ecology
At Leiden University, Institute of Environmental Sciences (CML)

and

Delft University of Technology, Faculty of Technology, Policy and Management

by

Mila Hendrikse
LU: s2348330
TUD: 4306146

To be defended in public on 26th of November 2021

FIRST SUPERVISOR	DR. PRADEEP K. MURUKANNAIAH
SECOND SUPERVISOR	DR. STEFANO CUCURACHI
THIRD SUPERVISOR	MSC MICHIEL VAN DER MEER

Layout design by Ronny Lassooij

Acknowledgements

I would like to thank my supervisors, Pradeep K. Murukannaiah, Stefano Cucurachi and Michiel van der Meer, for their supervision, availability and access to the TU Delft Proshare server. Thank you, Desiree, for being my first thesis buddy and going on this journey with me with your compassion, your warmth and support. Thank you, Kim, for being my second thesis buddy, for proof reading my report, for your complete open-mindedness and for somehow having an answer to all my questions and a solution to all my problems. You have made these last few months even more educative and enjoyable. Thank you, Jenny, for your input, kindness and your supervision a year before the actual start of my thesis. A thanks to the 'thesis circle': Lotte, Francien, Ronny, Elvira, Paula and Xander, for keeping me company during the last months of our theses. Thank you, Sahiti, for being the extremely skilful academic that you are and saving the storyline of this thesis. Thank you, Francesca, for your support and help, and charmingly putting up with me during my most stressed and grumpy days.

Abstract

In May 2019, the Dutch Council of State rejected the national approach for reducing nitrogen emissions in Dutch nature. Farmers were targeted by the policy change: all licenses for agricultural expansion were revoked, affecting the financial livelihoods of farmers. Farmers did not take this well and, using social media platforms, started organising large-scale demonstrations. Twitter flushed with posts about the demonstrations, and pictures and videos of the event went viral. Demonstrations result from social unrest, and social unrest starts with public dissatisfaction. The Nitrogen Crisis showed it is in the interest of decision-makers to monitor what negative feelings the public holds towards policies and act on these, before these feelings grow into social unrest. Twitter is a social media platform many users come to for expressing their opinions. Because of this, this study looks at what insights can be derived from Twitter about events, like demonstrations, that took place during the Nitrogen Crisis. For this, this study applies two Natural Language Processing methods: sentiment analysis and topic modelling (LDA). These methods are combined in order to create more insightful and interpretable results than the methods individually could provide. Two interviews are held with an expert on the Nitrogen Crisis to provide context on the crisis and to identify major events that received a lot of media attention. The events are plotted with- and compared to the results of sentiment analysis and topic modelling. In doing so, the following research question is answered:

How can sentiment analysis and topic modelling be applied to Twitter data to provide insights for decision-makers retrospectively about major events during the Dutch Nitrogen Crisis?

For sentiment analysis, two Dutch sentiment analysis tools are implemented and compared to the sentiment scores of 100 tweets by three annotators to select the best performing one. For topic modelling, a grid search is performed to choose the combination of timeframe and number of topics that result in the set of topic models that have the highest mean topic coherence. Also, a method is proposed for using topic models to represent changes in the topics discussed on Twitter over time. This is used not only to compare subsequent topic models per sentiment that are one week apart, but also topic models that are 4 weeks apart.

This research develops a fully functioning pipeline for collecting and processing tweets, applying sentiment analysis and topic modelling and plotting the outcomes. This pipeline has been validated at various points, leading to a scientifically viable methodology.

Unexpectedly, it is not sentiment analysis or topic modelling results that have the most obvious connection with the events identified: it is an increase in tweets during events. Therefore, while lacking better tools, decision-makers are recommended to monitor pre-determined topics on Twitter and implement a way to be notified when a significant change in volume of tweets takes place. The combination of sentiment analysis and topic modelling as implemented in this research is either not advanced enough to provide useful information to decision-makers, or sentiment analysis and topic modelling simply cannot provide

insightful results on the Dutch Nitrogen Crisis. However, because there is an extensive amount of research applying these methods to social media data around various political events with valuable results, it is recommended to perform more experiments with this approach, and the quality of each research step needs to be further improved in order to draw final conclusions on the usefulness of the combination of sentiment analysis and topic modelling for decision-makers during policy crises. Various improvements for each step in this research are suggested to gain more precise, insightful and interpretable results. In summary, this study and the pipeline it proposes can serve as a solid basis for further development into a process that provides ready-to-use information to decision-makers.

Table of Contents

List of Figures

List of Tables

List of Acronyms

CJEU	Court of Justice of the European Union
FDF	Farmers' Defence Force
PAS	Programma Aanpak Stikstof (Eng.: Integrated Approach to Nitrogen)
PCC	Pearson Correlation Coefficient
SS	SentiStrength
STD	Standard deviation
TU DELFT	Technical University of Delft

Positioning

This brief chapter describes how my background as a researcher influence the topics discussed, the methods applied, the narrative of this study and how biases could be present in it. Research is often claimed to be neutral and objective. However, studies on the decolonisation of knowledge institutions have shown that there is no such thing, and that research has perpetuated inaccurate stereotypes and discrimination (Thambinathan & Kinsella, 2021). As there is no such thing as a neutral or unbiased researchers, I believe positioning myself as researcher is an important step to take at the start of this report. Also, I will reflect on some lessons I learned at university, how they influence the environment I find myself in. I relate these lessons to events during the Nitrogen Crisis that caught my attention during this research and ask several questions about their implications.

My name is Mila Hendrikse. I am a Dutch student of 26 years old of Dutch and Italian descent. I grew up in a prosperous town called Heemstede and enjoyed education of the highest level before coming to TU Delft to study Computer Science. It seemed like a very interesting and difficult area of study and I wanted to challenge myself. During my bachelor, I moved to Lisbon, Portugal for half a year and studied Life Sciences. After finishing Computer Science, I briefly studied Logic at the University of Amsterdam, before returning to Delft to study MSc Industrial Ecology (IE) at the TU Delft and Leiden University and MSc Engineering and Policy Analysis (EPA) at the TU Delft. Although all different, all these degrees at different universities taught me to think from an academic perspective.

At Industrial Ecology, we look at climate change and complex environmental problems from technical, environmental and social perspectives on a high level. At EPA, the focus lies on tackling large scale complex problems such as, but not limited to, climate change. This is done through the modelling of such complex problems, analysing the problems with these models and using the outcomes for supporting conclusions and advice. The political landscape matters in this academic field, and several courses are taught on how to identify key stakeholders and navigate the political landscape while achieving pre-set goals. At both masters, 'the art of political framing' is taught by professor of governance Hans de Bruijn (De Bruijn, 2019). The art of framing is a technique for reframing a debate in your favour by redefining, in a situation, who the 'hero', the 'victim' and the 'villain' are. I remember feeling surprised during these lectures about the suggestion that we should at all times play 'the political game': it did not sound just that the person that is most skilled at (re)framing a debate is the person who is most likely to be successful and look wise and informed, regardless of their work and contributions. When vocalising my doubts in class, the teacher had a response ready: you have to learn how to play this political game, because everyone does. If you do not learn how to play, you will lose, regardless of your work or intentions. He was fast to respond, because I was obviously not the first student to question this approach. However, although I wish it were different, I have come to understand what he meant. He was right, at least for the environment I find myself in as a governance student at TU Delft and future policy maker in the Netherlands: to be successful at my work, it seems I need to learn how to play this political game.

The choice of the topic, methods and narrative of this study is highly influenced by my background. I chose the topic, the Nitrogen Crisis, because it links sustainability with policy, the two topics of the Masters I take. I chose the methods, NLP methods, because I already gained experience with applying NLP during my bachelor. And I chose the target audience, decision-makers, because policy-making is a skill that is taught at EPA.

In this research I talk about the Dutch Nitrogen Crisis. This was a crisis that started when the Dutch government hastily implemented measures that threatened the financial livelihood of Dutch farmers after the Court of Justice of the European Union rejected the Dutch policy proposal for managing Nitrogen emissions. Dutch farmers considered this unfair, as they were not the cause for the bad policy proposal, and because the Dutch government had received several warnings about the issue in the years preceding the crisis. The crisis could have been avoided, and it is the government that could have done this, not the farmers (Schreuder, 2019). While learning more about this crisis and the dissatisfied responses of the Dutch farmers, I came across hints that it is exactly the political game that frustrated the farmers. Their reactions seem to indicate frustration with the government and the organisations that represent the interests of the agriculture sector (Visser, 2020). The political game itself is something that excluded the farmers in the run-up to the Nitrogen Crisis and whose outcomes hurt them. After all, to play the political game, you need to be invited to the table and know the rules or be able to influence them. Farmers' Defense Force, the most radical new farmers' organisation, has often been criticised for showing its dissatisfaction in words and actions that are not deemed appropriate (Omroep West, 2020; Driessen, 2019; Kalkhoven, 2021; Wijnants, 2020; Winterman, 2019). Could one play the political game accordingly, though, when not being in charge of, not knowing or not agreeing with its rules? Who decides what is deemed 'appropriate'? And is this process fair? These are questions that arose while learning about the Nitrogen Crisis.

However, I have not nearly read and watched enough media from the perspective of the farmers to understand how they feel and why. Also, I am not a farmer myself nor did I grew up around them. As this is academic research, apart from news articles, I mostly cite and rely on academic sources. To create a more complete picture of the Nitrogen Crisis, I would have liked to read more from the perspective of the farmers and interviewed farmers as well. This research aims to help decision-makers, but decision-makers make decisions that influence the farmers' livelihood. The opinions and experiences of farmers should be represented in the decision-making process for the future of agriculture in the Netherlands.

2 Introduction

2.3 The Dutch Nitrogen Crisis and farmers protests

In May 2019 the Dutch Council of State[1] rejected the national Nitrogen approach to reducing nitrogen emissions in Dutch nature (Programma Aanpak Stikstof or PAS in Dutch), following a rejection from the Court of Justice of the European Union (CJEU) of the PAS in 2018 (Natuurmonumenten, 2018; NOS, 2019). The PAS permitted licenses to expanding businesses, even when the expansions would lead to more nitrogen emitted, as long as there were plans for the future compensation for these nitrogen emissions. Although various preceding cabinets were warned about the flaws of this policy, it turned out the Dutch government was completely unprepared for the possibility of its rejection by the EU. The initial response was to cancel all licenses for emission rights, which blocked around 18.000 building and infrastructure projects. The most vocal and visual stakeholders who spoke out about the harm by the sudden policy change, though, were the Dutch farmers (RTL Nieuws, 2019).

Farmers were greatly targeted by the policy change: all licenses for agricultural expansion were revoked. In search for solutions, politicians started talking about halving the total livestock in the Netherlands and large-scale buyout of farmers. Farmers did not take this well and started organising demonstrations (Leeuwarder Courant, 2019). These events were largely organised on social media platforms Facebook and WhatsApp (Kalkhoven, 2021). The biggest demonstrations took place in the city where the Dutch House of Representatives is located, the Hague, but several demonstrations were held in front of various Province Houses in the Netherlands as well. Dozens of tractors blocked the busiest highways on their way to the Hague

[1] The highest administrative court in the Netherlands

to occupy Malieveld[2]. Social media platforms flushed with posts about the demonstrations, and pictures and videos of the event went viral.

Politicians started reacting to the demonstrations. Some politicians were given permission by the protesting farmers to address the crowd onstage during demonstrations and Prime Minister Rutte and minister Schouten of Agriculture, Nature and Food Quality invited leaders of the farmers' organisations to meet and discuss the Nitrogen Policy (Schelfaut, 2019b; Dagblad, 2021). Although the farmers' demonstrations had some success and received much public and political attention, dissatisfaction about each new proposal and how the Cabinet handles the Nitrogen Crisis kept resurfacing. Farmer protests took place in October and December 2019, February and June 2020, and July 2021. The demonstrations got increasingly aggressive: heavy agriculture machinery was used to block roads, break through fences and some politicians started receiving threats (Wijnants, 2020; Kos, 2020; Winterman, 2019). The most radical farmer organisation, Farmers Defence Force (FDF), expressed the belief they need to organise increasingly radical protests to keep pressuring the government to listen to their requests (Kalkhoven, 2021).

When citizens feel the decisions of policy makers do not represent the need of the people, this can cause tremendous uproar. As seen during the Nitrogen Crisis, the actions of citizens today, for example demonstrations, can set the political agenda and even tomorrow's policy decisions (Klumpenaar & Van Laarhoven, 2019). And as we have also seen during the Nitrogen Crisis, today's demonstrations can be organised in a very short time on social media. As this combination of a fast citizen response and the large impact of this response can disrupt policy-making, it is in the interest of decision-makers to monitor what negative feelings the public holds towards policies and act on these, before these feelings grow into social unrest (Kalkhoven, 2021).

2.4 Twitter as a platform for public opinion

Twitter can serve as an excellent platform for monitoring social dissatisfaction and unrest. Twitter is one of the most wide-used social media platforms worldwide. While Facebook traditionally focuses on connecting people and sharing updates on users' lives, Twitter is mostly used as a platform for sharing opinions. Users do this through short messages, called Tweets, which have a limit of 280 characters each and can include media, like photos, videos and URLS as well.

There are advantages to analysing tweets to monitor public sentiment. First, in contrast to traditional public opinion polling methods like surveys and interviews, posting on and interacting with Twitter takes its users relatively little time. Since all messages are shorter than 280 characters, a Tweet with a thought, opinion or fact is written and send out into the world

[2] A grass field in the centre of The Hague known for being the location for festivals and big demonstrations.

easily and in no time. Second, extracting and analysing Twitter data requires no engagement with Twitter users. Third, Twitter is widely used as a platform for discussing and arguing about important topics on a large scale. These topics can include controversial policies or responses to global crises, like the Covid-19 pandemic. Last, because of the Covid-19 crisis, and its restrictions on social life and daily interactions, Twitter has recently seen a noteworthy increase in use (Miao, Last, & Litvak, 2020). The Covid-19 crisis and the accompanying increase in use of Twitter starts a few months after the start of the ongoing Nitrogen Crisis. Last but not least, huge volumes of data are posted on Twitter on a daily, even hourly basis. Where polling methods take much effort and time to provide limited information, Tweets can be easily downloaded and used for large scale data analysis.

2.5 Analysing social media discourse

The number of tweets that are posted to Twitter each day is too high to analyse manually. With the rise of the internet came the development of computational text analysis techniques, often referred to as Natural Language Processing (NLP) (Jurafsky & Martin, 2020). The two most common NLP tasks are sentiment analysis and topic modelling.

Sentiment analysis is the task of automatically calculating sentiment in a piece of text, and one of the most rapidly growing research areas (Mäntylä, Graziotin, & Kuutila, 2018). Sentiment analysis relies on pre-annotated lexicons with frequently used words and their sentiment score. This lexicon is then used to calculate the sentiment of a piece of text. The earliest sentiment analysis methods were bag-of-words approaches, relying purely on these lexicons and therefore the 'positivity' and 'negativity' of words in a sentence to score the sentiment. However, more recent methods take the order of words and the effect of intensifying words (e.g., "very" and "super") into account, leading to a higher score of the sentiment of words that follow intensifying words. Furthermore, it takes into account negative words (e.g., "not" and "nothing"), which gives an opposite sentiment score to the sentiment of the word that follows the negative words. Therefore, these more recent methods are better able to handle the nuances of human language (Trilling & Boumans, 2018).

Topic modelling is the procedure for automatically determining a set of topics from a dataset consisting of text documents. The most popular algorithm for this is Latent Dirichlet Allocation (LDA), a machine learning algorithm (Blei, Ng, & Jordan, Latent Dirichlet allocation, 2003). For the basic form LDA, the amount of topics present in the dataset must be defined at the beginning, say k. Then, each word in each document is assigned to one of the topics randomly, and iteratively the probability of each word in every document belonging to each topic is updated by multiplying the proportion of words in the document assigned to this topic and the proportion of documents assigned to this topic. After the algorithm converges, we are left with k topics, which are each represented by a list of words. These words are weighted: the highest weight is more likely to represent the topic than words with lower weights. However, one

important step that basic LDA does not perform is the labelling of each topic. When done by hand, this is a very time-consuming task, and sometimes the collections of words LDA produces for a topic are not coherent enough to be interpretable (Allahyari, Pouriyeh, Kochut, & Arabnia, 2017; Röder, Both, & Hinneburg, 2015). This can limit the practical use of LDA for processing text.

Topic modelling and sentiment analysis alone are not suitable for the purpose of analysing changes in public opinion on Twitter. Sentiment analysis can provide information on how people feel, but not on *what* they are discussing. Topic modelling can provide information on what people are discussing on social media, but it is hard to decide when to calculate a topic model, how many days its input data should span and when the decision-makers should take the time to look at and interpret topic models. Even if decision-makers would like to monitor online discourse through topic modelling, it would be foolish to spend an endless amount of time looking at topic models, not knowing exactly what to look for. Therefore, this study combines sentiment analysis and topic modelling, aiming to provide a method that points decision-makers to interesting developments that could signal social unrest on Twitter. In doing so, the following main research question will be answered:

MQ

How can sentiment analysis and topic modelling applied to Twitter data to provide insights for decision-makers retrospectively about major events during the Dutch Nitrogen Crisis?

The main research question encompasses various components that will now be unpacked. As introduced, the case for this study will be the Dutch Nitrogen Crisis, as it is a recent sustainability policy crisis that received a lot of media attention on both traditional news media and Twitter. The data that will be analysed for this study are Tweets during Nitrogen Crisis, because people widely shared their opinions on the Nitrogen Crisis on Twitter. The Nitrogen Crisis started in 2019, and although it is still ongoing at the time of writing, there is a lot of Twitter data available from 2019 to 2021 covering many events, such as demonstrations and policy decisions. The methods applied to this data are sentiment analysis and topic modelling. To gain a better understanding of the political dynamics of the Nitrogen Crisis and what kind of insights decision-makers would find useful, two interviews are held with expert Han de Groot, who interviewed many stakeholders of the Nitrogen Crisis. The main question that is asked is *how* these methods can provide insights for decision-makers: a processing pipeline is set up to combine sentiment analysis and topic modelling, aiming to plot sentiment and changes in topics discussed over time. These results are then compared to major events that received much attention from news media, to see if a connection can be found between the outcomes of the application of the methods and the occurrence of such events. In this way, retrospective insights will be gained, that could form the basis for future work on flagging social unrest over sustainability policies using Twitter data.

2.6 Sub research questions

To answer the main research question, the following sub questions will be answered:

SQs	
SQ 1	*How did sentiment on Twitter develop over time during the Nitrogen Crisis?*
SQ 2	*How can topic modelling be applied to analyse changes in topics discussed on Twitter over time?*
SQ 3	*How did the topics discussed during the Nitrogen Crisis on Twitter change over time?*
SQ 4	*Can sentiment analysis and topic modelling results provide markers for major events during the Nitrogen Crisis?*

To analyse how sentiment developed over time (SQ 1) two sentiment analysis tools are applied to the tweets and evaluated by comparing the sentiment scores of the tools to the sentiment scores of three annotators. After the best sentiment analysis tool is picked, the average sentiment of tweets and the volume of tweets per sentiment throughout the Nitrogen Crisis is plotted. Also, two subsets of the datasets are created: one containing all the positive tweets, and one containing the negative tweets. Answering SQ 2 involves developing a method for using topic models to analyse how topics discussed change over time. This method is then applied to plot the differences in topics discussed on Twitter over time, for the full tweet dataset, the dataset consisting of positive tweets and the dataset consisting of negative tweets, answering SQ 3. In answering SQ 4, the outcomes of sentiment analysis and topic modelling are combined and analysed around the major events, to explore whether the major events could have been identified via Twitter data and if there were events on Twitter that were not identified during desk research and the interviews. Together, the answers to these sub questions provide the answer to the main research question, with the overall goal to find NLP processing methods applicable to Twitter that are insightful to decision-makers in a situation similar to the Dutch Nitrogen Crisis.

2.7 Structure of report

In chapter 3*: Related work*, key concepts and methods are introduced and related research area are discussed. Chapter 4*: Methods* describes the research approach of this study, how data is

collected and analysed and how the steps in this research are evaluated. In chapter 5*: Results* the main findings of this study are presented and reflected on. In chapter 6*: Discussion, limitations and future research* the findings are interpreted and discussed in a larger context, after which the limitations are elaborated and recommendations for future work are presented. Chapter 7*: Conclusion* the research is summarised and the research questions are answered. The report ends with chapter 8: Appendix and, finally, the list of references.

Key Findings of Chapter 1: Introduction

1. Subject: Dutch Nitrogen Crisis
2. Methods: Sentiment analysis, topic modelling and two interviews
3. Aim: Generate Twitter insights for decision-makers
4. Main research question:

   How can sentiment analysis and topic modelling applied to Twitter data to provide insights for decision-makers retrospectively about major events during the Dutch Nitrogen Crisis?

3 Related work

This chapter first summarizes current knowledge on the interplay between social media platforms and politics. Then, as the contribution of this research comprises the combination of sentiment analysis and topic modelling, both methods are introduced and their limitations are discussed.

3.3 Social media and polarisation

Social media platforms are new and alternative communication platforms where users consume news. With millions of users on these platforms globally attending them daily, social media have gained a prominent position in society. However, what exactly is the societal relevance of social media is debated. For example, after 'fake news' on social media has been found to play a part in the 2016 US presidential election, many studies followed on the influence of social media on ideological polarization (Allcott & Gentzkow, 2017; Looijenga, 2018; Waikhom & Goswami, 2019). The latter studies show there is a correlation between the uprise of social media use and societal ideological polarization (Spohr, 2017; Hong & Kim, 2016). As polarization can change a political debate, this example shows that social media can fuel societal developments (Hong & Kim, 2016; Lee, Shin, & Hong, 2018; Lee F. , 2016; Spohr, 2017).

What the exact mechanisms of the interplay are between these platforms and societal developments is debated. Already in 2010, a study showed that, in Germany, Twitter was widely being used for discussing politics and political preferences (Tumasjan, Sprenger, Sandner, & Welpe, 2010). A study from 2018 on social media usage in South Korea points out that the use of social media is correlated with an increased political engagement, both online and offline. It seems exposure to political content online might lead to an increase in offline political engagement, like having political discussions and voting (Lee, Shin, & Hong, 2018). Another study on this phenomenon, based on data from Hong Kong, asserts that the polarizing effect of social media is intensified during periods of amplified political tensions (Lee F. , 2016). It is clear

there are connections between the global rise of social media and political engagement, yet the dynamics between what is said on social media and societal developments are only partially known. Furthermore, these dynamics may differ heavily from country to country, and results from other countries may not be directly applied to the Dutch context.

3.4 Twitter as representative of public opinion

Twitter is often used for studying public opinion. The reasons for this are mentioned previously (section 2.4 in the *Introduction* and sections 3.3 and 3.4 in *Related work*). However, there are some important limitations a researcher should consider when using Twitter as a data source for decision making. First of all, there is an overrepresentation of users that are younger than 40 (Statista, 2021a; Statista, 2021b). Second, the internet tends to have higher engagement of a small group of frequent users, while many users post very little (Mustafaraj, Finn, Whitlock, & Metaxas, 2011; Tumasjan, Sprenger, Sandner, & Welpe, 2010). This fact about internet use seems in line with the fact that 44% of all Twitter accounts never post Tweets after creation (Omnicore Agency, 2021). Although many politicians, athletes, artists and other influencers take to Twitter these days, there are influential people that are not social media. This important to note when using twitter for researching public opinion, otherwise important key actors or influencers can go unnoticed. Finally, social movements like #MeToo and Black Lives Matter succeeded in creating a lot of awareness and attention using social media (Manikonda, Beigi, Kambhampati, & Liu, 2018; Mundt, Ross, & Burnett, 2018). However, converting these decentralized movements to actual policy change is not at all straightforward (Malchik, 2019). Interestingly, during the Nitrogen Crisis the farmers had some success with influencing regional policy with their demonstrations (Klumpenaar & Van Laarhoven, 2019). However, the Nitrogen Crisis is still ongoing on a national level. When using twitter as a data source for analysing public discourse for decision making, a decision-maker should consider all the limitations of Twitter, and preferably complementary sources should be inquired for more robust decision making.

3.5 Politicians on social media

In terms of the political playground, nowadays there is interaction between politicians and social media platforms. On the one hand, it has been shown that the participation of politicians on social media can lead to spikes in the discussion of presented topics and the increase in populist views on these platforms (Engesser, Ernst, Esser, & Büchel, 2016). On the other hand, social media platforms can affect the behaviour of politicians and agenda setting as well. A comparative study on online populism and disinformation between the Netherlands and the USA shows that, in both countries, there have already been instances of conservative populist politicians choosing the 'common opinion' of social media users in their country over expert

views and factual evidence (Hameleers, 2020). Thus, there are indications that there is a two-way interaction between politics and social media.

3.6 NLP applied

NLP techniques have been applied in research on news articles, framing, social media trend discovery, health, climate change debates and many more. A study on public opinion on the Dutch measle outbreak studied co-occurrences of reports on measle infection cases and opinion patterns on social media. It showed that the monitoring of social media platforms could help in the formulation of vaccination policies and, in that way, contribute to vaccination acceptability (Mollema, et al., 2015). Comparably, Blankers et al. found the monitoring of online forums could contribute to early detection of increased popularity of novel psychoactive drugs (Blankers, Van der Gouwe, & Van Laar, 2019). NLP has also been applied to paint a picture of climate change denial and debates in the US (Boussalis & Coan, 2016). Finally, topic modelling has been applied to inform policy-making from US citizen e-participation data (Hagen, et al., 2015).

In this study, the NLP methods sentiment analysis and topic modelling are applied, which are discussed in more depth in the following sections.

3.7 Sentiment analysis

Although various sentiment analysis approaches give different results on different scales, most tools return the sentiment scores on a linear scale from negative to positive (Mäntylä, Graziotin, & Kuutila, 2018). However, there are newer approaches to sentiment analysis which consider, for example, that a message can contain both positive and negative sentiment at the same time and return two scores, or go further than classifying emotions as positive and negative, but try to distinguish between various specific emotions (Cambria, Gastaldo, & Bisio, 2015; Thelwall, Buckley, Paltoglou, & Kappas, 2010).

3.7.1 Sentiment analysis and Twitter

In early years of sentiment analysis, the most researched datasets comprised online reviews, for example, for movies or, more interestingly for commercial purposes, of products. Nowadays, though, the vast majority of sentiment analysis studies focus on social media platforms like Twitter and Facebook (Mäntylä, Graziotin, & Kuutila, 2018). Sentiment analysis can reveal information on positive or negative attitudes towards, for example, support of social movements, political parties or events. In 2018, one of the three most cited papers in Scopus and Google Scholar on sentiment analysis examine whether sentiment in tweets can help predict election

results in Germany (Tumasjan, Sprenger, Sandner, & Welpe, 2010; Mäntylä, Graziotin, & Kuutila, 2018). Another such study, looking at the popularity of Italian political leaders in 2011 and the voting intention of French Internet users in the 2012 presidential ballot, found the results of their analysis showed a high correlation with the data of more official mass surveys. Additionally, they found their analysis showed predictive ability for the election outcomes (Ceron, Curini, Iacus, & Porro, 2014). This shows the potential of applying sentiment analysis to social media for characterizing the attitudes of its users.

3.7.2 Challenges of applying sentiment analysis on social media

Although the application of sentiment analysis on social media data, like Twitter, is very promising, there are some important challenges to consider (Zhang, Xu, & Jiang, 2018). Tweets often contain slang, misspelled words and overflow with emojis, and are, of course, very short pieces of text, which can make it difficult for the sentiment analysis tool to correctly score the sentiment of the Tweet (Giachanou & Crestani, 2016; Zhang, Xu, & Jiang, 2018). Other great challenges to sentiment analysis which may lead to misclassification of the tweets are undetectable sarcasm and how to treat synonyms (González-Ibáñez, Muresan, & Wacholder, 2011). To date, no perfect solutions have been found, which complicates the interpretation of insights from applying of sentiment analysis to social media data.

3.8 Topic modelling

Topic modelling is a procedure for automatically calculating a set of topics from a dataset comprising text. To understand topic modelling it is essential to understand the following terms (Vayansky & Kumar, 2020):

- *Corpus*: the dataset, consisting of documents
- *Word*: each unique word in a corpus is indexed
- *Document*: a set of words, in bag-of-words representation, representing
- *Topic*: distribution of a pre-set vocabulary

The idea behind topic modelling is to calculate what topics are present in the corpus. The words and their probabilities of belonging to a certain topic are calculated based on the cooccurrences of words in the documents. This way, it is possible that a word belongs to multiple topics with varying probabilities. A corpus can consist of, for example, movie reviews, journal articles or, in the case of this study, tweets (Vayansky & Kumar, 2020).

3.8.1 LDA

The most popular algorithm for topic modelling is Latent Dirichlet Allocation (LDA). Originally developed by Blei et al. in 2003, LDA is a topic modelling algorithm that takes as input a corpus,

a vocabulary matrix β (beta) and a parameter for the number of topics that it should find (Blei, Ng, & Jordan, 2003). It returns the word prevalence distributions per topic (words that belong to the topic and the likelihood they belong to that topic, see Figure 3.1 for an example) and the topic prevalence distribution per document (topics that belong to the document and the likelihood they belong to that document) (Blei, Ng, & Jordan, Latent Dirichlet allocation, 2003). This is done through an iterative process where initially each word in the vocabulary matrix is arbitrarily appointed to one of the topics, and the distributions are randomly assigned. Then, for each word per document, both the chance that the word's assigned topic belongs in that document and the chance of that word actually belonging to its assigned topic are calculated. Based on these outcomes, the distributions are updated and the process continues until the algorithm converges (Liu, Tang, Dong, & Yao, 2016). Unlike sentiment analysis, LDA, being a statistical approach to NLP, is language independent: as long as the corpus comprises documents in the same language, topic modelling can be applied to corpuses of any language. With topic modelling you can calculate what the distribution is of certain topics over the whole corpus, or look at each individual document and identify the distribution of topics present in it

For a more extensive explanation of LDA and its details, the following publications are recommended: (Blei, Ng, & Jordan, Latent Dirichlet allocation, 2003; Roberts, Stewart, & Airoldi, A Model of Text for Experimentation in the Social Sciences, 2016; Liu, Tang, Dong, & Yao, 2016).

```
------ Topic 1 ------
Stikstofcrisis(0.486) boer(0.404) snelheid(0.290) kabinet(0.183) verminderen(0.067)...

------ Topic 2 ------
boer (0.45) landbouw(0.394) subsidie(0.276) maatregel(0.151) negeren (0.090)...

------ Topic 3 ------
boerenprotest(0.427) crisis (0.357) denhaag(0.237) pfa(0.138) groot(0.012)...
```

Figure 3.1 | Example of three topics and their word prevalence distributions.

Although LDA often has good results, there are challenges to applying it to short texts, like tweets (Steinskog, Therkelsen, & Gambäck, 2017). A study comparing the content of tweets to a traditional news medium, the New York Times, found that the standard LDA model performs poorly on tweets (Zhao, et al., 2011). Multiple studies have focussed on improving LDA results, for example by clustering tweets into bigger pieces of text (tweet pooling) and using hashtags for automatic labelling (Luyi & Wei Song, 2016; Mehrotra, Sanner, Buntine, & Xie, 2013; Zhao, et al., 2011). These extensions to the classic LDA algorithm often lead to better results. However, in this study, due to limited time resource, the classic LDA approach is used.

One of the biggest downsides of classic LDA is that the generated topics are not labelled, hence human interpreters are needed to label topics. There are various studies on automatic labelling of topics (Allahyari, Pouriyeh, Kochut, & Arabnia, 2017; Mehrotra, Sanner, Buntine, & Xie, 2013). However, they have mixed results and some require external data. Even if automatic labelling

was flawless and could be easily implemented, in case of a crisis that is being discussed on Twitter it still begs the question: when should decision-makers look at these topics discussed online? To focus this question, in this research the difference between topics over time is more of interest than the details of each individual topic model and its topics. This, however, is not an extensive research field. The developers of the original LDA algorithm have created an extension to it called 'dynamic topic modelling'. Where in traditional LDA the document order in the corpus is irrelevant to the algorithm, dynamic topic modelling looks at differences in how topics are composed. This method adds to the traditional LDA approach a new way of browsing large unstructured datasets, and can be used as predictive models (Blei & Lafferty, 2006). However, the necessity to evaluate individual topics also applies to dynamic topic modelling, as is shown by Bashar, Nayak, & Balasubramaniam (2020). There is no example found in literature of an application of topic modelling that allows detecting differences in topics over time without analysing the individual topics in topic models.

4 Methods

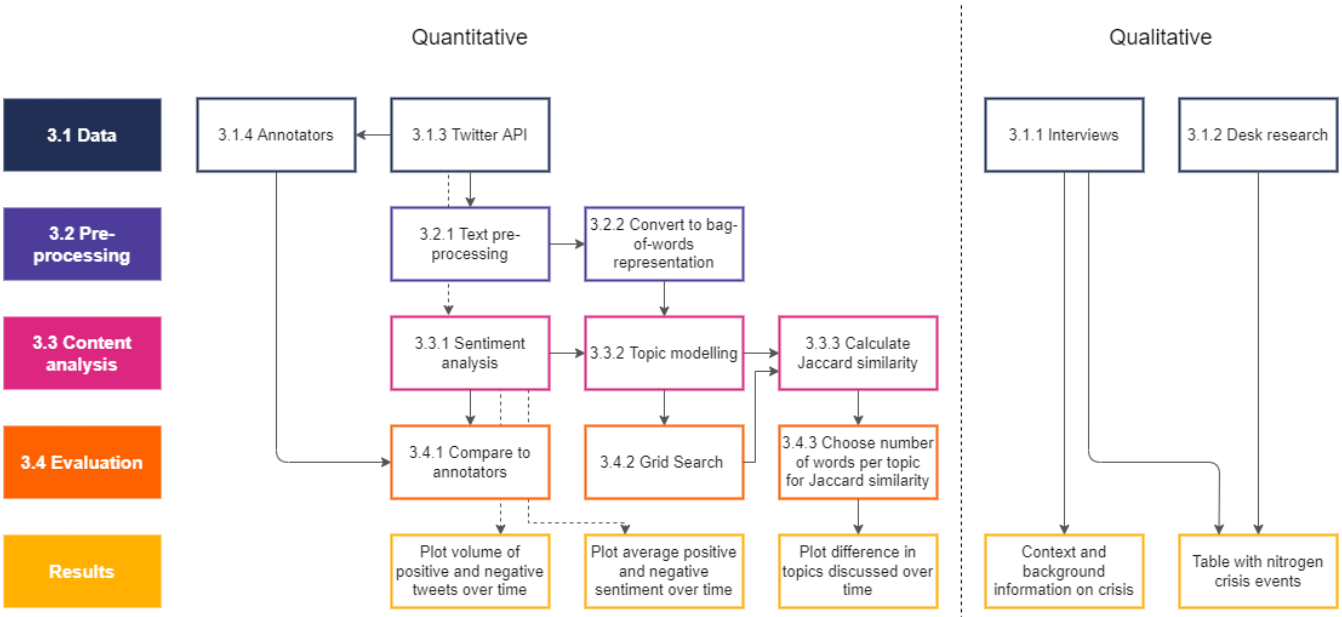


Figure 4.1 | Research approach, showing how information flows through the steps in this research.

This chapter discusses every step taken in this research. Figure 4.1 above shows an overview of the methodology of this research, and how information flows through it. Figure 4.1 includes the chapter section per step that covers it. On the left it shows the quantitative side, and the qualitative side is depicted on the right. Each color represents a layer of the study. First, all the data sources and how data was retrieved are discussed in section 4.3: *Data*. Secondly, section 4.4: *Pre-processing: preparing tweets for analysis* describes how the data was processed to prepare it for the NLP methods, which are described in section 4.5: *Content* analysis. Thereafter, this chapter will describe how the methods applied in the content analysis are evaluated in 4.6: *Evaluation*. This section describes how the best performing sentiment analysis tool is chosen, and how the topic modelling configurations leading to the best topic models are found. Beneath the Evaluation layer, Figure 4.1 also shows how the results of the content analysis and evaluation layers feed into the results. These results are covered in the next chapter, *Results*. Lastly, this chapter concludes with section4.7: *Data and code* which describes where to find the data and code used in this research.

4.3 Data

4.3.1 Interviews

Two interviews were held with Han de Groot, who completed a project as an external advisor for the DG Stikstof (governmental organisation responsible for tackling the Nitrogen Crisis), supervised by the minister of Agriculture, Nature and Food Quality. For the assignment, he talked to many important stakeholders involved in the Nitrogen Crisis and mapped their desires and proposed strategies for overcoming stakeholder challenges (De Groot, 2021). The first interview with De Groot took place on the 25th of May 2021. It was a semi-structured videocall interview (the questions list is enclosed in Appendix A: Questions from first interview Han de Groot) and served to provide the following:

- In-depth knowledge on the stakeholder project
- Extend the list of keywords used in the query for data collection

In short, the interview provided context on the Nitrogen Crisis, the stakeholders involved and their stances, and was used to sharpen the data collection process by looking at the search query.

The second interview took place on the 1st of June (questions are in Appendix B: Questions from second interview Han de Groot). The purpose of this interview was to:

- Identify significant events during Nitrogen Crisis
- Ask about topics that De Groot expected to be discussed on social media
- Provide context on what specifically went wrong in terms of policy-making in causing the Nitrogen Crisis
- Ask what De Groot thinks the perks are of the Twitter analysis and the volume and speed of its content.
- Present the choice of methods, and ask De Groot how the application of these methods could be useful to policy-makers

The second interview was important for defining how to proceed with the chosen methods, sentiment analysis and topic modelling, and what type of insights could serve decision-makers during a crisis like the Dutch Nitrogen Crisis.

4.3.2 Desk research: Identifying events with high news coverage

Desk research was performed to identify the events during the Nitrogen crisis that received a lot of attention from the news media. News articles were the primary source of information. However, (news) sources with an overview of all major events throughout the whole Nitrogen Crisis were scarce. Some news sources had a special page on their website with a collection of their news articles on the Nitrogen Crisis, but these would only contain news articles from 2019, while the Nitrogen Crisis is still ongoing at the time of writing. After a long search, it turned out

that Wikipedia had the most complete, well-sourced and comprehensive overview of the Nitrogen Crisis and the demonstrations that took place (Wikipedia, n.d.). Therefore, this page was taken as the main source for identifying important events.

4.3.3 Twitter API

After some experimentation with various data collection methods, the Twitter Developer Portal was used for the collection of Tweets. In order to request data from the Portal for non-commercial purposes, a researcher needs to fill in an application and explain what the Portal will be used for and what the research is about. If this request gets approved, one gets access to the 'Standard product track', one of various types of Portal access with each different rate limits.

Upon being granted access, a Bearer token was generated and a Python script with a query was written to retrieve tweets from the past using the Twitter API. In order not to exceed Twitter API rate limits, the script only does 299 requests per 15 minutes[3]. The parameters inserted define the timeframe, type of tweets and the key words used to gather tweets. Keyword parameters were found through a process of trial and error: starting with the words "stikstof" (Eng.: Nitrogen) and "Stikstof Crisis" (Eng.: Nitrogen Crisis) all tweets were collected that included either of these terms. Looking at the resulting tweets with the aim of extracting more terms associated with the Dutch Nitrogen Crisis, a list of terms was created with keywords, as seen in Table 4.1. The `OR` operator is used to collect tweets that contain either one or more of the keywords in the list. When keywords are put between quotation marks, only tweets containing the keywords in that exact same order are returned. When keywords are between brackets, Tweets are returned when all the keywords between the brackets are present in the tweet, regardless of the sequence of the words. As is visible in Table 4.1, many keywords are in the query twice: once as a word and once as a hashtag. This did not apply to all keywords though, as for example the word "pas", apart from being an abbreviation for the Dutch Nitrogen policy, is also a very common Dutch word with many meanings (e.g., "step", "just now", "pass", "only", "just" etc.). This is why the word "pas" is only included in the query when it is preceded by a hashtag: when "#PAS" is present in a tweet it will very likely be about the Nitrogen Crisis. Similarly, the word "stikstof" (Eng.: Nitrogen) on its own would result in too many tweets about, amongst others, biological processes that have nothing to do with the Crisis. Lastly, it is important to note that only tweets that are not retweets are collected.

[3] https://developer.twitter.com/en/docs/rate-limits

Table 4.1 | Query parameters for the retrieval of the dataset.

PARAMETER	VALUE INSTERED		
Start_time	2019-01-01T00:00:00Z		
End_time	2021-07-31T00:00:00Z		
Language	nl		
type	all but retweets		
Exclude	Mercosur OR #mercosur		
Include	stikstofcrisis OR PFAS OR #PFAS OR boerenprotest OR boerenprotesten OR #boerenprotest OR #boerenprotesten OR bouwprotest OR #bouwprotest OR grondinverzet OR grondverzet OR #stikstof OR #stikstofcrisis OR #PAS OR	"programma aanpak stikstof" OR "programma aanpak #stikstof" OR #stikstofbudget OR stikstofbeleid OR #stikstofbeleid OR (stikstof uitstoot) OR (stikstof crisis) OR (stikstof uitspraak) OR (stikstof probleem) OR stikstofdepositie OR #stikstofdepositie OR #grondinverzet OR #grondverzet OR #boerenterreur OR	(boeren terreur) OR #stikstofdebat OR stikstofdebat OR stikstofprobleem OR #stikstofprobleem OR stikstofgedoe OR #stikstofgedoe OR (stikstof gedoe) OR farmersdefenceforce OR #farmersdefenceforce "Stikstof-probleem" OR stikstofbudget OR

Table 4.2 shows the number of nitrogen tweets per month (see Figure 8.1 in Appendix 1 for a plot). Because up until June the number of tweets per month is lower than 1000 tweets per month these are removed preventatively (as indicated by the striped background), in order for these low numbers not to skew the sentiment analysis and topic modelling results later. In total, 1636 tweets are deleted.

Table 4.2 | Number of tweets per month in the dataset.

YEAR	MONTH	NUMBER OF TWEETS
2019	January	175
	February	336
	March	217
	April	205
	May	703
	June	2034
	July	1331
	August	1526
	September	15224
	October	85449
	November	31824
	December	29036
2020	January	8079
	February	15422
	March	6005
	April	4415
	May	2669
	June	4604
	July	23810
	August	6398
	September	4202
	October	5719
	November	11594
	December	6213
2021	January	2555
	February	2129
	March	3975
	April	2055
	May	3141
	June	4933
	July	7903

4.3.4 Annotators

In this study, two sentiment analysis tools are applied. To select the best of those two, 100 Tweets have been annotated by three annotators and the sentiment scores of the annotators have been compared to the scores by the sentiment analysis tools. The annotators are two TU Delft Masters students and one recent TU Delft graduate. They are Dutch females aged between 22 and 28 and from similar social and cultural backgrounds. The writer of this study is among them.

4.4 Pre-processing: preparing tweets for analysis

In order to use text as input for machine learning algorithms regardless multiple cleaning and pre-processing steps are required. The next sections describe these steps to prepare tweets for topic modelling.

4.4.1 Text pre-processing

Figure 4.2 shows the steps taken for the pre-processing of text in preparation for topic modelling. From left to right, the first step in 'Text pre-processing' is the removal of line breaks. Then, the accents on characters are removed. For example, the word "misère" (Eng.: misery) is turned into "misere". This is done because, on Twitter, while some users are very strict with traditional spelling, many users are more careless and will, among other things, ignore the characters on words. After this, all characters are converted to lower case, so, for example, the word "Dutch" and "dutch" will not be treated as two distinct words. Subsequently, special characters, like "&" and "#", are removed. After this, all words are lemmatised. Lemmatisation converts words into their root form, known as lemma (Nandathilaka, Ahangama, & Weerasuriya, 2018). This means all plural forms of words will be converted to their singular form (e.g., "trees" → "tree"), verbs are converted to the first person regular (e.g., "demonstrating" → "demonstrate"), etc. Lemmatisation tries to take the context of a word into account. For example, the word "running" can lemmatised differently depended on context. In the sentence "I am running to the train station" the lemma is "run", while in the sentence "The actor is in the running for an Oscar" the lemma would be "running". For this step, the lemmatizer of Python library SpaCy is applied.

The next step, is to remove stop words. In NLP, these are words so common and frequently used that they add little meaning to a sentence. Examples of stop words are 'a', 'the', 'of' and 'it'. There are various ready-made lists of stop words. Here, the Dutch stop words list from NLTK, a rich Python library for NLP, is used (Bird, Klein, & Loper, 2009). If, in the process of analysing the results, more words are identified that have little added value, these words can be added to the list of words that are filtered out. The words that the list of stop words was extended with are listed in Appendix D:List of stop words.

4.4.2 Convert to bag-of-words representation

For topic modelling, two more steps are required, which are shown under 'Prepare for topic modelling' in Figure 4.2. First, the tweets are tokenized, meaning that instead of being a long string, the sentences are split on every space into single words. Then, a Dictionary is created with each unique word and how often it occurs in the corpus. In order to reduce the size of the corpus, shorten processing time for topic modelling and increase the chance of coherent models, both the most occurring words and the least occurring words are filtered out: words that occur a lot will probably not have a specific relation to a topic, and words that are barely used are of little value. Here, words which occur in less than 2 documents and words which occur in over 50% of all documents are deleted. Then, on the right of Figure 4.2 the bag-of-words representation of the corpus remains, which is ready to be used for topic modelling.

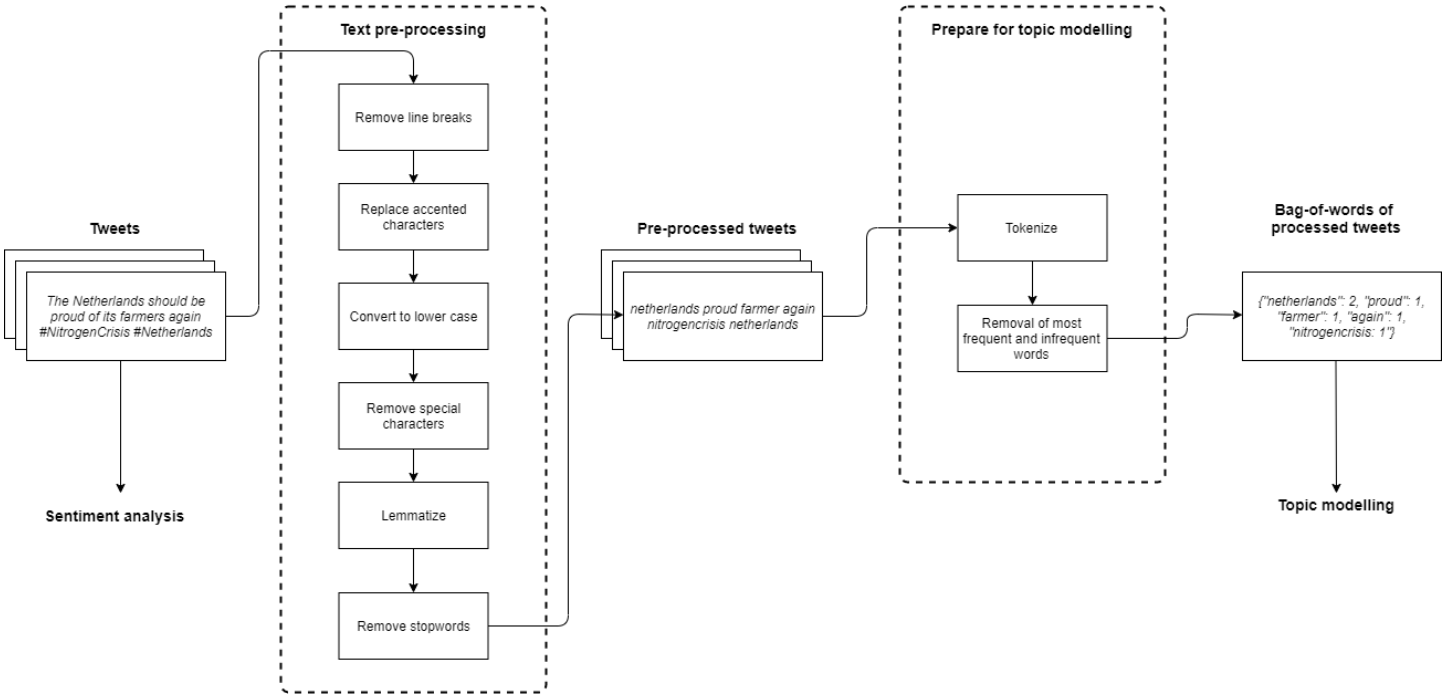


Figure 4.2 | Steps in pre-processing tweets.

4.5 Content analysis

4.5.1 Sentiment Analysis

As Dutch is a language spoken natively by roughly only 23 million people, its sentiment analysis tools are limited compared to the English tools available (Trilling & Boumans, 2018). Unfortunately, because of inherent differences between languages, a sentiment analysis tool for one language can usually not be blindly applied to another (Zhang, Xu, & Jiang, 2018). There is no one-size-fits-all sentiment analysis tool readily available in Dutch, so in order to apply sentiment analysis to a Dutch piece of text, it is best to pick a sentiment analysis tool that is especially developed for the type of text that is being analysed (Trilling & Boumans, 2018). In the following sections, the two sentiment analysis techniques used in this study are described: SentiStrength and Pattern. These tools were selected because they are available for free and easy to implement. Also, they both scale sentiment as 'positive' or 'negative', although on different scales, as discussed in the next sections.

SentiStrength

SentiStrength (SS) is a tool originally developed in 2010 for social media text (Thelwall, Buckley, Paltoglou, & Kappas, 2010). Interestingly, SentiStrength outputs its sentiment results in two separate sentiment strength scores: one from -1 to -5 for negativity, and the other from 1 to 5 for positivity. The developers explain their choice for using two separate scores instead of one by referring to psychological research that shows 'mixed emotions' are real: a person can experience two opposing emotions at the same time (Berrios, Totterdell, & Kellett, 2015).

The core of SentiStrength is an annotated word list, which is annotated by human evaluators with scores of +/-2 to +/- 5. Then, the tool contains several other annotated word lists for several purposes. There is a 'booster word' list containing words that can either enhance or reduce the sentiment intensity of the following word (e.g., 'extremely' increases the sentiment score of the following word, while 'some' lowers it) (Thelwall, Buckley, Paltoglou, & Kappas, 2010, p. 2549). SentiStrength also takes punctuation into account (e.g., an exclamation mark increases the sentiment score of the sentence), contains a list of sentiment scores of emoticons and corrects spelling mistakes. The emoticon list is not used in this study, as by 2021 emoticons are hardly used: social media users use *emojis* now (e.g. 🙃).

SentiStrength has expanded to classify various languages other than English. Several studies have applied SentiStrength in multiple languages, for example for looking at gender bias in English sentiment analysis and for approximating national happiness in Turkey (Durahim & Coşkun, 2015; Thelwall, 2018).

Pattern

Pattern is a Python package that provides several NLP options, amongst which sentiment analysis. Like SentiStrength, it has an annotated list of adjectives at its core, and for the Dutch Pattern version this list was generated by collecting the 1000 most frequent adjectives in a mined dataset of Dutch book reviews (De Smedt & Daelemans, 2012). These adjectives were annotated by human evaluators, and the list has been further expanded with annotations from another corpus. Pattern'' sentiment analysis function returns a value between -1 and 1 (continuous scale), where -1 is very negative and 1 is very positive. Pattern is unfortunately not actively maintained and updated anymore (De Smedt & Daelemans, 2012).

4.5.2 Topic modelling: LDA

While there are many implementations of LDA, Gensim LDA, which is an open-source library for unsupervised topic modelling and natural language processing is chosen (Řehůřek & Sojka, 2011). This is a well-documented and widely used library with many advanced functionalities, available in Python. Because of this, it is the choice for this study.

Because in this study it interests to look at differences in topic models over time, instead of exploring individual topic models in depth, multiple topic models need to be generated. For this, two choices need to be made:

- The time windows to slice the data in
- The number of topics

These choices will be made by performing a grid search. Because the grid search is part of the Evaluation layer of this research, it is discussed later on in section 4.6.2. For now, keep in mind that the choice for the time window and number of topics is made so that the resulting topic models will be of the highest quality. Now, the next section describes how the difference in topics discussed over time is calculated with these topic models.

4.5.3 Calculating difference between topic models: mean Jaccard similarity

The grid search establishes what time window and number of topics best to pick to generate topic models for each time window in the dataset. So, after performing it, there are three topic models for each time window: one for the full dataset, one for the positive and one for the negative. Now, it is essential to see how topics develop over time. To compare subsequent topic models from subsequent time windows, their Jaccard similarity is calculated (Jaccard, 1912). This works as follows: imagine two topics are compared, of which the words that compose the first are called set A and the second set B. Then the Jaccard similarity is calculated as follows:

$$J(A,B) = \frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A} \cup \mathrm{B}|}$$

Or, in words, the Jaccard similarity between set A and set B is calculated by dividing the *intersection* (the number of words that are present in *both* topics) of the two sets by the *union* (the number of *unique words* in the two topics combined) of the two sets.

Topics are distributions of long lists of words, therefore the number of words that are chosen to calculate the Jaccard similarity between two topics with is important to the outcome. To look at the behaviour of the Jaccard similarity formula, Figure 4.3 shows the Jaccard similarity of two topics depending on the number of words the topics have in common for two different topic sizes: 10 words per topic on the left, and 20 words per topic on the right. When the number of overlapping words is low, the Jaccard similarity stays low. However, the more overlapping words, the steadier the increase of the Jaccard similarity. Note that the higher the number of words per topic, the higher the Jaccard similarity can be. It is also important to consider that the Jaccard similarity does not consider the distribution of the words in each topic. This way, some information contained in topics is lost.

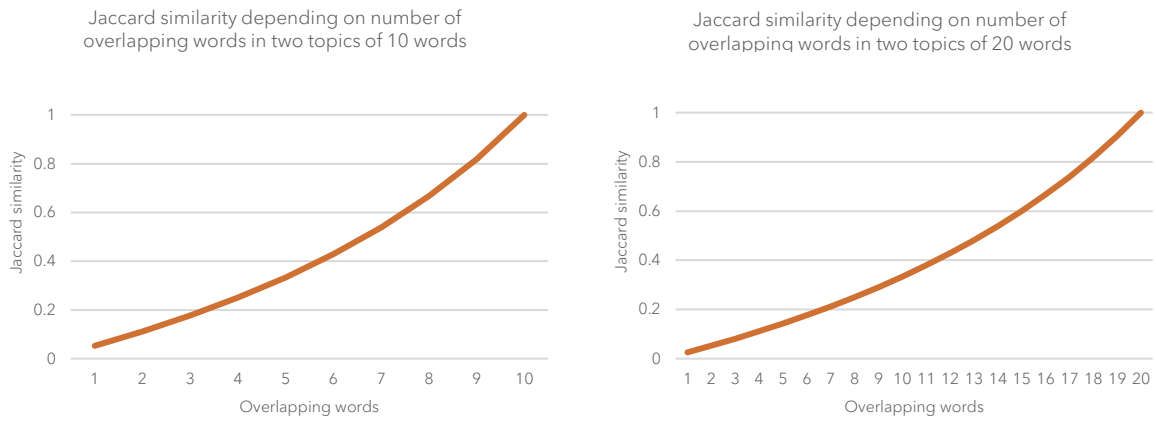


Figure 4.3 | Jaccard similarity per number of words that overlap in two topics.
On the left: the Jaccard similarity when topics comprise 10 words.
On the right: the Jaccard similarity when topics comprise 20 words

Now, two full topic models can be compared by first calculating the Jaccard similarity of *each combination of topics in each model,* as shown in Figure 4.6, and then *taking the mean* of this sequence of Jaccard similarities. This means that if we have topic models of 15 topics, $15x15 = 225$ similarities are calculated, of which the mean will be the mean Jaccard similarity between the two topic models.

The mean Jaccard similarity is an easily implementable index that can aggregate the complex data structure of a topic model in a value, which makes it elegant and interpretable.

4.6 Evaluation

4.6.1 Best Sentiment Analysis method: Compare to annotators

In order to evaluate the quality of the two sentiment analysis tools and pick the one most suitable for this research, the following sections show how scores of each method are compared to the annotations of three people.

Dataset for annotation: 100 tweets

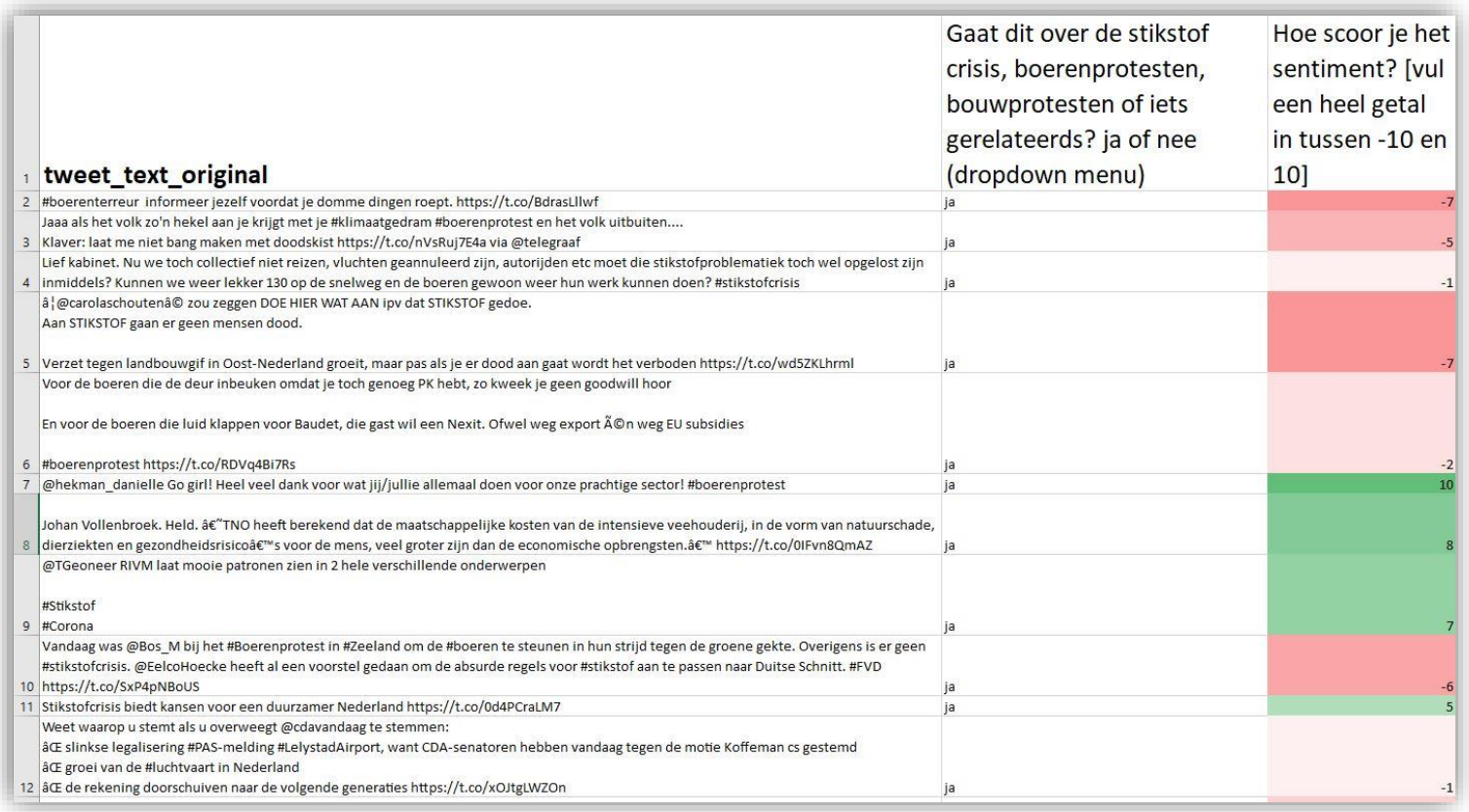
In order to pick the most suitable of the two sentiment analysis techniques, Pattern and SS, three annotators were asked to score 100 tweets on sentiment.

These tweets were selected as follows:

- 20 randomly picked tweets with a positive Pattern score
- 20 randomly picked tweets with a negative Pattern score
- 20 randomly picked tweets with a higher positive than negative SS score
- 20 randomly picked tweets with a higher negative than positive SS score
- 20 randomly picked tweets from the full dataset

The first 80 tweets are selected this way to make sure that in the annotated dataset there are tweets that score both high and low on according to both sentiment analysis methods. The 100 tweets were shuffled before being presented to the annotators.

Figure 4.4 shows a snapshot of the excel files that the annotators received for scoring the sentiment of tweets. The first column shows the tweet. In the second column the annotator indicated whether they think the tweet is about the Nitrogen crisis or not ('ja': yes, 'nee': no or 'onduidelijk': unclear). In the third column they are asked to score the sentiment in the tweet on a scale from -10 (very negative) to 10 (very positive). As Figure 4.4 shows, when inserting a value for the sentiment the cell changes colour according to the inserted value. It goes in a gradient from intense red (-10) slowly to white (0) and then to intense green (10).



	tweet_text_original	Gaat dit over de stikstof crisis, boerenprotesten, bouwprotesten of iets gerelateerds? ja of nee (dropdown menu)	Hoe scoor je het sentiment? [vul een heel getal in tussen -10 en 10]
2	#boerenterreur informeer jezelf voordat je domme dingen roept. https://t.co/BdrasLllwf	ja	-7
3	Jaaa als het volk zo'n hekel aan je krijgt met je #klimaatgedram #boerenprotest en het volk uitbuiten.... Klaver: laat me niet bang maken met doodskist https://t.co/nVsRuj7E4a via @telegraaf	ja	-5
4	Lief kabinet. Nu we toch collectief niet reizen, vluchten geannuleerd zijn, autorijden etc moet die stikstofproblematiek toch wel opgelost zijn inmiddels? Kunnen we weer lekker 130 op de snelweg en de boeren gewoon weer hun werk kunnen doen? #stikstofcrisis	ja	-1
5	â¦@carolaschoutenâ© zou zeggen DOE HIER WAT AAN ipv dat STIKSTOF gedoe. Aan STIKSTOF gaan er geen mensen dood. Verzet tegen landbouwgif in Oost-Nederland groeit, maar pas als je er dood aan gaat wordt het verboden https://t.co/wd5ZKLhrml	ja	-7
6	Voor de boeren die de deur inbeuken omdat je toch genoeg PK hebt, zo kweek je geen goodwill hoor En voor de boeren die luid klappen voor Baudet, die gast wil een Nexit. Ofwel weg export Ã©n weg EU subsidies #boerenprotest https://t.co/RDVq4Bi7Rs	ja	-2
7	@hekman_danielle Go girl! Heel veel dank voor wat jij/jullie allemaal doen voor onze prachtige sector! #boerenprotest	ja	10
8	Johan Vollenbroek. Held. â€™TNO heeft berekend dat de maatschappelijke kosten van de intensieve veehouderij, in de vorm van natuurschade, dierziekten en gezondheidsrisicoâ€™s voor de mens, veel groter zijn dan de economische opbrengsten.â€™ https://t.co/0IFvn8QmAZ	ja	8
9	@TGeoneer RIVM laat mooie patronen zien in 2 hele verschillende onderwerpen #Stikstof #Corona	ja	7
10	Vandaag was @Bos_M bij het #Boerenprotest in #Zeeland om de #boeren te steunen in hun strijd tegen de groene gekte. Overigens is er geen #stikstofcrisis. @EelcoHoecke heeft al een voorstel gedaan om de absurde regels voor #stikstof aan te passen naar Duitse Schnitt. #FVD https://t.co/SxP4pNBoUS	ja	-6
11	Stikstofcrisis biedt kansen voor een duurzamer Nederland https://t.co/0d4PCraLM7	ja	5
12	Weet waarop u stemt als u overweegt @cdavandaag te stemmen: âŒ slinkse legalisering #PAS-melding #LelystadAirport, want CDA-senatoren hebben vandaag tegen de motie Koffeman cs gestemd âŒ groei van de #luchtvaart in Nederland âŒ de rekening doorschuiven naar de volgende generaties https://t.co/xOJtgLWZOn	ja	-1

Figure 4.4 | Snapshot of the (filled in) excel sheet with 100 tweets.
Annotators fill in whether the tweet is about the Nitrogen Crisis and score the sentiment on a scale of -10 to 10.

Calculating correlation between annotators: intraclass correlation

Using the annotated tweets of the three annotators for comparison, one of the two sentiment analysis tools is chosen. For this, the results of each method need to be compared to the dataset of annotated tweets. To find this correlation, the IntraClass Correlation (ICC) is calculated. This is a formula designed especially for assessing the soundness of ratings by different subjects, by comparing the variability of the variance of each individual rating to the variance over all ratings and all subjects. There are various versions of the ICC, of which one needs to be chosen based on if the annotators represent a population or whether they are the only people of interest, and whether there were one or more measurements (Shrout & Fleiss, 1979). In this study, based on the three annotators, one sentiment analysis tool is selected as 'better' so the annotators represent how Dutch people would score the sentiment of these tweets. Also, they were asked to score the tweets just once. Therefore, this study uses ICC2: Single random raters.

The ICC gives insight into how much the annotators agree with each other in the scoring of 100 tweets, after which the sentiment analysis tool whose scores overlap best with the average score of the annotators is selected for the continuation of this study. The higher the ICC, the more the annotators agree on how tweets are scored, and the more confidently the annotator scores can be used for picking the best sentiment analysis tool.

Calculating correlation between annotators and sentiment analysis techniques

After the ICC is calculated to evaluate the annotations, the average scores of the annotators per tweet is calculated and compared to each sentiment analysis score. For the latter, the Pearson Correlation Coefficient (PCC) is calculated. The PCC gives insight into the linear relation between two datasets and returns a normalized value between -1 and 1: -1 shows a perfect negative relation, 1 a perfect positive relation, and 0 shows no correlation. The closer the value is to 1, the more the annotators and the sentiment analysis tool agree. Generally, if the coefficient value lies between +/-0.5 and +/- 1, a strong correlation is assumed, while between +/-0.3 and +/- 49 it is considered moderate, and below +/-0.29 low.

For both SS and Pattern, the PCC is calculated between the average annotators' score and SS, and the average annotators' score and Pattern score of the 100 tweets. The sentiment analysis tool that has the highest correlation with the annotations is chosen as the preferred method in this study.

4.6.2 Choosing number of topics and time windows: Grid search

To find the right settings for time window and number of topics, a grid search is performed. A grid search is a technique for finding the optimum value for parameters by systematically trying out many configurations for these parameters and choosing the configuration that leads to the best results. First, to compare the quality of topic models based on different data slices and with varying number of topics, a metric for topic model quality is determined.

Evaluating the quality of topic models: topic coherence

LDA does not guarantee 'coherent' results: that each topic returned comprises words that are contextually related and properly represent a topic that was described in the corpus that served as training set. The state-of-the-art approach for evaluating the coherence of topics in topic models is human evaluation, a method that is costly, sensitive to biases and, in the case where many topic models need evaluation, extremely time-consuming (Röder, Both, & Hinneburg, 2015). Newman et al. developed a metric that approximates human annotation, called topic model coherence (Newman, Lau, Grieser, & Baldwin, 2010). A high coherence score stands for a more coherent and interpretable topic model. There are various ways to calculate topic coherence, and here the best performing on is selected according to Röder et al.: c_v (Röder, Both, & Hinneburg, 2015). From now on, I will be referring to the c_v topic coherence score

when using the words 'topic coherence'. By selecting topic models based on topic coherence, topic models are picked with a high chance of being interpretable by humans, and thus decision-makers. Topic coherence is a general metric that is widely used for estimating the quality and coherence of topic models and it is implemented in Gensim, hence easily applicable. For an evaluation of other topic coherence metrics, see Röder et al., 2015 (Röder, Both, & Hinneburg, 2015).

Now the quality of a topic model can be approximated through its coherence score, a grid search can be performed. This will decide, firstly, what the time window is to slice data in datasets and calculate a topic model per time window and, secondly, the number of topics these models need to generate. For the data slices, four time periods are tested: 7, 14, 21 and 28 days. This means topic models are generated for every 1, 2, 3 and 4 weeks over the whole dataset. Similarly, the following number of topics are tested: 2, 4, 6, 8, 10, 12, 14 and 16. Due to RAM limits on the remote server, the grid search script ran on, only topic models of up to 16 topics could be calculated.

Figure 3.5 depicts this process and will be described step by step from left to right. On the left, we see the dataset, in its processed bag-of-words representation. Figure 3.5 shows only 6 stacked tweets, in reality there are of course many more, but for illustration purposes we assume there are 6. The lighter are older ones. Now, we cut the dataset per time window. Let us assume the time window for this run is 7 days. It turns out, there are 2 tweets per week, and so for each week we calculate a topic model. However, as the stacked second step in the process depicts, we do this for each number of topics. So, for every week with each 2 tweets, we calculate a topic model with a number of topics = 2, a number of topics = 4, a number of topics = 6 etc. As we try 8 different numbers of topics and we have 3 data slices (one per week) when using a time window of 1 week, we end up with 8*3=24 topic models.

Then, for each topic model for each data slice for each number of topics, the coherence of that topic model is calculated. The coherence score of the sequence of topic models per number of topics is aggregated by calculating the mean of all subsequent topic models. In our example, the three topic models have coherences of 0.39, 0.34 and 0.36, which results in a mean of ≈0.36. The other numbers of topics had averages of 0.40 and 0.38. Now that we have an average coherence score per number of topics, we plot the results on a heatmap (as seen on the right).

This entire process is repeated per time window size. In the example, the time window was one week, so for each number of topics topic models were generated for each week. However, in the next run, the process repeats with a time window of 14 days, and then for 21 and 28 days. In summary: topic models for each time period in the dataset and, per time period, 8 different numbers of topics, are generated and their coherence score calculated. Then, per time window size and number of topics, we take the average of all coherence scores, to compare the *average* quality of topic models per combination of time window and number of topics choice.

Lastly, the process described above is repeated three times: the first run is with the full dataset, the second with only the tweets that are scored positive and the third with only the tweets that

are scored negative. The sentiment score based on which these datasets will be separated depends on which of the two sentiment analysis methods shows most overlap with the human annotators. This way, apart from generating information on what topics are discussed throughout the whole dataset, topic modelling and sentiment analysis are combined here to additionally provide information on the topics discussed per sentiment.

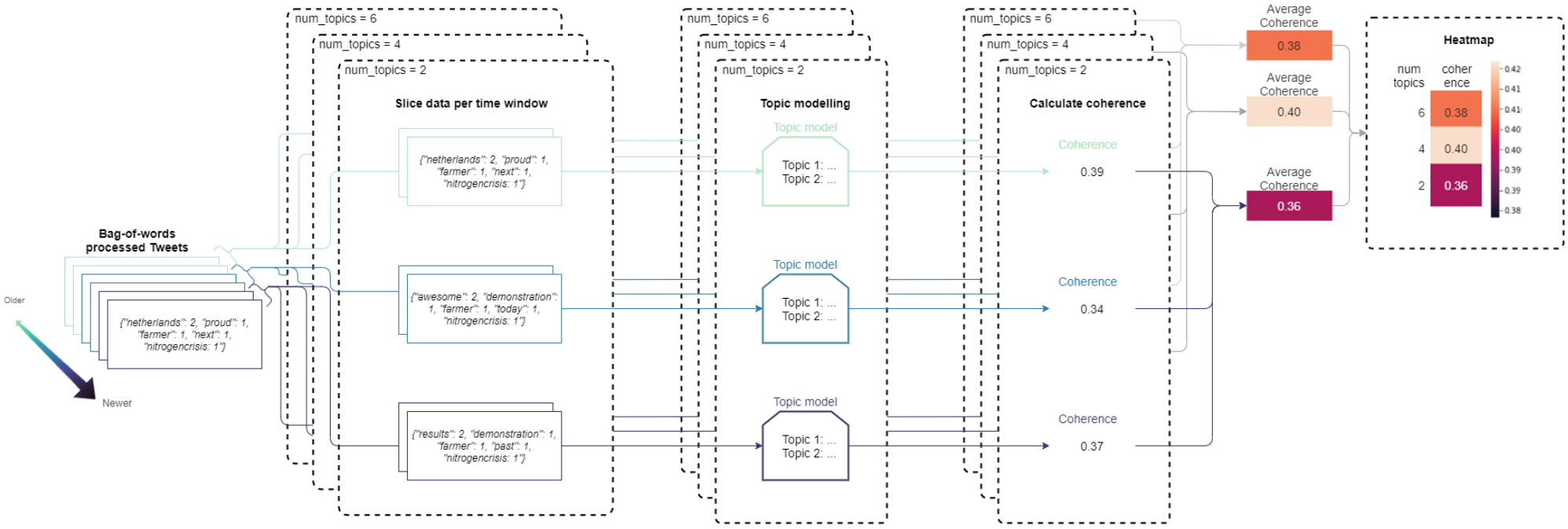


Figure 4.5 | Process steps for the grid search that results in choosing a time window for the dataset data slices and the number of topics to train topics models with.

4.6.3 Choosing the number of words to calculate Jaccard similarity

To illustrate how the Jaccard similarity between topic models changes per choice of words per topic, Figure 4.6 shows 5 different matrixes with the Jaccard similarity between the same two topic models. From left to right, the Jaccard similarity is calculated with the top 10, 20, 30, 40 or 50 words per topic. A high similarity between two topics is represented by the colour blue, while a low similarity will show in red. Figure 4.6 shows that 10 words per topic will lead to a more 'turbulent' matrix: It is the matrix with both the darkest red and the darkest blue. As we will take the mean of the matrix, this turbulence would not be represented in the outcome. When topics comprise 20 words, the matrix is more uniform. Then, very little changes between topics consisting of 20, 30, 40 and 50 words. The choice of the number of words to consider per topic is important: choosing too little words might cause high similarities between topic models if the most important words of different topics are often similar (think of 'farmer' or 'nitrogen'). However, choosing too many words per topic will give words with a low weight the same influence on the outcome as words with a high weight, as the distributions of words per topic are ignored by the Jaccard similarity. The number of words per topic will be chosen in section 5.5.2, eyeballing the plots of the average Jaccard similarity over time for 10, 20, 30 and 40 words per topic. The plot with the most varying behaviour for the mean Jaccard similarity will be chosen as the number of words to include per topic. After all, this research aims to find ways for identifying changes in topics discussed online.

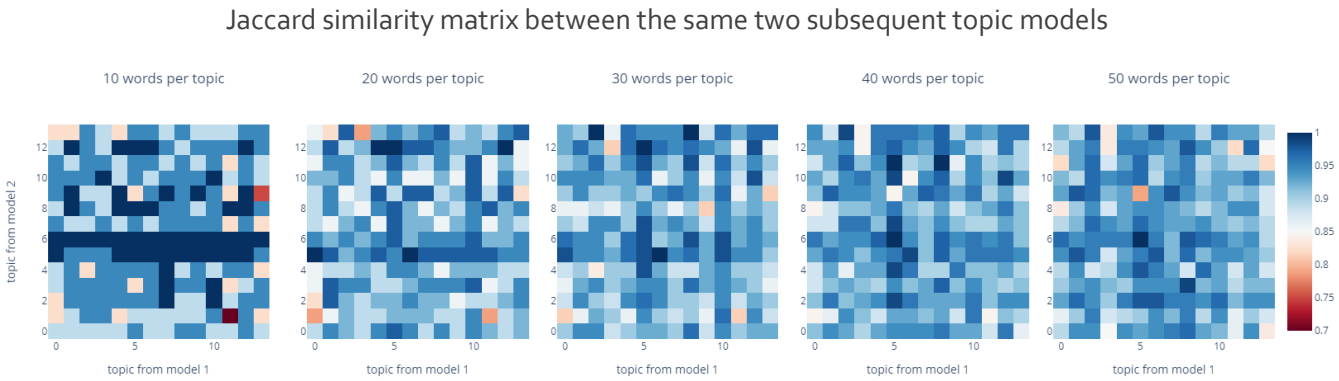


Figure 4.6 | From left to right: the Jaccard similarity matrix between the same two subsequent topic models when using 10, 20, 30, 40 and 50 words as topic representation.

4.7 Data and code

The data and scripts used in this study are available at repository 'ThesisNLP 'on GitHub, under the account of username milahendrikse[4].

Key Findings of Chapter 3: Methods

1. Two interviews are held with Han de Groot to provide context on the Nitrogen Crisis and to specify how the results of this study could be useful to decision-makers
2. A query is set up to collect tweets about the Nitrogen crisis
3. The outcomes of two Dutch sentiment tools are compared with sentiment scores of three annotators. Only the results of the best performing will be included in chapter 4: *Results*
4. The time window and number of topics for topic models will be established with a grid search
5. Sentiment analysis results are used to create two additional datasets: one comprising of positive and one of negative tweets
6. The mean Jaccard similarity is calculated per subsequent topic models to represent changes in topics discussed over time, for each sentiment dataset

SQ2

[4] https://github.com/milahendrikse/ThesisNLP.git

5 Results

This chapter describes the results of the interviews and of the application of sentiment analysis, topic modelling and the calculation of the difference between topic models on the dataset of tweets. First, key insights from the interviews with Han de Groot are described in section 5.3. Then, the identified events that received a lot of media attention are arranged in a table in section 5.3.4. These events will later be plotted in the results of both quantitative methods. Of these methods, first sentiment analysis results are discussed in section 5.4: Sentiment analysis. Based on the tweet annotations, one of the sentiment analysis tools is selected in section 5.4.1, after which the sentiment analysis results are plotted in section 5.4.2, answering RQ1. Lastly, in section 5.5: Topic modelling, the best settings for topic modelling are picked per sentiment based on the grid search in section 5.5.1, the best number of words for Jaccard similarity is chosen in section 5.5.2, and the results are plotted of the Jaccard similarity over time in section 5.5.3, answering RQ3.

5.3 Interviews Han de Groot: Key takeaways

The following key takeaways are based on two interviews held with Han de Groot. They provide background information on the Nitrogen Crisis, characteristics of the interplay between the discourse on Twitter and the behaviour of politicians, and how analysing Twitter discourse could be helpful to decision-makers.

5.3.1 What preceded the crisis

Sometimes, it is easiest for decision-makers to postpone all action on an issue. For issues like sea level rise, it is very hard and costly to plan and execute proper policies and mitigation strategies. While working for the Ministry of Agriculture, Nature and Food Quality, Han de Groot heard rumours about so-called "crisis teams" within the government that were frowned upon

for trying to put off actions on certain issues hoping that a next minister would solve it. Usually, several lawyers would be members of these teams, who would seek grey areas of the law to justify the inaction. This might have been how the flaws in the Dutch Nitrogen policy got postponed up to the point of escalation.

This strategy of putting off issues until they really no longer can be ignored, creates a tricky political situation: at some point, these issues come to light accompanied with the knowledge that the government knew about these issues for a long time and did not act. At that point, an issue will have reached such a dramatic condition that immediate action is required for it not to turn into a crisis. However, governmental organisations are unprepared and there is a huge political pressure to 'make the right decisions', and the situation becomes unmanageable. The Dutch Nitrogen Crisis was born in a similar fashion.

5.3.2 Key characteristics of the Crisis

This crisis is memorable for various reasons. For example, for the fact that agriculture and building permits were halted, which temporarily froze the Dutch economy. Or for the farmers' protests of almost militant proportions. However, there is another, lesser covered aspect to the Dutch Nitrogen Crisis which is noteworthy: the birth of highly influential new organisations replacing the function of older, established membership organisations on social media. For example, traditionally the Land- en Tuinbouworganisatie (Eng.: Agricultural and Horticultural Organization, abbr.: LTO) was the organisation representing Dutch farmers and entrepreneurs in the agricultural sector. With over 35.000 members and connections to various governmental organisations, the LTO officially represented Dutch farmers. However, during the Nitrogen Crisis many Dutch farmers got so upset with policy-making and governmental institutions, that many turned their backs to the LTO for "being run by former politicians" and not representing the needs of the Dutch farmers" (FDF Board, 2021). Somehow, as a result of this unrest, two brand new farmers' organisations were born: Farmers Defence Force (FDF) and Agractie. Both these organisations gained many followers in a short time, and they were the organisers of the large-scale farmers protests that took place on the Malieveld in the Hague.

FDF and Agractie mobilised numerous farmers in a very short time, and by doing so suddenly became key actors in the political arena. Carola Schouten, minister of Agriculture, Nature and Food Quality, scheduled meetings with both organisations to hear their input. However, at least one meeting with FDF was cancelled out of concern for the Minister's safety, because of threatening tweets posted by FDF members. Of the two organisations, FDF is known for formulating more extreme opinions. For example, in December 2019, Mark van den Oever, president of FDF at the time, caused large commotion when he compared the Nitrogen policies in the Netherlands with the holocaust (Driessen, 2019).

5.3.3 How Twitter discourse analysis can help decision-makers

During a crisis like the Nitrogen Crisis, politicians are under high pressure to come up with solutions fast. However, in the modern digital age, politicians' actions are immediately visible to the public, either through news media or social media. Ministers, like Carola Schouten, are responsible for a fast solution, but when they do or say something one day, it gets criticised on the media the same day, and other politicians will pick this up and ask questions about her actions the next day in the House of Parliament. It makes it very hard for the ministers to decide what to do, or whom to talk to. Therefore, it would be helpful for them to get insights into the support base of various stakeholders and various plans. This is helpful in two ways. First, as we saw in the Nitrogen Crisis, new stakeholders can arise online and gain support in a short time. It is important that these news stakeholders are on the radar of the decision-makers as soon as possible, so meetings can be held soon with stakeholders with high support bases and these key stakeholders do not feel unseen or ignored. Secondly, it is also important to note when traditionally important stakeholders, like the LTO, lose their support base. This helps decision-makers prepare for meetings and negotiations with all stakeholders. Additionally, it is interesting for decision-makers to know about the reach of different stakeholders. How many people read their messages? How many are reacting to them, and retweeting them? Finally, about sentiment analysis, De Groot says that knowing about sentiments on Twitter can be helpful, but only if it can be linked to a *who:* who is sharing content of what sentiment? Does it differ per group? Finally, De Groot notes that there are also important stakeholders, like Johan Vollenbroek. Johan Vollenbroek does not have a social media presence, but has played a key role before and during the Nitrogen Crisis (Hakkenes, 2019). In 2015, when the PAS was announced, Johan Vollenbroek and his environmental organisation objected it (Hakkenes, 2019). When these objections did not lead to a change in the policy, he, together with a few others, sued the Dutch government for not adhering to European nitrogen emission rules. He won the trial, and the Nitrogen Crisis was born. De Groot points out that for decision-makers to only focus on social media risks not gaining awareness of the presence and influence of stakeholders like Johan Vollenbroek.

5.3.4 Important events during the Nitrogen Crisis

Table 5.1 shows the identified important events during the Nitrogen Crisis with brief descriptions, based on the interviews and on news sources. These events are later plotted in the sentiment analysis and topic modelling results, in section 5.4.2: Sentiment analysis results and section 5.5.2: Evaluation: Choosing number of words per topic for Jaccard similarity, to look for co-occurrences of these events and patterns in sentiment analysis and topic modelling results.

Table 5.1 | Major events during the Nitrogen Crisis.

	DATE	EVENT	DESCRIPTION	SOURCES
2018	7th November 2018	CJEU rejects the PAS	The CJEU rules that the licenses granted through the PAS to businesses that will increase nitrogen emissions are against European nature legislation	(NOS, 2019a)
2019	29th May 2019	Dutch Council of State	In line with the CJEU ruling, the Dutch Council of State rejects the PAS. All pending license applications are put on hold.	(Raad van State, 2019; NOS, 2019a)
	4th October 2019	First proposal new Nitrogen policy	The Dutch Cabinet presents an initial proposal for nitrogen reduction. This includes a lower highway speed limit on roads close to Natura2000 areas, and buy-out of livestock farms.	(NOS, 2019a; NOS, 2019b)
	14th-17th October 2019	Large scale farmer protests	Farmers raise demonstrations nationwide. The most impressive demonstrations take place in the Hague, though many farmers travel across the country to protest in front of various Provincial Government Buildings.	(Klumpenaar & Van Laarhoven, 2019)
	25th October 2019	Farmer protests in North Brabant	Farmers protest the Nitrogen policy and, this time, also the planned policies of the Dutch dairy sector. They block the entrance to the head office of Friesland Campina, the biggest Dutch dairy firm.	(Schelfaut, 2019a)
	18th December 2019	Farmer protests	Farmers blocked highways and supermarket distribution centers, even though a Dutch judge had forbidden it.	(RTL Nieuws, 2019)
2020	18th and 19th of February, 2020	Farmer protests	More farmer protests, some fines are handed out to farmers entering the highways with their tractor. In the Hague a few hundred people demonstrate.	(RTL Nieuws, 2020)

	8th July 2020	Minister Schouten cancels her visit to Zeeland due to security risks	Carola Schouten, Minister of Agriculture, Nature and Food Quality, cancels her visit to the province Zeeland on the advice of the police of Zeeland. The police reported that farmers had tracked the location of her destination and had formed a large group on the location, awaiting her.	(Wijnants, 2020) H. de Groot, personal communication, May 26, 2021)
	October 2020	Farmer show up at home of politician	Farmers show up late at night at the house of Rob Jetten, a D66 politician who wants to reduce the livestock in the Netherlands. Jetten is in quarantaine because of COVID, and the farmers bring him a package of food with, among others, meat, even though Jetten is a vegetarian. Jetten reports he thinks this gesture goes too far: intimidation wrapped in a nice gesture.	(Kos, 2020)
	November 2020	Farmer protests	Farmers go to the Malieveld with their tractors to protest the Nitrogen policy.	(Omroep West, 2020)
	December 2020	Farmer protests	Farmers protest both the Nitrogen policy and the low prices Dutch supermarkets ask for their products. Various supermarket distribution centres were barricaded, and the protests got a lot of criticism for going too far and the protesters were criticised for not following COVID regulations	(NOS, 2020)
2021	7th July 2021	Farmer protests	Members of Agractie came to the Hague once again to protest on the Malieveld, but this time the atmosphere was more friendly. FDF protested on multiple other locations in the Netherlands.	(Eijsink, 2021; NOS, 2021)

5.4 Sentiment analysis

In this section, first the two sentiment analysis tools, SentiStrength and Pattern, are evaluated and compared to the sentiment scored by the three annotators in section 5.4.1. The tool that scores most similar to the annotators, is selected as the best tool. Lastly, the sentiment analysis results with this tool are presented in section 5.4.2.

First, to give a few examples of tweets and the scoring behavior of SS and Pattern, Table 5.2 shows a selection of 4 tweets and their sentiment scores, translated in English in *Italic*. Remember the scales: SS scores from -1 to -5 for negativity, and from 1 to 5 for positivity. Pattern returns just one value, between -1 and 1, 0 being neutral.

Tweet 1 is scored negatively by both SS and Pattern, getting the lowest possible score of -1 by Pattern. The tweet is a complaint and contains words like "absurd" and "ridiculous". Tweet 2 is scored positive by both tools, and is an exclamation of how "beautiful" the farmers protest is. Tweet 3 is scored as neutral, getting a 0 from Pattern and the minimal scores from SS: -1 and 1. This tweet seems an enumeration of facts. Tweet 4, is scores lightly positive by both sentiment tools, and contains a call for 'respectful' help. On the last Tweet, tweet 5, the two sentiment analysis tools disagree: Pattern scores it slightly positive with 0.275, while SS scores it convincingly negative with negative score of -3, while a neutral positive score of 1.

Table 5.2 | Examples of tweets and their sentiment scores. Positive scores are coloured on a blue spectrum, while negative scores are coloured on a red spectrum. The higher the score, the more intense the colour. Neutral scores are white. Translation by Google Translate.

	TWEET	PATTERN	SS POSITIVE	SS NEGATIVE
1	@The_realist_31 Inzet van defensie bij de boerenprotesten was absurd en belachelijk. En dan is het in die context NOG belachelijker dat defensie nu niet ingezet is geweest. *Translation: @The_realist_31 Deployment of defence in the peasant protests was absurd and ridiculous. And then in that context it is EVEN more ridiculous that the defence has not been deployed now*	-1	1	-3
2	Prachtige boerenprotest in Nederland! *https://t.co/QpT39f05Gw* *Translation: Beautiful farmers protest in the Netherlands! https://t.co/QpT39f05Gw*	1	4	-1
3	Van de Nederlandse uitstoot bestaat 60% uit ammoniak (NH3) en 40% uit stikstofoxiden (NOx). De landbouw zorgt voor 61% van de stikstofuitstoot (door mest, maar ook uit kassen en door	0	1	-1

	landbouwvoertuigen), het wegverkeer voor 15%. De uitstoot wordt ook wel stikstof-emissie genoemd. *Translation: 60% of Dutch emissions consist of ammonia (NH3) and 40% of nitrogen oxides (NOx). Agriculture is responsible for 61% of nitrogen emissions (from manure, but also from greenhouses and agricultural vehicles), road traffic for 15%. The emissions are also called nitrogen emissions*			
4	Net voor dit #boerenprotest de A29 bij de Haringvlietbrug op. Alle respect en blijf ons helpen door te innoveren met respect voor natuur en dieren. *https://t.co/7GoRqwhJTw* *Translation: Just before this #farmersprotest on the A29 at the Haringvlietbrug. All respect and keep helping us by innovating with respect for nature and animals. https://t.co/7GoRqwhJTw*	0.55	3	-1
5	Goed nieuws in informatiebrief mbt boerenprotest 16 oktober jl. De samenwerking tussen hulpdiensten en boeren liep voorspoedig en geen enkele vorm schade in en om Wassenaar. *https://t.co/rGNyAGVRWP* *Translation: Good news in information letter regarding farmers' protest on 16 October.* *The cooperation between emergency services and farmers went smoothly and no damage was done in and around Wassenaar. https://t.co/rGNyAGVRWP*	0.275	1	-3

5.4.1 Evaluation

In order to pick the best sentiment analysis tool, this section describes the outcomes of the comparison between the sentiment scores by the 3 annotators, and the scores by the two tools: SS and Pattern.

First, it is useful to look at how much the sentiment scores of the annotators overlap with each other. Table 5.3 shows the ICC2 of the annotated tweets by the three annotators. The ICC2 of 0.776 is high, meaning there is a significant correlation between the annotators and there is high agreement on how sentiment in tweets is scored.

Table 5.3 | Intraclass Correlation Coefficient 2 of annotated tweets.

TYPE	DESCRIPTION	ICC
ICC2	Single random raters	0.776

Comparison annotation and sentiment analysis techniques

Now it is clear that the annotators relatively agree on how to score sentiment in tweets, the mean of their scores is calculated and compared with the scores of each individual sentiment analysis tool. Table 5.4 shows the PCC between the two sentiment analysis methods and the mean of the annotation per tweet. SS has a higher correlation with the annotated tweets, with a value above 0.5 being considered high. Because of this, SS will be used for sentiment analysis of the tweet dataset in the continuation of this research.

Table 5.4 | PCC between both sentiment analysis techniques and the mean annotated sentiment scores.

	SS & ANNOTATORS	PATTERN & ANNOTATORS
PCC	0.540259055	0.328589

Assigning tweets to the positive or negative dataset using SS scores

In order to compare the volume and the average sentiment of negative tweets to the volume and average sentiment of positive tweets, the dataset is split into two datasets per sentiment analysis method: one positive and one negative. For SS this is tricky, as it scores sentiment on two scales. Assigning tweets to a sentiment based on their SS score is done as follows: the highest of the two sentiment scores will decide whether the tweet is considered positive or negative. To illustrate, if a tweet scores 3 (on the positivity scale) and -1 (on the negativity scale), the tweet will be considered positive and its sentiment score will be 3. All the tweets with an equal positive and negative score are considered emotionally neutral: the intensity of the opposing sentiment scores phase each other out. This results in the deletion of 116855 (40%) tweets, out of the total 292245 tweets. In the end, the positive dataset consists of 39398 tweets, while the negative dataset consists of 135992 tweets. This means the negative dataset is roughly 3.5 times bigger than the positive dataset.

5.4.2 Sentiment analysis results

Figure 5.1 on the next page shows, on top, in Figure 5.1.1, the number of tweets per week as a reference. The events from Table 5.1 are shown as tagged vertical lines. In the plots below that, the results of sentiment analysis are depicted. The second graph 'Weekly number of tweets per sentiment', shows the number of weekly tweets *per sentiment* (Figure 5.1.2). On the third graph, the weekly mean sentiment is plotted (Figure 5.1.3). To see the difference in average sentiment more clearly, Figure 5.1.4 shows the difference between the average positive and negative sentiment score: the plot is coloured orange when the average negative sentiment is higher than the average positive sentiment, and blue vice versa. Below that, the same weekly mean is plotted, but with the standard deviation (std) plotted as a bar at each datapoint (Figure 5.1.5).

Interestingly, what becomes evident when looking at the total volume of tweets in Figure 5.1.1 is that there is some relation between a peak in tweet volume and the identified events. During 7 out of the 11 identified events there is a peak in the number of tweets that week. Figure 5.1.2 shows that, in terms of volume of tweets per sentiment, there are more negative than positive tweets at all times, often more than double. Also, Figure 5.1.3 and Figure 5.1.4 show that tweets are more negative than positive in mean sentiment, almost at all times. The only exception is mid August, when the volume of tweets was still low, which does not overlap with one of the identified events and so cannot be explained by them. It follows that, according to the sentiment analysis results, tweets are more negative than positive throughout (nearly) the whole duration of the crisis, both in volume and in mean weekly sentiment. Figure 5.1.5 on the bottom, shows the same weekly mean sentiment as Figure 5.1.3, but shows the std of that mean as a bar per datapoint. Figure 5.1.5 shows that there is high of dispersion in the tweet sentiment score for both positive tweets and negative tweets. The mean std throughout the whole dataset is 0.66 for positive tweets and 0.64 for negative tweets, on a sentiment scale from 2 to 4 (remember that tweets that tweets with neutral sentiment were deleted).

The results of the Pattern sentiment analysis can be found in Appendix E: Pattern results. As is expected after the comparison with annotators, the results from Pattern are different from SS: the pattern that is seen in the SS results of tweets being more negative both in sentiment and in volume, is not seen in the Pattern results.

Sentiment Analysis

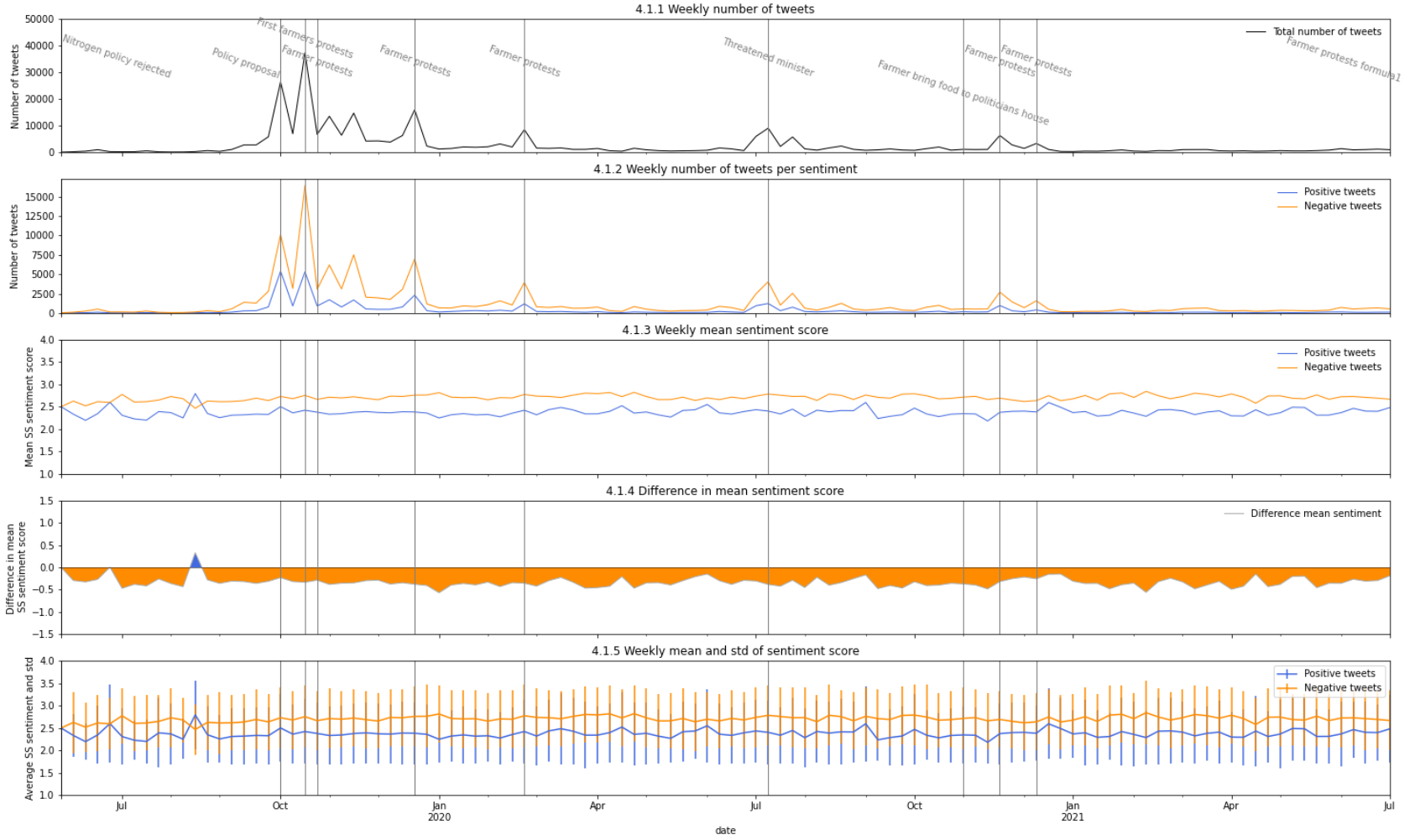


Figure 5.1 | Sentiment analysis results. From top to bottom: 4.1.1. weekly volume of tweets, 4.1.2. weekly volume of tweets per sentiment, 4.1.3. Weekly mean sentiment score, 4.1.4. difference in mean sentiment score difference in mean sentiment score difference in mean sentiment score difference in mean sentiment score difference in mean sentiment score and 4.1.5 Weekly mean and std.

5.5 Topic modelling

Now the sentiment analysis results are discussed, this following section covers the results of the application of the other NLP method: topic modelling. First, section 5.5.1 covers the outcomes of the grid search, which leads to the choice of *number of topics* and *time window* to slice the data per sentiment. In the next section, section 5.5.2, the number of words to represent a topic to calculate the mean Jaccard similarity is chosen, by plotting the mean Jaccard similarity for topic models with 10, 20, 30 and 40 words per topics. Using the settings chosen in the previous two sections, section 5.5.3 shows two topic models, one with a high topic coherence score and one with a low topic coherence, to show what a topic model looks like and what the differences can be between models with different coherence scores. Lastly, with the parameter settings chosen in sections 5.5.1 and 5.5.2, section 5.5.3 shows the topic modelling results for the complete dataset, separately per sentiment. Here, the change over time in topics discussed during the Nitrogen Crisis is plotted per dataset.

5.5.1 Evaluation: Grid search for choosing topic modelling settings

Figure 5.2 on the next page shows the average coherence scores of all topic models per combination of time window and number of topics (num_topics) in three heat maps. Note that every cell in the heat map represents an entire sequence of topic models, that are trained on dataset chunks of either 1, 2, 3 or 4 weeks worth of tweets. On the left, Figure 5.2 shows that the highest average coherence scores belong to topic models with a time window of 7 days, and 16 topics. In the middle, Figure 5.2 shows the highest average coherence for the positive dataset with a time window of 7 days, and 4 topics. For the negative tweet subset, the highest average coherence score is achieved through topic models with a time window of 7 days, and 16 topics. These values will be chosen as input settings for the topic models in the next sections. Note that the scales are not the same: ranging between 0.443 and 0.479, the positive dataset generates topic models with an average coherence that is higher than the negative and the total dataset, which both range between 0.3 and 0.385.

In Figure 5.3, the standard deviation of the three grid searches is plotted. On all three heatmaps, the whole upper area is very dark, which means the stds are relatively low when the average coherences are high in Figure 5.2. It follows that among the sequence of topic models with settings that lead to higher mean coherence, the coherence values fluctuate less per topic model. Although the range of mean coherence scores is higher for the positive dataset, the std range is slightly higher. This means that the coherence scores of topic models in the positive dataset are slightly more dispersed than in the other two datasets.

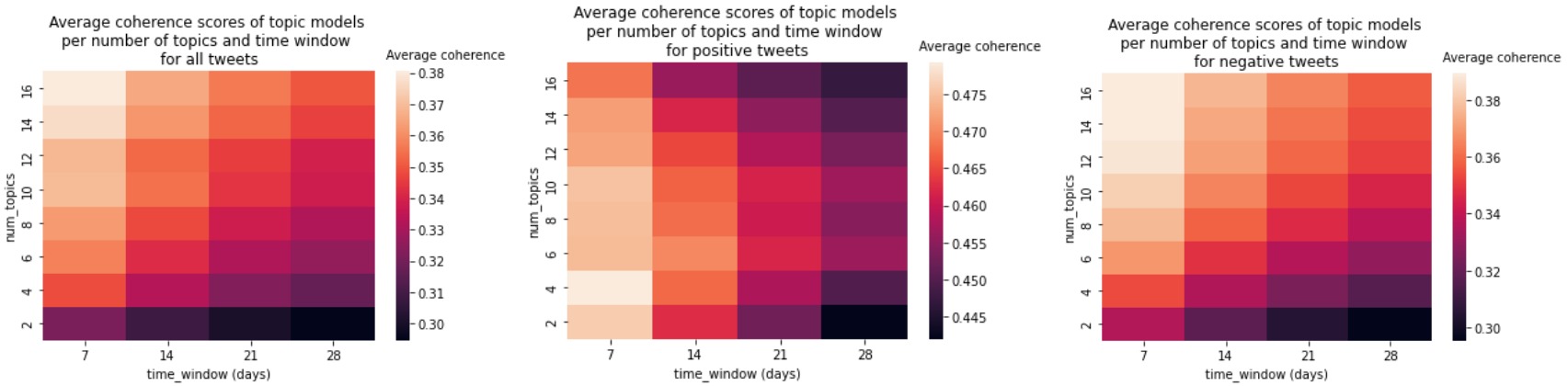


Figure 5.2 | From left to right, the average coherence scores of topic models for all tweets, only the positive tweets and only the negative tweets. The brighter the colour, the higher the average coherence score.

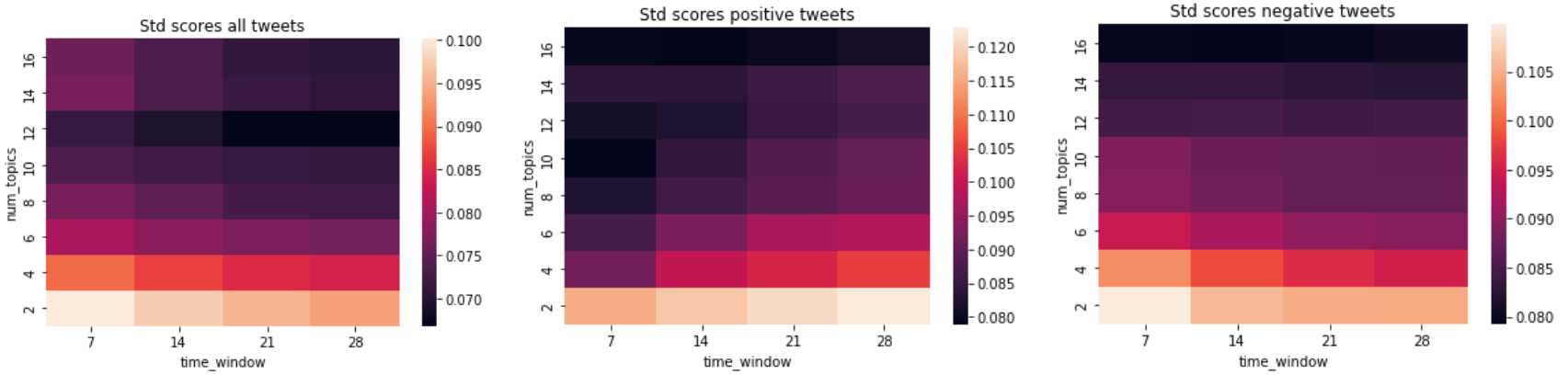


Figure 5.3 | From left to right, the stdev of the coherence scores of topic models for all tweets, only the positive tweets and only the negative tweets. The brighter the colour, the higher the stdev.

5.5.2 Evaluation: Choosing number of words per topic for Jaccard similarity

Figure 5.4 shows the mean Jaccard similarity of subsequent topic models for the complete tweet dataset. In this case, by 'subsequent models' two models are meant with either 1, 2, 3 or 4 weeks in between. This way, not only short-term differences in topics discussed can be shown, but also changes that happen on a longer term. For example, when the mean Jaccard similarity is consistently high over a period of time, on a week-to-week basis very little changes in the topics that are discussed. However, when within that period of time, the mean Jaccard between two topic models 4 weeks apart is lower, it means that topic discussed are changing little week by week, but are definitely changing.

Figure 5.5 shows the stds of the Jaccard similarities from Figure 5.4. In Figure 5.4, it is shown that the more words are included in a topic before computing the mean Jaccard similarity, the less variance is shown in the plot. Figure 5.5 confirms this, as the more words per topic, the lower the stds. Figure 5.5 also shows that if more words are considered per topic, this does not mean that the mean Jaccard similarity goes up. One might expect that when the words with less weight are considered and the total number of words considered goes up, that the overlap of words between topics might get relatively higher, but this is not the case. What does happen when more than 10 words are chosen to represent a topic, shown in Figure 5.5, is that the std of the Jaccard similarity is slightly lower on average. In the Methods chapter, section 4.6.2, five matrices were plotted with the Jaccard similarities between two topic models (Figure 4.3). Each matrix showed the Jaccard similarities with a different number of words per topic, and the more topics were included per topic, the more uniform the matrix was coloured (which means a lower std). This means the more words are chosen to represent a topic, the better the mean Jaccard similarity represents the distance between two topic models, as the std is lower. However, choosing more words will also mean that words that represent the topic less than the top 10 are included, and that they are given the same weight as the top 10 words in the Jaccard similarity formula. Changes in topics have less effect on the mean Jaccard similarity and therefore the mean Jaccard similarity changes less, as is seen in Figure 5.4. Because the aim of this research is to provide results that are meaningful to decision-makers, the number of words per topic is chosen that leads to the most change in mean Jaccard similarity over time: 10 words per topic.

Finally, Figure 5.5 shows that the std does not vary much based on whether the mean Jaccard similarity is calculated between topic models with 1, 2, 3 or 4 weeks in between. Therefore, the number of weeks chosen in between topic models does not influence how representative the mean Jaccard similarity is if the similarities between each individual topic of the two topic models. Plotting all 4 mean Jaccard similarities in one plot creates a less easily interpretable plot (too many lines), and the mean Jaccard similarities of topic models of 2 and 3 weeks apart seem to follow a similar pattern. For these reasons, in the plots with all topic modelling results in section 5.5.4, the choice is made to only plot the mean Jaccard similarity of topic models that are either 1 or 4 weeks apart. Similar plots as Figure 5.4 and Figure 5.5 for the positive and negative dataset are shown in Appendix F: Jaccard distance depending for number of words per topic = 10 to 40, and comparing various weeks.

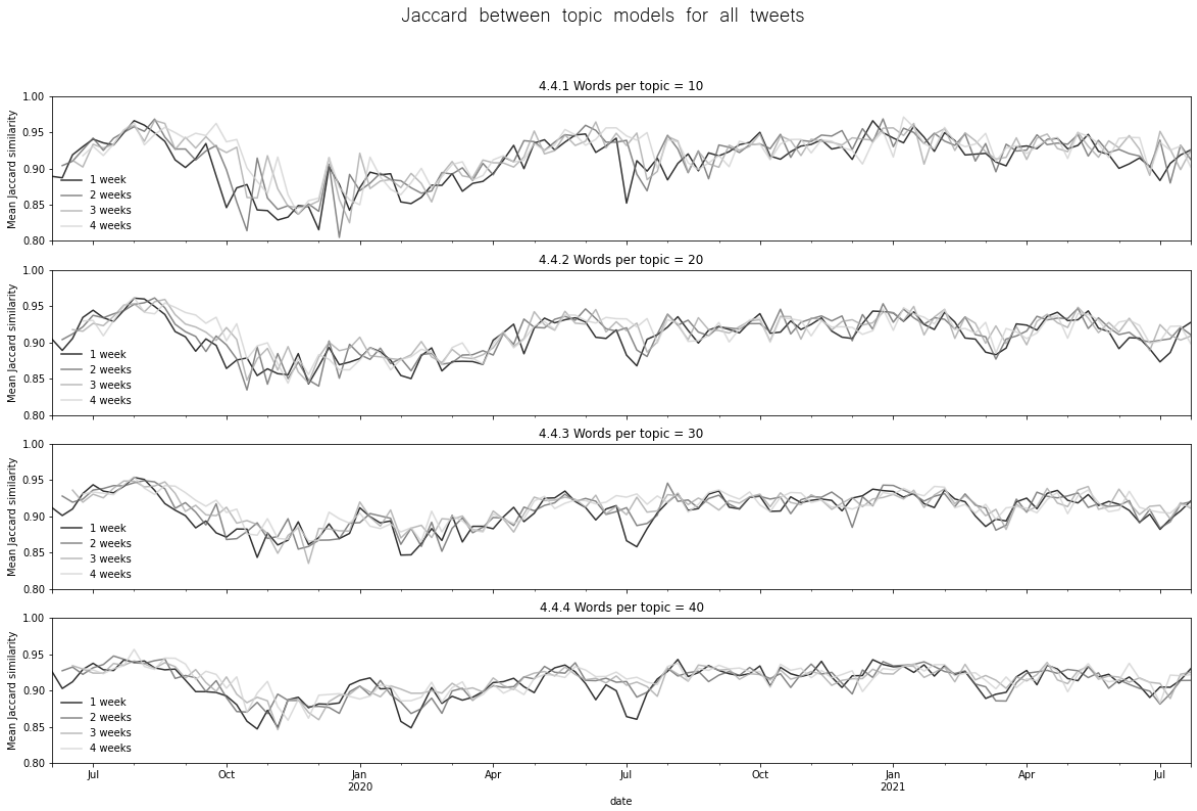


Figure 5.4 | Mean Jaccard similarity for all tweets, with either 1, 2, 3 or 4 weeks time in between two topic models. Each plot shows the mean Jaccard similarity of two subsequent topic models when topics are represented by either 10, 20, 30 or 40 words.

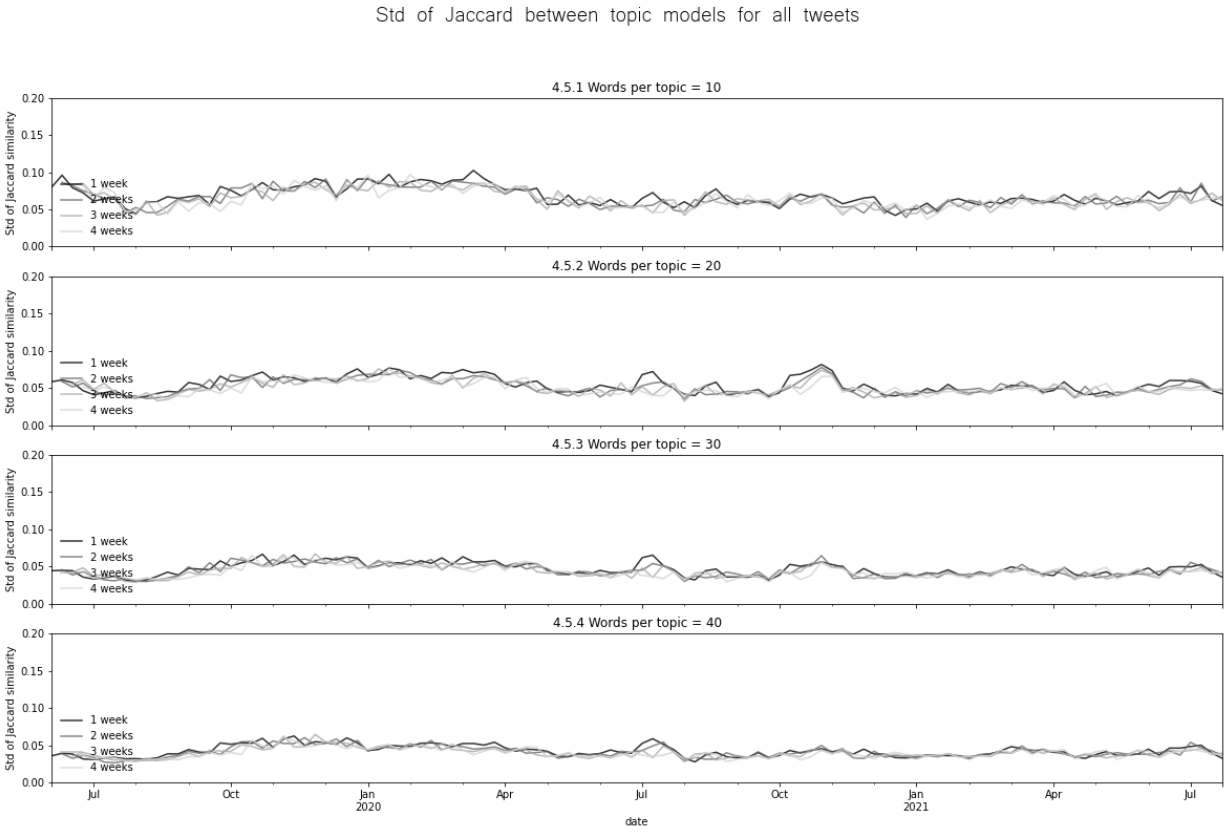


Figure 5.5 | Std of Jaccard similarity for all tweets, with either 1, 2, 3 or 4 weeks time in between two topic models. Each plot shows the st dev of theJaccard similarity of two subsequent topic models when topics are represented by either 10, 20, 30 or 40 words.

5.5.3 Examples of topic models

Now the number of topics, time window and number of words to represent a topic are chosen, the topic models for each sentiment can be plotted. Before plotting the mean Jaccard similarities for each sentiment in the next section, this section shows two examples of topic models are shown. It has been stated that research aims not to focus on the time-consuming evaluation of the content of topic models but rather on the changes in topics discussed over time. However, to give the reader an idea of what the many topic models with varying coherence scores in the next section look like and allow them to inspect and compare the topics they contain, Figure 5.6 and Figure 5.7 shows word clouds of the topic models with the highest and the lowest coherence score from the dataset containing all tweets. The font size of words in topics represents their weight on the topic distribution. To allow non-Dutch speaking readers to interpret the topic models, the translation of the Dutch words is given in brackets after each word (translated by Google Translate). These translations are quite accurate as far as it is possible to eyeball this without the context of each word. To illustrate the difficulty with this, in topic 1 in Figure 5.6 the Dutch word "weer" is translated as "again", while the word could also mean "weather", and "waar" is translated as "true", while it could also mean "where". Two examples of the translation being inaccurate is "mens" being translated to "man" instead of "person", and "natura", the name of Dutch protected natural areas, gets unjustly translated to "kind". Regardless, the translation seems accurate for the most part.

To evaluate the interpretability of the topic models, I tried to label each topic per topic model and included that in Table 5.5 for Figure 5.6 and in Table 5.6 for Figure 5.7. It is recommended to have a go at labeling topics as reader before looking at the tables with my interpretations. It was interesting to see that I managed to label more topics in the topic model with higher coherence (a question mark shows the topics that I was unable to label). Differences that van be seen between the two word clouds is that Figure 5.6, with the higher coherence, has bigger fonts (higher weights per word) and more variance between font sizes than Figure 5.7. Also, I found it harder two distinguish between topics in Figure 5.7 as the topics contained more of the same words. For example, the word "stikstofcrisi" (lemmatized version of stikstofcrisis, which means Nitrogen Crisis) is present in the top 10 words of all 16 topics in Figure 5.7.

It is interesting to see that quite a few names of Twitter account are present in the word clouds, like 'jinek_rtl' (a talkshow)[5], 'ftm_nl' (a news website)[6] and 'horecaned' (a twitter account with memes and news about the Dutch Hospitality Industry)[7].

In addition to the word clouds shown in this section, four word clouds of the topic models with the highest and lowest coherence for both the positive and negative sentiment are found in Appendix G.

[5] https://twitter.com/jinek_rtl
[6] https://twitter.com/ftm_nl
[7] https://twitter.com/horecaned?lang=en

Figure 5.6 | Word Cloud with the highest coherence score from the topic models of the full tweet dataset.

Table 5.5 | Label per topic: an interpretation of the topics in Figure 5.6 by the writer.

TOPIC 1	TOPIC 2	TOPIC 3	TOPIC 4
?	Grab the polluter!	The different nitrogen policies	Conservatism
TOPIC 5	TOPIC 6	TOPIC 7	TOPIC 8
Why focus on Agriculture, when human emit a lot of Nitrogen with, for example, flying?	Negotiations and scepticism	Farmer protest at municipality of Oud-Ijsselstreek	Low turnout for Agractie protests
TOPIC 9	TOPIC 10	TOPIC 11	TOPIC 12
	What do Nitrogen emissions really do to natural areas?	News media	Nitrogen crisis: agriculture vs. construction
TOPIC 13	TOPIC 14	TOPIC 15	TOPIC 16
Visibility of agractie farmer protests	?	Who is responsible for leading and supervising change to prevent environmental disasters	How facts and calculations are used within the agriculture nitrogen emission debate

Figure 5.7 | Word Cloud with the lowest coherence score from the topic models of the full tweet dataset.

Table 5.6 | Label per topic: an interpretation of the topics in Figure 5.7 by the writer.

TOPIC 1	TOPIC 2	TOPIC 3	TOPIC 4
	The Cabinet comes with a solution that costs/saves a few billion		
TOPIC 5	TOPIC 6	TOPIC 7	TOPIC 8
A column written by one Arno Wellens		Politics and Brussels deciding on policies	
TOPIC 9	TOPIC 10	TOPIC 11	TOPIC 12
		The Nitrogen crisis and farmers protests	
TOPIC 13	TOPIC 14	TOPIC 15	TOPIC 16
	Carless Sunday as a solution to the Nitrogen Crisis	Politics and politicians	

5.5.4 Topic modelling results

In this last section of the Results chapter, the mean Jaccard similarity (MJS) is plotted in Figure 5.8. Just like with the sentiment analysis results, Figure 5.8.1 shows the weekly volume of tweets as reference. First, the y-axis of Figure 5.8.2, 4.8.3 and 4.8.4 shows a range between 0.75 and 1. Looking back at Figure 4.3 on the left (because topics are represented by 10 words) in section 4.5.3, it follows that when the MJS is 0.7, the number of overlapping words is 8.6, and when the MJS is 1, the number of overlapping words is 10. Because Figure 5.8 shows the *mean* Jaccard similarity, this means that on average the number of words that overlap per topic vary between 8.6 and 10 for all topic models. If on average between as many as 8.6 and 10 words overlap per pair of topic models, this means topics models only change slowly.

Secondly, the behaviour of the MJS of topic models with 1 week distance (1 week MJS) is similar for the three datasets. For all datasets, the MJS decreases just before the 'Policy proposal' event in early October 2019, which means that new topics are introduced, then drops slightly again during the farmers protests in December 2019, and then drops again right before the 'Threatened minster' event in July 2020. These events, or the anticipation of these events, seem to spark a slight change in what topics are discussed in all three datasets. However, the change might also be caused by the increase of tweets around these events, which could influence the outcome of the topic models. Other than those three events, there is no visible connection between the presence of an event and changes in the MJS. Therefore, none of the MJS values of the datasets can be used as a marker for these events. Yet again, the volume of tweets looks like a better marker.

The behaviour of the MJS of topic models with 4 weeks distance (4 week MJS) varies more between the datasets. When looking at these lighter lines, it is interesting to examine whether the 4 week MJS is higher or lower than the 1 week MJS, and whether a drop in the 1 week MJS is followed by a drop in the 4 week MJS. When a drop takes place in the 1 week MJS and goes back up the week after, then in the week of the drop new words are introduced in the topics, that after that week will probably keep being used, because the MJS goes back up again. If this drop is also visible 4 weeks after the drop in the 4 week MJS, then the words introduced 4 weeks ago continued being used up to 4 weeks after the drop, because the similarity between this week and 4 weeks ago is low. This can be seen for the drop in October in Figure 5.8.3 (positive dataset).

However, if no such drop is visible after 4 weeks, this means that the topic discussed after the drop changed back over the weeks to the topics discussed before the drop. This can be seen in Figure 5.8.3 as well, when the drop in June 2020 does not results in a drop 4 weeks after.

In terms of differences between the three sentiment datasets, Figure 5.8 shows that MJS of the positive dataset (blue) is less consistent than the full dataset (**black**) and negative dataset (orange). This could be because for the positive dataset, it varies more per week whether Twitter users change the topics they discuss or discuss similar topics to last week. However, this could also be explained by the fact the topic models of the positive dataset contain only 4 topics, while

the other each contain 16 topics per topic model (this was determined in section 5.5.1). Therefore, the MJS in Figure 5.8.3 will change more with a few changes in words per pair of topics than the MJS of the other datasets will.

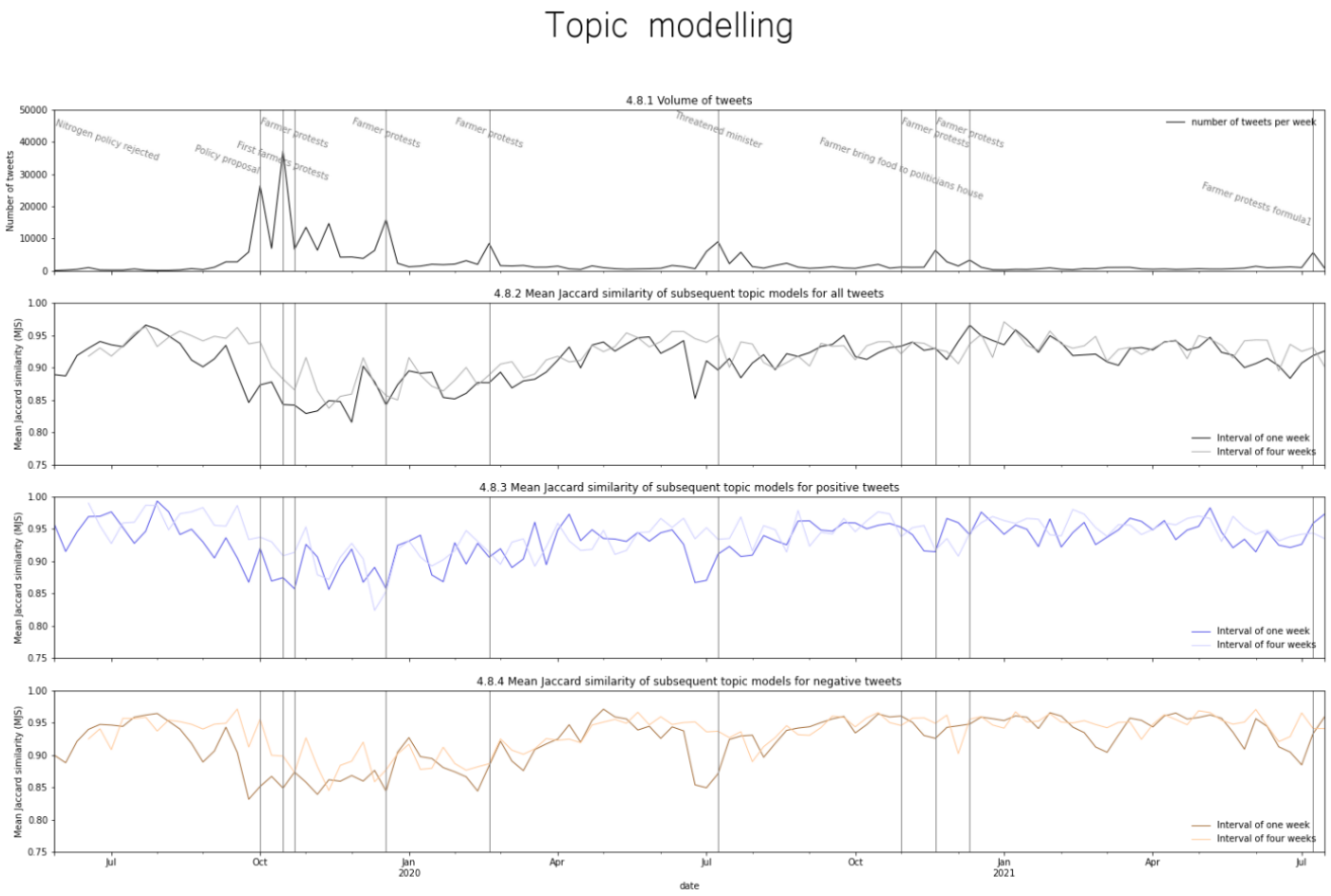


Figure 5.8 | Mean Jaccard similarity for topic models of 1 and 4 weeks distant per sentiment dataset.

Key Findings of Chapter 5: Results

1. 12 events were identified based on the interviews and desk research
2. *SentiStrength* performs *better than Pattern* and is chosen as sentiment analysis tool
3. Tweets are *more negative than positive* throughout the whole duration of the crisis (except during 1 week) both in volume and in average sentiment — SQ1
4. The *window size* that results in the best topic models is of *7 days*
5. The number of topics that results in the best topic models varies per sentiment: — SQ3
   for all tweets: number of topics = 16
   for positive tweets: number of topics = 4
   for negative tweets: number of topics = 16
6. Having a topic represented by the 10 most important words leads to the most variation between subsequent topic models over time
7. It is possible to identify human interpretable topics based on the topic model with the highest scoring topic coherence
8. The range of difference in the mean Jaccard similarity is small [0.75, 1]
9. A spike in the weekly volume of tweets seems the best indicator for an event, not sentiment analysis or topic modelling results

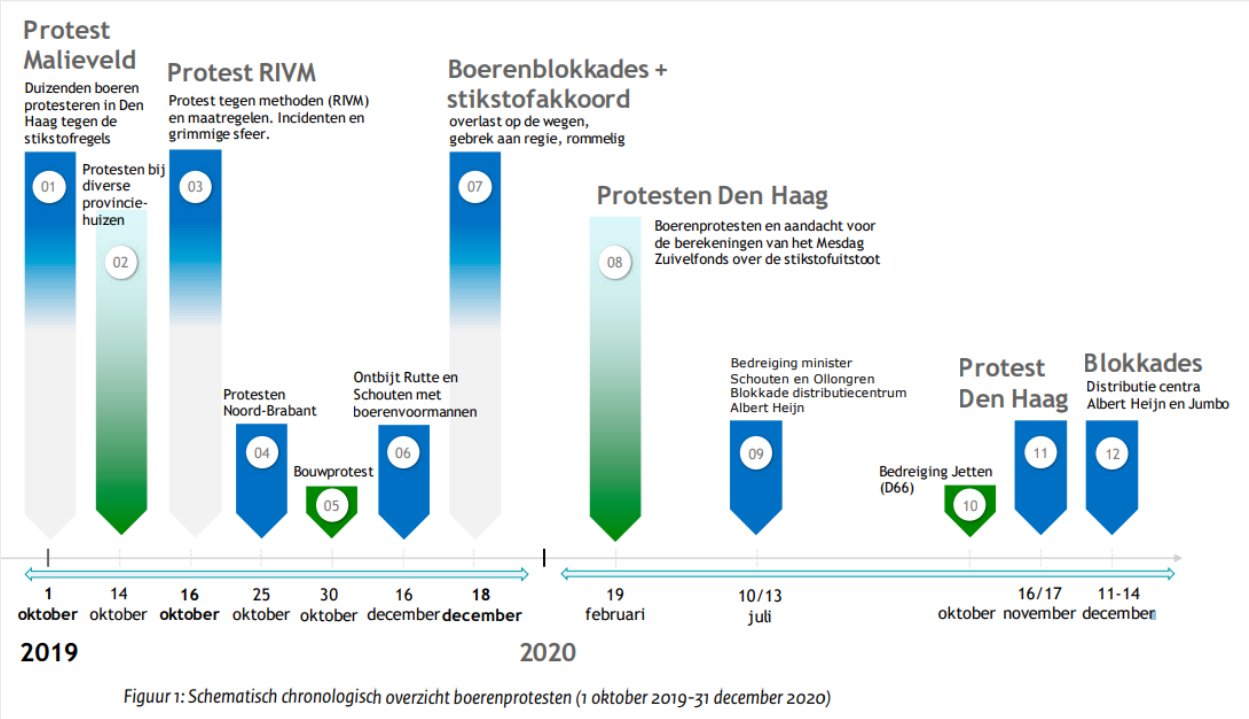
6 Discussion, limitations and future research

This research aims to apply sentiment analysis and topic modelling to Twitter data on the Dutch Nitrogen Crisis in order to provide insights for research makers. Sentiment analysis on its own does not provide insights in the content of tweets and while topic modeling does, interpreting these topic models is an ambiguous and time-consuming task. Therefore, these two NLP methods are combined in this research to see if together they lead to results that can be insightful and less time-consuming to decision-makers. In this chapter, the results are discussed in a larger context and interpreted. The meaning, importance and relevance of the results of this research are elaborated on. Then, the limitations of this research and recommendations for future research are formulated. This chapter follows the structure of discussing the key insights listed in the Results chapter on the previous page one by one. Lastly, this chapter is closed with a section on general overarching limitations of this study and recommendations for future work.

6.1 12 events were identified based on the interviews and desk research

In addition to the interviews, it was a Wikipedia page that provided the most comprehensive overview of events that took place during the Nitrogen Crisis as described in the Methods chapter (Wikipedia, n.d.). However, in the meantime, the Dutch Ministry of the Interior and Kingdom Relations published a media analysis of the farmers' protests, written by Lieuwe Kalkhoven (Kalkhoven, 2021). Just like this study, this report identifies key events during the Nitrogen Crisis and contains a (albeit short) social media analysis.

There is a high overlap between the main events identified in the report by Kalkhoven, shown in Figure 6.1, and the events identified in this research (Table 5.1). Ten out of the twelve events plotted in the figure from the Kalkhoven report were identified in this research, though two events in Figure 6.1 were combined into one event in this research. The two events that were not considered in this research which were identified by Kalkhoven were the construction worker protest and the breakfast meeting of PM Rutte and Minister Schouten with representatives of the farmers' organisations, both in December 2019.



Figuur 1: Schematisch chronologisch overzicht boerenprotesten (1 oktober 2019-31 december 2020)

Figure 6.1 | Schematic chronological overview of farmers' protests (1 October 2019-31 December 2020) (Kalkhoven, 2021).

6.2 SentiStrength performs better than Pattern and is chosen as sentiment analysis tool

SS was developed for short social media text while Pattern was based on book reviews, which could explain why SS scores, when applied to tweets, correlates more with the sentiment scores by the annotators than Pattern scores (De Smedt & Daelemans, 2012; Thelwall, Buckley, Paltoglou, & Kappas, 2010). Another explanation might be that Pattern was developed in 2012 and does not get updated anymore (De Smedt & Daelemans, 2012).

Another Master thesis that compared SS and Pattern was written by Lennart van Winzum for his Masters in Dutch studies (Van Winzum, 2021). In his comparison, Pattern turned out to correlate best with the annotations. However, Van Winzum's data comprised news article headlines, not social media text, which could explain this. Another difference between Van Winzum's approach and the one in this research is that here, only the most extreme SS is chosen to represent the sentiment, meaning if a tweet scores 4 on the positive scale and -2 on the negative, the final score is 4. Van Winzum took the sum of the two scales, in this example that would result in a tweet sentiment score of 4 + -2= 2. For future research, it can be seen if using Van Winzum´s

approach for combining the two SS sentiment scales, summing, leads to a higher correlation between the SS scores and the annotator sentiment scores.

In this research, the results of the two sentiment analysis tools were compared to the annotations of only three annotators. These annotators were all female, from similar social backgrounds and of similar age. To calculate the correlation between the sentiment scores of the three annotators, ICC2 was calculated. This calculation for ICC2 assumes the group of annotators is a random sample of a larger population (Shrout & Fleiss, 1979). This was not the case, as the three annotators were the writer of this report and two of her friends. To properly represent a larger population, the group of annotators should be larger and more diverse. Based on the annotations of a more representative group, there is more substantiation for the choice best sentiment analysis.

Even though SS correlates better with the sentiment scores of the annotators, a correlation of 0.540259055 (see Table 5.4) only shows a moderate correlation. Therefore, SS can be applied for sentiment analysis of the Nitrogen Crisis tweets, but a more accurate sentiment analysis tool would lead to more precise results. This can be done in several ways. First, instead of using a pre-developed tool, researchers could train their own sentiment analysis model. This can be done in a supervised or unsupervised manner (Pang, Lee, & Vaithyanathan, 2002). For supervised training of a sentiment analysis model, a large number of tweets needs to be annotated, the larger the better. Then, a classifier algorithm, e.g., Naive Bayes, maximum entropy classification, support vector machines, Decision Trees or Random Forest, is applied to train a sentiment analysis model based on the annotations (Pang, Lee, & Vaithyanathan, 2002; Rothfels & Tibshirani, 2010; Guia, Silva, & Bernardino, 2019). For an unsupervised method, the problem of lack of annotated data needs to be overcome before a classifier algorithm can be trained. One way to do this is by defining seed words, one for negative sentiment and one for positive sentiment, to calculate the semantic orientation of each document (Turney, 2002).

A different approach that could improve sentiment scoring accuracy but that does not require training a new sentiment analysis model, is to translate all tweets to English and apply a pre-trained English sentiment analysis method (Mohammad, Salameh, & Kiritchenko, 2016). English sentiment tools are generally of higher quality than sentiment analysis tools for languages that are less widely spoken (Dashtipour, et al., 2016). Also, various papers assessing the quality of automatic translators have found that the quality of translation suffices for applying sentiment analysis and retrieve reliable results, sometimes even better results than the sentiment tools in the original languages (Balahur & Turchi, 2012; Araujo, Pereira, & Benevenuto, 2016).

6.3 Tweets are more negative than positive throughout the whole duration of the crisis (except during 1 week) both in volume and in average sentiment

The higher negativity of the dataset could be explained by the overarching topic of the tweets: the Nitrogen *Crisis*. It is possible that during times of crisis Twitter users are more inclined to discuss the topic negatively than positively. The results suggest that not only the chance of a tweet being negative is higher than the tweet being positive, the intensity of the negativity is slightly higher on average than the intensity of the positivity. However, for both sentiments, the intensity of the sentiment does not vary much: at no point in time does the mean sentiment reach 3 or higher. On a scale from 2 to 5 (remember the tweets with the neutral score of -1 or 1 area not included in the plot), this variance is low.

The only time the mean positivity was higher than the mean negativity, in mid-August 2019, cannot be explained by the identified events. When searching online for events that might have taken place in August 2019, the only news article discusses how, because of the Nitrogen Crisis, in August, fewer building permits were granted (Doodeman, 2019). This does not sound like an event that would spark positivity, therefore by itself it cannot explain the increase in positivity. To research this further, it would be advised to look into the topic models of that week and find out what topics were discussed positively.

Interestingly, the previously mentioned media analysis of the farmers' protests published by the Dutch Ministry of the Interior and Kingdom Relations included an analysis for establishing the public support base for the farmers (Kalkhoven, 2021). The analysis concluded that the public support base was very high at the beginning of the crisis, but, as the actions of the farmers turned more and more radical, it lowered significantly. This particular conclusion was not based on social media posts but on surveys and news articles. However, it could be possible that when the public support for the farmers lowers, there would be an increase in negativity on Twitter. The sentiment analysis results do not depict this reduction in support base. For future research, it would be interesting to isolate the tweets that are specifically about the farmers' protests, and see if the volume and intensity of the negative tweets increases.

Various studies on positivity and negativity on Twitter conclude that negativity spreads much faster and broader than positivity (Schöne, Parkinson, & Goldenberg, 2021; Bellovary, Young, & Goldenberg, 2021; Jiménez-Zafra, Sáez-Castillo, Conde-Sánchez, & Martín-Valdivia, 2021). These studies look at the spread of sentiment of tweets from Twitter influencers or news media Twitter accounts, two of them in connection to an important political event. Although this study does not focus on the spread of sentiment, it would be interesting in future work to look what Twitter accounts had the highest reach on Twitter, and whether the sentiment of the posts of these accounts can be connected to the higher negativity during the Nitrogen Crisis.

Finally, it is interesting to note that the Pattern sentiment results (shown in *Appendix E: Pattern results*) were entirely more positive than negative in volume, and in mean sentiment roughly half

of the time. SentiStrength scores correlated more than Pattern with the three annotators, however such a significant difference begs the question of how good SentiStrength really performs. The various suggestions for improving sentiment analysis mentioned in the previous section (6.2) could generate sentiment scores that vary more in sentiment and possibly represent sentiment on Twitter better.

6.4 The window size that results in the best topic models is of 7 days

For all three datasets, the mean coherence score is highest when the data is cut in chunks of one week worth of tweets, regardless of the number of topics per topic model (Figure 5.2). This means the most coherent topic models are calculated when the tweets originate from only one week, and thus a smaller dataset. In contrast, in literature, topic coherence is expected to get higher the bigger the size of the corpus (Omar, On, Lee, & Choi, 2015).

A possible explanation for the phenomenon seen here is that when only taking tweets from one week, time does not influence the topic discussed as much compared to when taking a larger time window. This could make it easier for the LDA algorithm to assign certain words to topics. However, when there are more topics present in the dataset or some topics that are present have more version in the dataset, it will become harder for the LDA algorithm to assign a word to a topic and give it a weight, because there are more topics, or versions of topics, the same word could belong to. We see that in the examples of the topic models with low coherence (Figure 5.7, Figure 8.8 and Figure 8.10) there are multiple words, like "farmer" and "nitrogen crisis" that are present in many, if not all, topics.

The results imply that performing LDA on tweets results in the highest topic coherence score when performed on datasets of small time frames. This can be taken as a takeaway point for future research that seeks to get insights into the topics discussed on social media through topic modelling. Especially when aiming to perform topic modelling close to real-time, this implication is favourable: data of short time periods can be collected faster than if researchers need to wait for a longer time before a coherent topic model can be calculated.

For future research, it is recommended to compare the topic coherence scores and the weekly volume of tweets, to further examine the connection between the topic coherence and the size of the dataset the topic model is generated with.

6.5 The number of topics that results in the best topic models varies per sentiment

The number of topics that result in the highest mean coherence for the full and the negative dataset is 16 (Figure 5.2). For the positive dataset, this value is 4. These results imply that less topics are discussed in positive tweets than negative tweets. One way to interpret this is to say that, during the Dutch Nitrogen Crisis, people have more to complain about than to praise. As the overarching topic of this dataset is a crisis, that makes sense.

Note, though, that there are a few differences between the datasets. First of all, the negative dataset is bigger than the positive, consisting of roughly 3.5 times more tweets. Also, the full dataset is a little over twice as big as the negative dataset. Interestingly, in line with the finding discussed in the previous paragraph, the smaller datasets seem to have the higher mean topic coherence: the upper range of the negative dataset is about 0.05 higher than the full dataset, while the positive dataset has a max mean topic coherence of 0.095 higher than the full dataset.

It is possible that in the larger datasets there are simply too many topics: 16 is not enough to represent them. For future research, it is advised to train topic models of over 16 topics as well, as the topic coherence might increase in that direction. During this study, this was not done because it took too long for topic models of over 16 topics to converge. Furthermore, there are more hyperparameters for LDA topic models, like learning decay, learning offset and maximum iterations. These hyperparameters were set on default during this study. In future work, it would be interesting to tune these hyperparameters and see if this increases the topic coherence of topic models.

Finally, to improve the quality of the topic models, LDA MALLET could be used instead of the classic Gensim LDA. LDA MALLET works similarly to classic LDA, but uses a different sampling method which takes a longer time to compute and often derives more coherent topics than the classic LDA (Mccallum, 2002; Dawar, Samuel, & Alvarado, 2019). Another alternative to standard LDA is structural topic modelling: an approach that is similar to LDA, but differs because it incorporates metadata of documents such as for example the author's name, gender or political affiliation, to derive more precise topics (Kuhn, 2018; Roberts, et al., 2014). The Twitter API has a lot of such metadata available, and thus structural topic modelling could lead to better results than the topic modelling results of this study (Twitter, n.d.).

6.6 Having a topic represented by the 10 most important words leads to the most variation between subsequent topic models over time

It is inherent that when topics consist of 10 words, a change of, for example, one word in a topic will lead to a higher change in the MJS of two topic models than when one in 20 words changes. Seeing the MJS changes more extremely over time when topics consist of 10 words compared to when topics consist of 10+ words, means that increasing the number of words representing a topic does not necessarily mean that, on average, changing words are added to the topics as well.

As the highest weights are given to the top 10 words in a topic, it would be unfavourable to add more words at the risk of adding noise. However, the choice of 10 words was based on eyeballing the differences between the MJS plots and their stds. To test the assumption that the top 10 words have the highest weights, it is recommended to plot the distributions of topics. These distributions will show if indeed the top 10 words have generally substantially higher weights than the other words. If this is the case, the choice of words per topic is better substantiated.

6.7 It is possible to identify human interpretable topics based on the topic model with the highest scoring topic coherence

Topic coherence is a quality measure designed that approximates how interpretable humans find topics (Newman, Lau, Grieser, & Baldwin, 2010). Therefore, it is expectable that most topics in the topic models with the highest topic coherence score are interpretable. What is noticeable about the topic models with the lowest topic coherence, is that the topics look very similar, often containing various overlapping words. Also, the highest weights of the topics in the topic models with low topic coherence, are significantly lower than the highest weights of words in topics from topic models with high topic coherence.

To improve the interpretability of these topic models for the Nitrogen Crisis in future research, it is recommended to identify phrases that act like single words, and add them to the corpus concatenated (e.g., 'social_media' instead of the separate words 'social' an 'media') during the pre-processing of text. These are called collocations (Lau, Baldwin, & Newman, 2013). There are various methods for finding and selecting the most meaningful collocations, like frequency counting, hypothesis testing and Pointwise Mutual Information (Lau, Baldwin, & Newman, 2013; Kumova Metin S., 2010; Rao & Taboada, 2021). The exemplary topic model in Figure 6.2 shows a case where collocations would be useful: the highlighted words "den" and "haag", which together form "The Hague", are present separately in three of the 10 topics, because the always

co-occur (for none-Dutch readers: 'den' is not the regular word for 'the' in Dutch, 'then' is only used in old Dutch city names).

```
------ Topic 1 ------
boer stikstofcrisi wel houden politiek moeten land weer protest groot

------ Topic 2 ------
boer agractie den haag weg chaos tractor boeren kabinet stikstofcrisi

------ Topic 3 ------
boer boeren boerenopstand komen wel agractie jesseklaver moeten uitstoot zien

------ Topic 4 ------
boer den weer haag wel vandaag zeggen weg denken jullie

------ Topic 5 ------
boer komen stikstofcrisi agractie boeren staan laten malieveld ander weg

------ Topic 6 ------
boer den haag malieveld laten boeren agractie boerenopstand staan tractor

------ Topic 7 ------
boer stikstofcrisi jaar boeren wel allemaal crisis klimaat ander laten

------ Topic 8 ------
boer den wel jesseklaver haag jaar boeren zien alleen onze

------ Topic 9 ------
boer nederland boeren politie agractie wel moeten vinden jullie mens

------ Topic 10 ------
boer komen jullie wel zeggen boeren alleen laten weten waar
```

Figure 6.2 | Topic model from the negative dataset in September 2019, where number_topics = 10 window_slice = 18.

6.8 The range of difference in the mean Jaccard similarity is small [0.75, 1]

On average, between 8.6 and 10 words overlap per topic in two subsequent topic models. This is high considering that topics comprise 10 words. It means the mean change in words per topic is low. This means collectively, topics discussed change only a little per time interval. However, the MJS represents the change in *all* topics discussed in two topic models. For the MJS to be very low, from one week to another, *all* topics need to have changed significantly compared to last week or 4 weeks ago. However, the dataset that is used already described a specific topic: the Nitrogen Crisis. So, all topics identified by the topic models are within the specific context of the Nitrogen Crisis. Imagine, at some point in time, we have a topic model of 4 topics about 1. The Nitrogen policy, 2. The farmers' demonstrations, 3. The construction demonstrations and 4. How measurements on nitrogen emissions are performed. However, the week after, a brick is thrown through a window of the house of the minister of Agriculture, Nature and Food Quality. Then the first three topics of this week's topic model are the same as last week, but this week's topic 4. is about the shattered window of the minster. In this example, although the news of the shattered window gets Twitter coverage, the MJS does not change much because only one

topic in the topic model changes and not all four. It would even change less if the topic model does not comprise 4, put 16 topics.

The previous example could explain how the MJS could be consistently quite high even though new topics are discussed. It implies that the MJS would not properly capture this change. For future research, it is recommended to try out alternative formulas for calculating the difference in topic discussed between topic models. One disadvantage of using Jaccard similarity is that the weights of each word in a topic are ignored: every word in the topic is equally relevant. Several formulas that compare topic distributions are Kullback-Leibler divergence, Bhattacharyya distance and Hellinger distance (Hellinger, 1909; Paul, 2009; Bhattacharyya, 1946). It is advised to try these formulas out and see if topics discussed over time are better captured by any of them.

6.9 A spike in the weekly volume of tweets seems the best indicator for an event, not sentiment analysis or topic modelling results

During a crisis where many events received large-scale news media and social media attention, it is expectable that the volume of tweets goes up around those events. However, when events, like demonstrations, are discussed on social media extensively, it was expected that there would be some changes in the sentiment and topics discussed. If a change in topics discussed on Twitter around those events is really present during those events, implementing the improvements suggested in the previous sections of this chapter could lead to more precise results that could mark these events. It is also possible that around those events there was simply minor change in *what* topics were discussed, and merely a change in *how much* the topics were discussed on Twitter.

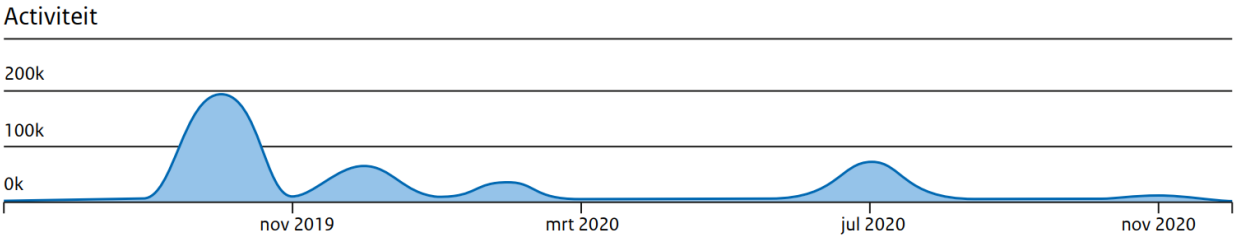


Figure 6.3 | Social media reporting activity between July 2019 - December 2020 (Kalkhoven, 2021)

The spikes in volume overlap well with the plot of volumes of tweets in the media analysis of the Ministry of the Interior and Kingdom Relations, shown in Figure 6.3 above (Kalkhoven, 2021).

However, in that analysis 380.000 tweets were collected, while in this study 292.245 tweets were collected. Interestingly, that analysis was only about the farmers' demonstrations, not about the full Nitrogen Crisis, unlike this study. Therefore, it is recommended to evaluate the query that was developed for the collection of Nitrogen Crisis tweets in this study. One way to do that is to use the entry in the excel sheets filled in by the annotators (Figure 4.4 from section 4.6.1 shows a snapshot), responding to whether the annotator thought the tweet was about the Nitrogen Crisis, to find what characterises tweets that are collected by the query but which are not about the Nitrogen Crisis.

There are three small spikes in tweet volume: two in November 2019 and on in July 2020. These are unexplained by the events identified in this research or in Kalhoven's analysis or on the WIkipedia Nitrogen Crisis overview page (Wikipedia, n.d.; Kalkhoven, 2021)..\ All three take place shortly after demonstrations. One interpretation is that some demonstrations and threats to politicians caused so much unrest that they continued being discussed in the following weeks. However, the MJS goes down after the second farmers' demonstrations in October, not up, contradicting this interpretation. Another interpretation could be that the demonstrations and threats received much media attention in the following weeks that tried to reconstruct how these demonstrations and the Nitrogen Crisis started, like this RTL Nieuws news article: (RTL Nieuws, 2019).

As a spike in the weekly volume of tweets seems the best indicator for an event, decision-makers are recommended to monitor pre-determined topics on Twitter and implement a way to be notified when a significant change in volume of tweets takes place, while lacking better tools. When receiving a notification, the processing pipeline of this study to look at the topic models of those weeks to see what people are discussing.

6.10 Additional remarks

One of the limitations of using Twitter as a data source for analysing public opinion (discussed in Chapter 2: Related Work, section 3.4) is that some powerful stakeholders are might not be present on Twitter. Han de Groot mentioned that is the case in the Nitrogen Crisis: Johan Vollenbroek (H. De Groot, personal communication, May 26, 2021). Johan Vollenbroek does not have a social media presence, but has played a key role before and during the Nitrogen Crisis (Hakkenes, 2019). This means, even though he has an immense footprint on the Nitrogen Crisis, he does not influence the discourse online directly. When is researching Twitter discourse on policy crisis, like the Nitrogen Crisis, it is therefore recommended to complement it with other data sources, like news articles or interviews.

The interviews held in this study were held with one interviewee. For a broader context on the policy crisis at hand, it is recommended to interview various stakeholders with different

perspectives to reduce the risk of selection bias. Also, as this research aimed at generating insights for decision-makers, for future research it is recommended to ask decision-makers of the Nitrogen Crisis how they interpret the results of this study and what could make it more useful to them.

According to Han de Groot, the insights this research provides are only useful if it can be connected to a *who*: who is discussing what topics? Who are feeling what? Who has much influence on Twitter? Therefore, a significant improvement in the usefulness to decision-makers could be made by extending this research with a network analysis. The data is already available: the metadata of each tweet in this dataset contains an id, the number of likes, the number of retweets and, if applicable, a list of Twitter users that are mentioned in the tweet's text. Tracking active stakeholders can be used by decision-makers in three ways. First, by monitoring a list of stakeholders that they believe are influential, and verify this by looking at their Twitter presence. Second, for discovering new influential stakeholders, when influential Twitter users are not on the list of pre-defined influential stakeholders. This way, new stakeholders can be added to the list early on, and decision-makers can decide to include them in the policy-making process. Third, by monitoring the Twitter presence of stakeholders, changes in support base of each stakeholder can be noticed early on. This is useful information for decision-makers during negotiations with those stakeholders.

There are more useful resources that this study has not used. First, hashtags emphasize words, therefore it is recommended to try giving words that are preceded with a hashtag higher weights, and see if this improves the topic coherence of the generated topic models. Second, retweets were not used in this study, while it has been claimed that sharing tweets without adding text can be a sign of agreeing with the tweet (Goldenberg, et al., 2020). Third, many tweets contain URLs, often linking to articles on news websites. It could be interesting for decision-makers to map what news sources are spread most widely, to see where different groups on Twitter are getting news from. This could help in understanding where potential unrest is originated. Fourth, this thesis does not consider emojis or emoticons in establishing sentiment scores. Emojis and emoticons serve writers to express sentiment without using words. Some sentiment libraries score the sentiment of emoticons. SentiStrength has a lexicon for emoticons (e.g. :-) or :,-(etc.) but emoticons are not used often anymore, so the lexicon is of little use (Thelwall, Buckley, Paltoglou, & Kappas, 2010). Nowadays, the old-fashioned emoticons have made place for the newer and popular emojis: pictograms of expressive faces or other objects that are available in a wide range of emotions which can be embedded in text. Emojis sometimes are interpreted differently by people of different backgrounds, ages, etc. Hence, it is not a straightforward task to classify the sentiment of each emoji (Ishmael, 2021). For future research, it is recommended to determine and use only the emojis on whose interpretation most Twitter users agree to improve overall classification of sentiment.

7 Conclusion

This study aims to answer the question: *How can sentiment analysis and topic modelling be applied to Twitter data to provide insights for decision-makers retrospectively about major events during the Dutch Nitrogen Crisis?* This chapter will summarize the main findings of this study, answer the research questions and mention how this study could form a starting point for future research.

SQ 1

How did sentiment on Twitter develop over time during the Nitrogen Crisis?

There was minor variation in the mean sentiment on Twitter over time. However, the std was consistently high, meaning there was a high variance in sentiment scores of tweets each week. In terms of quantity, the number of negative tweets per week was higher than the number of positive tweets throughout the whole crisis, except during one week. The events identified do not explain the short time period during which the average positive sentiment was higher than negative. To find an explanation for this short peak, it is recommended to use the topic models generated in this study to find out what topics were discussed that week.

I expected to see some variation in sentiment around major events. For example, more positivity in tweets when the farmers had high public endorsement during the first demonstrations. It is possible that the methods applied in this study, Pattern and SentiStrength, were not accurate enough for good sentiment classification, or that having three annotators score 100 tweets were not enough to evaluate which tool was best. For future application of sentiment analysis on tweets about a Dutch policy crisis, a different approach is recommended: either translate tweets to English and apply a more advanced English sentiment analysis tool for possibly more accurate results, or train a new sentiment analysis model specifically for the task at hand. Also, it is advised to work with over three annotators, whose demographics should approximate the demographics of Dutch Twitter users.

SQ 2

How can topic modelling be applied to analyse changes in topics discussed on Twitter over time?

First, this study proposes a pipeline with a grid search for selecting what time window to calculate topic models over and how to select the number of topics that will lead to topic models with, on average, the highest topic coherence. This is a contribution to future work, as applying this approach has the potential of saving time and systematically generating meaningful topic models with high topic coherence scores.

Second, this study proposes MJS as a measure for difference in topics discussed over time. Calculating MJS makes it possible to plot changes in topics discussed over time, without needing to look at each individual topic model and manually identify changes. Additionally, not only topic models of 1 week apart are plotted but also topic models of 4 weeks apart. Plotting this information shows if changes in topics discussed occurred incrementally and slowly, or sudden shocks. By doing this, this study shows a new and automated way of finding changes in topic models with no need for human inspection of topics in each topics model manually.

SQ 3

How did the topics discussed during the Nitrogen Crisis on Twitter change over time?

Collectively, according to the topic modelling and MJS plots, topics discussed changed only a little during the Nitrogen Crisis: the range in which the MJS varies is small. It is possible that the MJS simply does not effectively represent changes in topics discussed, in which case changes in some topics in a topic model do not change the MJS enough to be noticeable.

To represent changes in topics discussed more accurately in future work, this study makes various suggestions for improving the quality of topic modelling. For example, suggestions are made for fine tuning text preprocessing steps and for alternative topic modelling algorithms. Finally, alternative metrics for MJS are proposed to better capture change in topics discussed in future work.

To improve the traceability of changes in topics over time, this study recommends improving the quality of the topic models. Several suggestions are made for fine tuning text preprocessing steps and for alternative topic modelling algorithms. Finally, alternative metrics for MJS are proposed to better capture change in topics discussed in future work. When implementing these suggestions, the pipeline developed in this study could prove useful to decision-makers for plotting changes in topics discussed over time.

SQ 4

Can sentiment analysis and topic modelling results provide markers for events that received attention from news media during the Nitrogen Crisis?

Plotting the results of sentiment analysis and the MJS per sentiment datasets did not provide markers for most of the events. There is no observable pattern in the mean sentiment or volume of tweets per sentiment that suggests these sentiment results can provide markers. However, in this Master's thesis research many steps were proposed and executed. Because of limited time that needed to be divided over all steps, only so much time could be spent on each individual step. Because of this, the concept proposed in this study could further be explored by the improvements proposed for each step, to find out whether observable patterns appear in more precise results. If so, this can aid the development of tools for marking important events during a policy crisis for decision-makers.

MQ

How can sentiment analysis and topic modelling applied to Twitter data to provide insights for decision-makers retrospectively about major events during the Dutch Nitrogen Crisis?

Unexpectedly, it is not sentiment analysis or topic modelling results that have the most obvious connection with the events identified: it is an increase in tweets during events. Therefore, while lacking better tools, decision-makers are recommended to monitor pre-determined topics on Twitter and implement a way to be notified when a significant change in volume of tweets takes place. The combination of sentiment analysis and topic modelling as implemented in this research is either not advanced enough to provide useful information to decision-makers, or sentiment analysis and topic modelling simply cannot provide insightful results on the Dutch Nitrogen Crisis. However, because there is an extensive amount of research applying these methods to social media data around various political events with valuable results, I argue that more experiments need to be performed with this approach and the quality of each research step needs to be further improved in order to draw final conclusions on the usefulness of the combination of sentiment analysis and topic modelling for decision-makers during policy crises. In this study, many approaches for such improvements are suggested.

This study has only explored a fraction of the potential of Twitter data. I hope the concepts elaborated in this study and the expansions that are suggested will inspire future research to improve these concepts and use them to compute meaningful outcomes that are interpretable to- and useful to decision-makers.

Key Findings of Chapter 6: Conclusion

1. It is not sentiment analysis or topic modelling results that have the most obvious connection with the events identified: it is an increase in tweets during events
2. The combination of sentiment analysis and topic modelling as implemented in this research is either not advanced enough to provide useful information to decision-makers, or sentiment analysis and topic modelling simply cannot provide insightful results on the Dutch Nitrogen Crisis
3. This study and the pipeline it proposes can serve as a solid basis for further development into a process that provides ready-to-use information to decision-makers

8 Appendix

A. Questions from first interview Han de Groot

1. For your stakeholder research, how did you decide who were the stakeholders you wanted and should interview?
2. Were there stakeholders that you identified as important but didn't get to talk to? If so, why?
3. According to the report you wrote there is a need for more "integrality". What is integrality and what is its role during the Nitrogen Crisis?
4. "Koppel de stikstofaanpak waar enigszins mogelijk aan andere belangrijke milieu-trajecten , zoals Klimaat- en het Schone Lucht akkoord en vernieuwing van het ruimtelijk beleid" (De Groot, 2021, p. 21) – What do you mean by that?
5. What is the relation between nitrogen and PFAS?
6. I read that only 1% of the Dutch nitrogen emissions are emitted by the building sector, what does the building sector have to do with the nitrogen crisis?
7. "voor velen is zowel de ambtenaar van AGRO als die van DGS een vertegenwoordiger van de minister van LNV, Carola Schouten, Verschillen in aanpak of positie worden dan slecht of niet begrepen. " (De Groot, 2021, p. 7)– What are these "differences"?
8. "In de eerdere fase, na de Raad van State uitspraak, waren weliswaar de koppen bij elkaar gestoken maar was het interdepartementale niet echt gescheiden van het interbestuurlijke." (De Groot, 2021, p. 8) – What is the relation to the Nitrogen Crisis debate
9. "Tegelijkertijd moet elke overheid stoppen de ander als stakeholder te zien, dat staat het zicht op de echte stakeholders in de weg. " (De Groot, 2021, p. 8) – Is that possible in an organisation as big and diverse as the Dutch government? And does it have something to do with: "In de loop van 2020 is een aantal studies en beleidsnota's

verschenen over rollen en interactie tussen Rijk en medeoverheden. Door de oogharen bezien, wijzen die op de noodzaak van meer centrale regie bij het Rijk en een goede definitie binnen het 4W model van de antwoorden op: Wie?, Wat?, Wie doet Wat? en Waarmee?." (De Groot, 2021)

10. "Landbouw en boeren worden vaak als één gezien" (De Groot, 2021, p. 8) – What do you mean by this?
11. Who are the most important stakeholders in the Dutch agriculture and its sub sectors?
12. "Met FDF is het contact tussen LNV/DGS verbroken na incidenten en bedreigingen" (De Groot, 2021, p. 8) – What happened?
13. Do Farmers Defence Force (FDF) and Agractie only protest nitrogen policies or do they do more?
14. As one of the causes for the harmed relations between the government and the agricultural sector, you name the transfer of tasks to provinces. What kind of tasks do you mean and how did this happen?
15. Nowhere in the report you mention the Covid-19 outbreak. Did this pandemic have any influence on the Nitrogen Crisis?
16. (De Groot, 2021, p. 20): Why do the Industry and Energy sectors have 2 seats and the Mobility sector just one?
17. Is there a link between 'Mercosur' and the Nitrogen Crisis and debate?

B. Questions from second interview Han de Groot

These questions are translated from Dutch to English with Google Translate.

1. You once said that you would have liked to have been to a ministry in the Netherlands when the crisis started, because you already saw it coming:
    a. What made you see the nitrogen crisis coming?
    b. What would you do if you were there at the time? What kind of clusters of subjects do you expect to emerge?
2. What clusters of topics do you expect to emerge?
3. How were the opinions of citizens included/not included in the nitrogen policy?
    a. What could be done better there?
4. When do you think social media insight contributes most to crisis policy?
    a. Before a crisis
    b. During a crisis
    c. After a crisis, as an evaluation learning moment
5. Why is Twitter/social media specifically interesting for policy?
6. And HOW does social media insight contribute?
7. What do you think of hashtags? How relevant are they in your view?
8. Does it matter who the user is who writes the tweet?
    a. And if so, how does that matter?
9. What do you see as the (largest) contribution of these insights to policy?
10. Is the government open to monitoring on Twitter at all?

C. Plot of volume of tweets per month

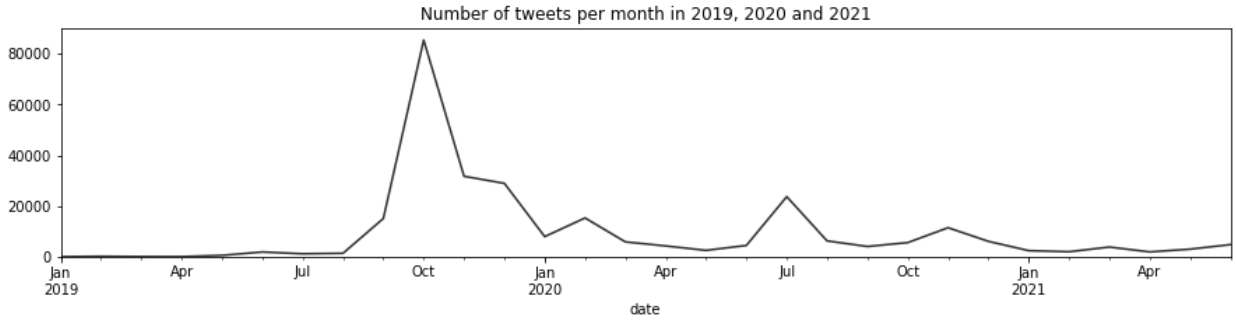


Figure 8.1 | the number of tweets about the Nitrogen Crisis per month, from January 2019 up to June 2021.

D. List of stop words

echt
gaan
goed
heel
maak
maken
per
probleem
stikstof
stikstofcrisis
stikstofprobleem
wil
willen
wilt
zal
zullen

E. Pattern results

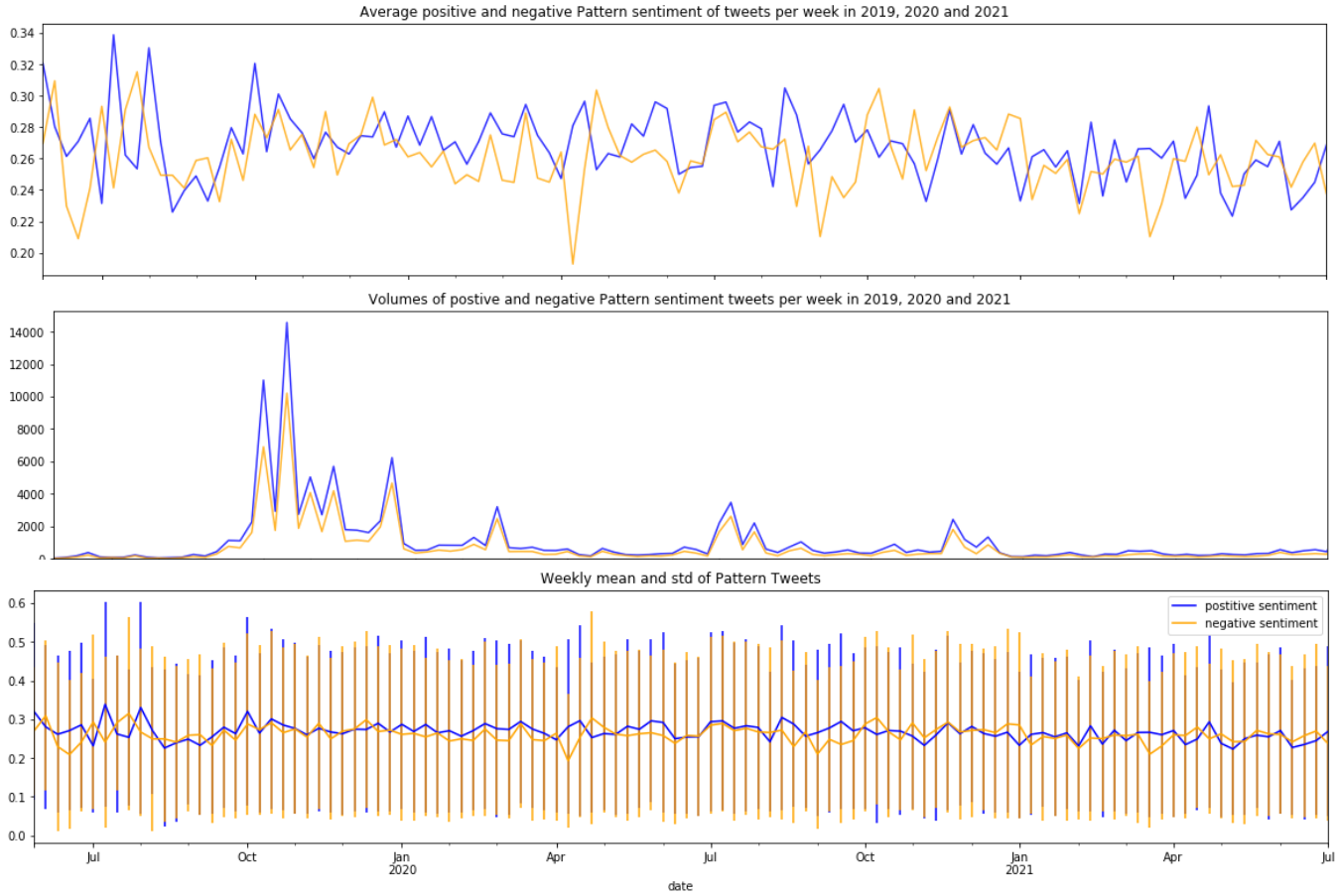


Figure 8.2 | Results of Pattern sentiment analysis.

F. Jaccard distance depending for number of words per topic = 10 to 40, and comparing various weeks

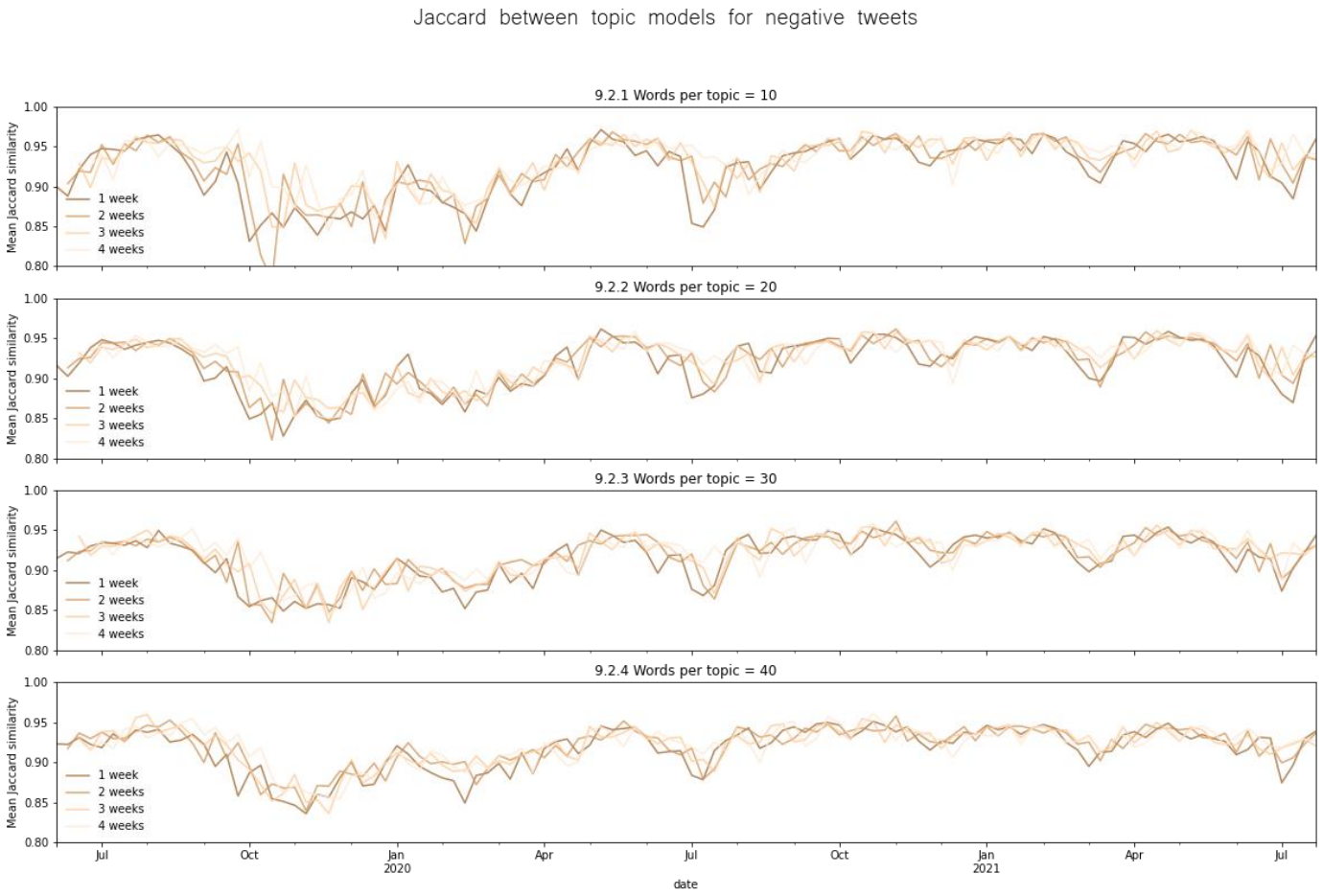


Figure 8.3 | Mean Jaccard similarity for negative tweets, with either 1, 2, 3 or 4 weeks time in between two topic models. Each plot shows the mean Jaccard similarity of two subsequent topic models when topics are represented by either 10, 20, 30 or 40 words.

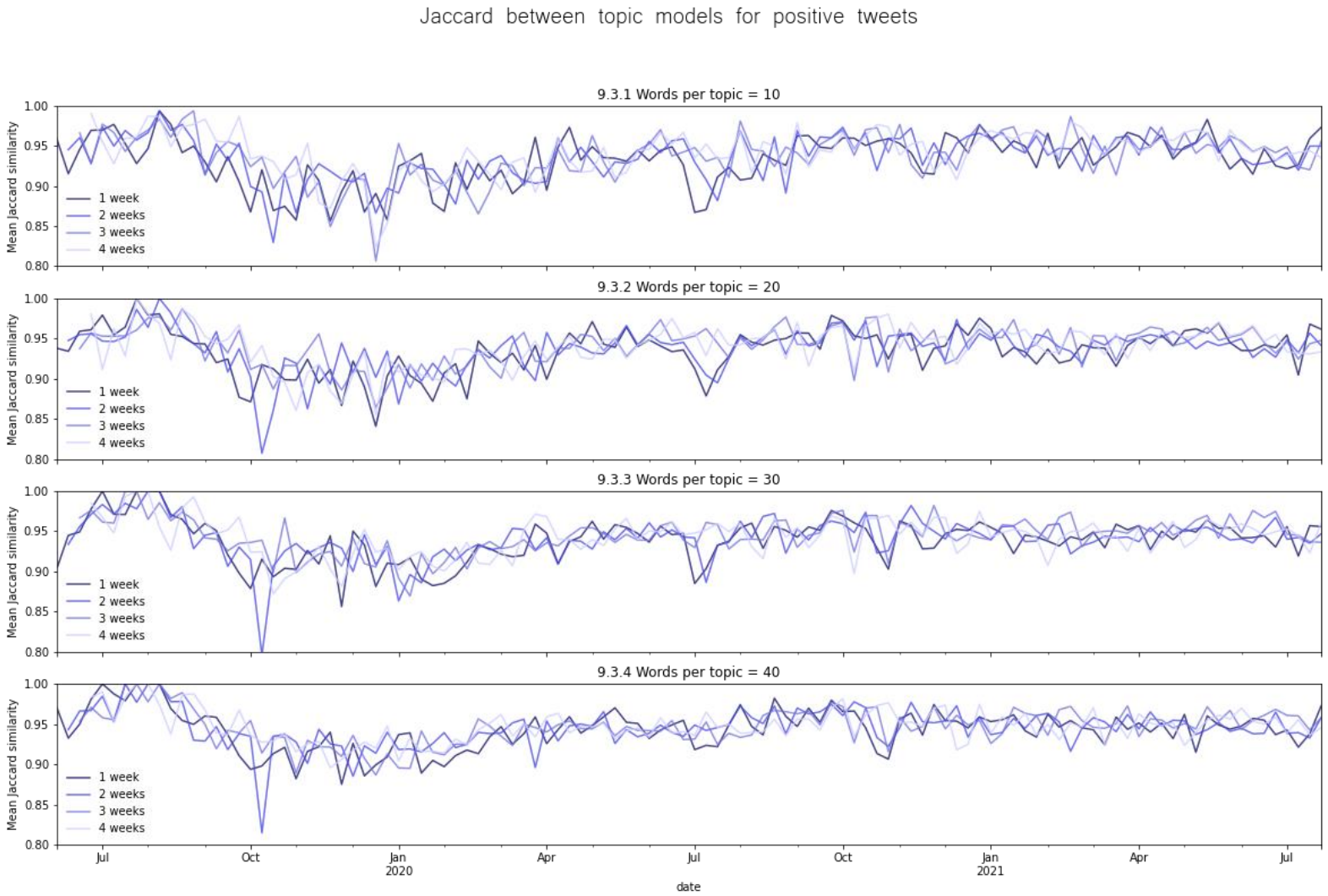


Figure 8.4 | Mean Jaccard similarity for positive tweets, with either 1, 2, 3 or 4 weeks time in between two topic models. Each plot shows the mean Jaccard similarity of two subsequent topic models when topics are represented by either 10, 20, 30 or 40 words.

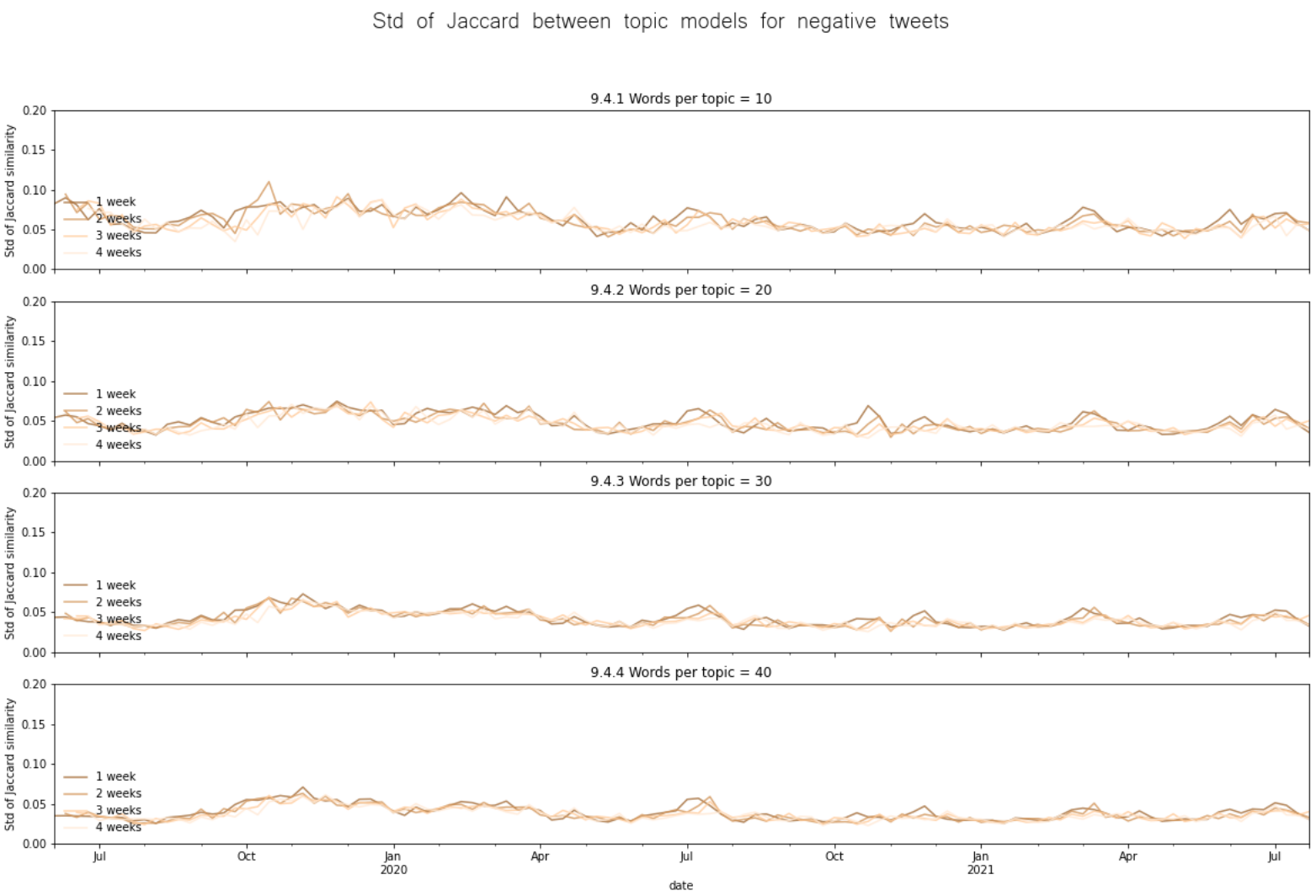


Figure 8.5 | Std for negative tweets, with either 1, 2, 3 or 4 weeks time in between two topic models. Each plot shows the std of Jaccard similarity of two subsequent topic models when topics are represented by either 10, 20, 30 or 40 words.

Figure 8.6 | Std for positive tweets, with either 1, 2, 3 or 4 weeks time in between two topic models. Each plot shows the std of Jaccard similarity of two subsequent topic models when topics are represented by either 10, 20, 30 or 40 words.

G. Examples of topic models per sentiment with high and low coherence

Figure 8.7 | Word Cloud with the highest coherence score from the topic models of the negative tweet dataset.

Figure 8.8 | Word Cloud with the lowest coherence score from the topic models of the negative tweet dataset.

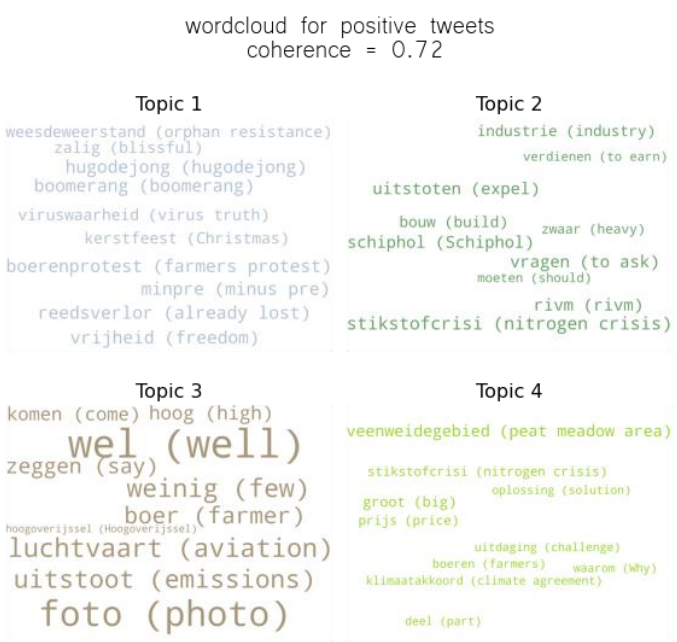


Figure 8.9 | Word Cloud with the highest coherence score from the topic models of the positive tweet dataset.

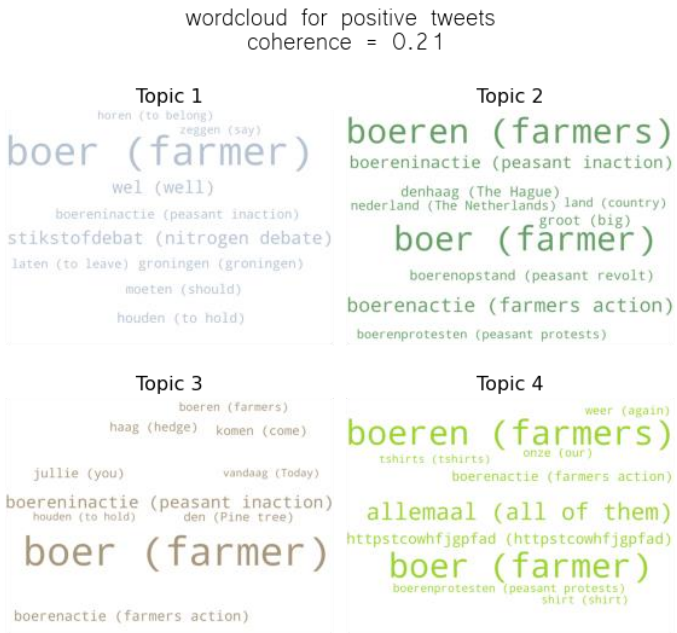


Figure 8.10 | Word Cloud with the lowest coherence score from the topic models of the positive tweet dataset.

9 References

Allahyari, M., Pouriyeh, S., Kochut, K., & Arabnia, H. (2017). A Knowledge-Based Topic Modeling Approach for Automatic Topic Labeling. *International Journal of Advanced Computer Science and Applications, 8*(9), 335-349. doi:https://doi.org/10.14569/IJACSA.2017.080947

Allcott, H., & Gentzkow, H. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives, 31*(2), 211-36. doi:https://doi.org/10.1257/jep.31.2.211

Anderson, A., Allan, S., Petersen, A., & Wilkinson, C. (2005). The Framing of Nanotechnologies in the British Newspaper Press. *Science Communication, 27*(2), 200-220. doi:https://doi.org/10.1177/1075547005281472

Araujo, M. R., Pereira, A., & Benevenuto, F. (2016). An evaluation of machine translation for multilingual sentence-level sentiment analysis. *Proceedings of the 31st Annual ACM Symposium on Applied Computing (SAC '16)* (pp. 1140–1145). New York, NY, USA: Association for Computing Machinery. doi:https://doi.org/10.1145/2851613.2851817

Balahur, A., & Turchi, M. (2012). Multilingual sentiment analysis using machine translation? *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis* (pp. 52-60). Jeju, Republic of Korea: Association for Computational Linguistics. doi:https://doi.org/10.5555/2392963.239297

Baldwin, M., & Lammers, J. (2016). Past-focused environmental comparisons promote proenvironmental outcomes for conservatives. *Proceedings of the National Academy of Sciences, 113*(52), 14953–14957. doi:https://doi.org/10.1073/pnas.1610834113

Bashar, M. A., Nayak, R., & Balasubramaniam, T. (2020). Topic, sentiment and impact analysis: Covid19 information seeking on social media. *arXiv preprint*. Retrieved from https://arxiv.org/abs/2008.12435

Bellovary, A., Young, N., & Goldenberg, A. (2021). Left- and Right-Leaning News Organizations Use Negative Emotional Content and Elicit User Engagement Similarly. *Affective Science*. doi:https://doi.org/10.1007/s42761-021-00046-w

Berrios, R., Totterdell, P., & Kellett, S. (2015). Eliciting mixed emotions: a meta-analysis comparing models, types, and measures. *Frontiers in Psychology, 6*(428). doi:https://doi.org/10.3389/fpsyg.2015.00428

Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations. *Sankhyā: the Indian Journal of Statistics*, 401-406.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc.

Blankers, M., Van der Gouwe, D., & Van Laar, M. (2019). 4-Fluoramphetamine in the Netherlands: Text-mining and sentiment analysis of internet forums. *International Journal of Drug Policy, 64*, 34-39. doi:https://doi.org/10.1016/j.drugpo.2018.11.016

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd international conference on Machine learning (ICML '06)* (pp. 113–120). New York, NY, USA: Association for Computing Machinery. doi:https://doi.org/10.1145/1143844.1143859

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 993–1022.

Borchardt, F. (1988). Neural Network Computing and Natural Language Processing. *CALICO Journal, 5*(4), 63-75.

Boussalis, C., & Coan, T. G. (2016). Text-mining the signals of climate change doubt. *Global Environmental Change, 36*, 89–100. doi:https://doi.org/10.1016/j.gloenvcha.2015.12.001

Cambria, E., Gastaldo, P., & Bisio, F. &. (2015). An ELM-based model for affective analogical reasoning. *Neurocomputing, 149*(Part A), 443-455. doi:https://doi.org/10.1016/j.neucom.2014.01.064

Camisani-Calzolari, M. (2012). *Analysis of Twitter followers of the US Presidential Election candidates: Barack Obama and Mitt Romney*. Retrieved from http://digitalevaluations.com/.

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media and Society, 16*(2), 340–358. doi:https://doi.org/10.1177/1461444813480466

Chen, Q., Min, C., Zhang, W., Wang, G., Ma, X., & Evans, R. (2020). Unpacking the black box: How to promote citizen engagement through government social media during the Covid-19 crisis. *Computers in Human Behavior, 110*, 106380. doi:https://doi.org/10.1016/j.chb.2020.106380

Dagblad, R. (2021). *Politici oogsten afschuw en applaus op het Malieveld*. Retrieved from Reformatorisch Dagblad: https://www.rd.nl/artikel/934806-politici-oogsten-afschuw-en-applaus-op-het-malieveld

Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. A., & Gelbuk, A. (2016). Multilingual Sentiment Analysis: State of the Art and Independent Comparison of

Techniques. *Cognitive Computation, 8*, 757–771. doi:https://doi.org/10.1007/s12559-016-9415-7

Dawar, K., Samuel, A. J., & Alvarado, R. (2019). Comparing Topic Modeling and Named Entity Recognition Techniques for the Semantic Indexing of a Landscape Architecture Textbook. *2019 Systems and Information Engineering Design Symposium (SIEDS)*, (pp. 1-6). doi:https://doi.org/10.1109/SIEDS.2019.8735642

De Bruijn, H. (2019). *The Art of Political Framing: How Politicians Convince Us That They Are Right*. Amsterdam: Amsterdam University Press.

De Groot, H. (2021). *Verkenning governance en participatie stikstofaanpak*. Retrieved from https://www.aanpakstikstof.nl/binaries/aanpakstikstof/documenten/rapporten/2021/03/19/verkenning-governance-en-participatie-stikstofaanpak/Verkenning+Governance+en+Particiatie+Stikstofaanpak.pdf

De Smedt, T., & Daelemans, W. (2012). Pattern for Python. *Journal of Machine Learning Research, 13*(66), 2063–2067.

Dirikx, A., & Gelders, D. (2010). To frame is to explain: A deductive frame-analysis of Dutch and French climate change coverage during the annual UN Conferences of the Parties. *Public Understanding of Science, 19*(6), 732–742. doi:https://doi.org/10.1177/0963662509352044

Doodeman, M. (2019, October). *Stikstofcrisis slaat toe: vergunningenverlening keldert met 34 procent in augustus*. Retrieved November 19, 2021, from Cobouw: https://www.cobouw.nl/marktontwikkeling/nieuws/2019/10/stikstofcrisis-slaat-toe-vergunningenverlening-keldert-met-34-procent-in-augustus-101277830

Driessen, P. (2019). *Holocaust-vergelijking kost boerenprotest sympathie: 'Rij die trekker van je eens naar Auschwitz'*. Retrieved September 10, 2021, from AD: https://www.ad.nl/binnenland/holocaust-vergelijking-kost-boerenprotest-sympathie-rij-die-trekker-van-je-eens-naar-auschwitz~a208b75b/

Durahim, A. O., & Coşkun, M. (2015). #iamhappybecause: Gross National Happiness through Twitter analysis and big data. *Technological Forecasting and Social Change, 99*, 92-105.

Eijsink, R. (2021). *Terugblik op boerenprotest*. Retrieved September 10, 2021, from Nieuwe Oogst: https://www.nieuweoogst.nl/nieuws/2021/07/07/terugblik-op-boerenprotest

Engesser, S., Ernst, N., Esser, F., & Büchel, F. (2016). Populism and social media: how politicians spread a fragmented ideology. *Information, Communication & Society, 20*(8), 1109–1126. doi:https://doi.org/10.1080/1369118x.2016.1207697

FDF Board. (2021). *ALLEEN BOEREN BEHARTIGEN BOERENBELANG!* Retrieved November 22, 2021, from Farmers Defence Force: https://farmersdefenceforce.nl/alleen-boeren-behartigen-boerenbelang/

Gallagher, R. J., Reing, K., & Kale, D. V. (2017). Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics, 5*, 529-542. doi:https://doi.org/10.1162/tacl_a_00078

Giachanou, A., & Crestani, F. (2016). Like It or Not: A Survey of Twitter Sentiment Analysis Methods. *ACM Computing Surveys, 49*(2), 1–41. doi:https://doi.org/10.1145/2938640

Goldenberg, A., Garcia, D., Halperin, E., Zaki, J., Kong, D., Golarai, G., & Gross, J. J. (2020). Beyond emotional similarity: The role of situation-specific motives. *Journal of Experimental Psychology: General, 149*(1), 138. doi:https://doi.org/10.1037/xge0000625

González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying Sarcasm in Twitter: A Closer Look. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 581-586). Portland, Oregon, USA: Association for Computational Linguistics.

Guia, M., Silva, R. R., & Bernardino, J. (2019). Comparison of Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest on Sentiment Analysis. *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2019)* (pp. 525-531). SCITEPRESS – Science and Technology Publications. doi:https://doi.org/10.5220/0008364105250531

Hagen, L., Harrison, T., Uzuner, O., Fake, T., Lamanna, D., & Kotfila, C. (2015). Introducing Textual Analysis Tools for Policy Informatics: A Case Study of e-Petitions. *Proceedings of the 16th Annual International Conference on Digital Government Research (dg.o '15)* (pp. 10–19). New York, NY, USA: Association for Computing Machinery. doi:https://doi.org/10.1145/2757401.2757421

Hakkenes, E. (2019). *Deze man maakte van de stikstofcrisis dé milieukwestie van het jaar.* Retrieved November 21, 2021, from Trouw: https://www.trouw.nl/duurzaamheid-natuur/deze-man-maakte-van-de-stikstofcrisis-de-milieukwestie-van-het-jaar~b07ae7ba/

Hameleers, M. (2020). Populist Disinformation: Exploring Intersections between Online Populism and Disinformation in the US and the Netherlands. *Leadership, Populism and Power, 8*(1), 146. doi:http://dx.doi.org/10.17645/pag.v8i1.2478

Hellinger, E. (1909). Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *Journal für die reine und angewandte Mathematik, 136*, 210-271. doi:https://doi.org/10.1515/crll.1909.136.210

Hong, S., & Kim, S. H. (2016). Political polarization on twitter: Implications for the use of social media in digital governments. *Government Information Quarterly, 33*(4), 777–782. doi:https://doi.org/10.1016/j.giq.2016.04.007

Hunt, E. (2021). *Words matter: how New Zealand's clear messaging helped beat Covid.* Retrieved September 2, 2021, from The Guardian: https://www.theguardian.com/world/2021/feb/26/words-matter-how-new-zealands-clear-messaging-helped-beat-covid?CMP=Share_iOSApp_Other

Ishmael, A. (2021). *Sending Smiley Emojis? They Now Mean Different Things to Different People.* Retrieved November 04, 2021, from Wall Street Journal: https://www.wsj.com/articles/sending-a-smiley-face-make-sure-you-know-what-youre-saying-11628522840

Jaccard, P. (1912). THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE. *New Phytologist, 11*(2), 37–50. doi:https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Jiménez-Zafra, S. M., Sáez-Castillo, A. J., Conde-Sánchez, A., & Martín-Valdivia, M. (2021). How do sentiments affect virality on Twitter? *Royal Society Open Science, 8*(4). doi:https://doi.org/10.1098/rsos.201756

Jurafsky, D., & Martin, J. (2020). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* (third ed.). Upper Saddle River, N.J: Pearson Prentice Hall.

Kalkhoven, L. (2021). *Media-analyse Boerenprotesten.* Den Haag: Ministerie van Binnenlandse Zaken en Koninkrijkrelaties.

Klumpenaar, S., & Van Laarhoven, K. (2019). *Provincies zwichten voor boerenprotest: geen nieuwe stikstofregels voor boeren.* Retrieved October 1, 2021, from NRC Handelsblad: https://www.nrc.nl/nieuws/2019/10/14/ook-provincie-drenthe-schort-stikstofregels-op-na-boerenprotest-a3976663

Kos, J. (2020). *Leden Farmers Defence Force brengen Rob Jetten thuis voedselpakket.* Retrieved October 1, 2021, from NRC Handelsblad: from: https://www.nrc.nl/nieuws/2020/10/28/vijf-leden-farmers-defence-force-brengen-rob-jetten-thuis-voedselpakket-a4017685

Kuhn, K. D. (2018). Using structural topic modeling to identify latent topics and trends in aviation incident reports. *Transportation Research Part C: Emerging Technologies*, 105-122. doi:https://doi.org/10.1016/j.trc.2017.12.018

Kumova Metin S., K. B. (2010). Collocation Extraction in Turkish Texts Using Statistical Methods. In R. E. Loftsson H. (Ed.), *Advances in Natural Language Processing. NLP 2010. 6233*, pp. 238-249. Springer, Berlin, Heidelberg. doi:https://doi.org/10.1007/978-3-642-14770-8_27

Lau, J. H., Baldwin, T., & Newman, D. (2013). ACM Transactions on Speech and Language Processin. *ACM Trans, 10*(3), 1-14. doi:https://doi.org/10.1145/2483969.2483972

Lee, C., Shin, J., & Hong, A. (2018). Does social media use really make people politically polarized? Direct and indirect effects of social media use on political polarization in South Korea. *Telematics and Informatics, 35*(1), 245-254. doi:https://doi.org/10.1016/j.tele.2017.11.005

Lee, F. (2016). Impact of social media on opinion polarization in varying times. *Communication and the Public, 1*(1), 56–71. doi:https://doi.org/10.1177/2057047315617763

Lee, F. (2016). Impact of social media on opinion polarization in varying times. *Communication and the Public, 1*(1), 56–71. doi:https://doi.org/10.1177/2057047315617763

Leeuwarder Courant. (2019). *Waarom protesteren de boeren en wat willen ze bereiken?* Retrieved September 1, 2021, from Leeuwarder Courant: https://lc.nl/friesland/Waarom-protesteren-de-boeren-en-wat-willen-ze-bereiken-24878113.html

Liu, L., Tang, L., Dong, W., & Yao, S. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus, 5*, 1608. doi:https://doi.org/10.1186/s40064-016-3252-8

Looijenga, M. (2018). *The Detection of Fake Messages using Machine Learning*. Retrieved September 1, 2021, from http://essay.utwente.nl/77385/

Luyi, Z., & Wei Song, W. (2016). LDA-TM: A two-step approach to Twitter topic data clustering. *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, (pp. 342-347). doi:https://doi.org/10.1109/ICCCBDA.2016.7529581

Malchik, A. (2019). *The Problem With Social-Media Protests*. Retrieved from The Atlantic: https://www.theatlantic.com/technology/archive/2019/05/in-person-protests-stronger-online-activism-a-walking-life/578905/

Manikonda, L., Beigi, G., Kambhampati, S., & Liu, H. (.-3.-3.-9.-6. (2018). #metoo Through the Lens of Social Media. In R. Thomson, C. Dancy, A. Hyder, & H. Bisgin (Ed.), *SBP-BRiMS 2018. Lecture Notes in Computer Science. 10899*, pp. 104-110. Cham: Springer International Publishing. doi:https://doi.org/10.1007/978-3-319-93372-6_13

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis–A review of research topics, venues, and top cited papers. *Computer Science Review, 27*, 16-32. doi:https://doi.org/10.1016/j.cosrev.2017.10.002

Mccallum, A. K. (2002). *MALLET: A Machine Learning for Language Toolkit*. Retrieved November 21, 2021, from Manning College of Information & Computer Sciences: https://people.cs.umass.edu/~mccallum/mallet/

McComas, K., & Shanahan, J. (1999). Telling Stories About Global Climate Change: Measuring the Impact of Narratives on Issue Cycles. *Communication Research, 26*(1), 30–57. doi:https://doi.org/10.1177/009365099026001003

Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013). Improving LDA topic models for microblogs via tweet pooling and automatic labeling. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '13)* (pp. 889–892). New York, NY, USA: Association for Computing Machinery. doi:https://doi.org/10.1145/2484028.2484166

Miao, L., Last, M., & Litvak, M. (2020). Twitter Data Augmentation for Monitoring Public Opinion on Covid-19 Intervention Measures. *Proceedings of the 1st Workshop on NLP for Covid-19 (Part 2) at EMNLP 2020*. Association for Computational Linguistics. doi:https://doi.org/10.18653/v1/2020.nlpcovid19-2.19

Mohammad, S. M., Salameh, M., & Kiritchenko, S. (2016). How translation alters sentiment. *Journal of Artificial Intelligence Research, 55*, 95-130. doi:https://doi.org/10.1613/jair.4787

Mollema, L., Harmsen, I., Broekhuizen, E., Clijnk, R., De Melker, H., Paulussen, T., . . . Das, E. (2015). Disease Detection or Public Opinion Reflection? Content Analysis of Tweets, Other Social Media, and Online Newspapers During the Measles Outbreak in the Netherlands in 2013. *Journal of Medical Internet Research, 17*(5). doi:https://doi.org/10.2196/jmir.3863

Mundt, M., Ross, K., & Burnett, C. M. (2018). Scaling Social Movements Through Social Media: The Case of Black Lives Matter. *Social Media + Society, 4*(4). doi:https://doi.org/10.1177/2056305118807911

Mustafaraj, E., Finn, S., Whitlock, C., & Metaxas, P. T. (2011). Vocal minority versus silent majority: discovering the opinions of the long tail. *Proceedings of SocialCom/PASSAT*, (pp. 103–110). Boston, MA, USA.

Nandathilaka, M., Ahangama, S., & Weerasuriya, G. (2018). A Rule-based Lemmatizing Approach for Sinhala Language. *2018 3rd International Conference on Information Technology Research (ICITR)*, (pp. 1-5). doi:https://doi.org/10.1109/ICITR.2018.8736134

Natuurmonumenten. (2018). *Europees Hof: Nederlands stikstofbeleid moet veel beter*. Retrieved November 20, 2021, from Natuurmonumenten: https://www.natuurmonumenten.nl/nieuws/europees-hof-nederlands-stikstofbeleid-moet-veel-beter

Negara, E. S., Triadi, D., & Andryani, R. (2019). Topic Modelling Twitter Data with Latent Dirichlet Allocation Method. *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, (pp. 386-390). doi:https://doi.org/10.1109/ICECOS47637.2019.8984523

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 100–108). Los Angeles, California, USA: Human Language Technologies.

NOS. (2019). *Raad van State: Nederlandse aanpak stikstof deugt niet*. Retrieved November 20, 2021, from NOS: https://nos.nl/collectie/13799/artikel/2286818-raad-van-state-nederlandse-aanpak-stikstof-deugt-niet

NOS. (2019a). *De stikstofcrisis, hoe heeft de politiek het zo ver laten komen?* Retrieved August 26, 2021, from NOS: https://nos.nl/collectie/13799/artikel/2304768-de-stikstofcrisishoe-heeft-de-politiek-het-zo-ver-laten-komen

NOS. (2019b). *Kabinet wil boeren uitkopen en op aantal wegen minder hard rijden.* Retrieved August 26, 2021, from NOS: https://nos.nl/collectie/13799/artikel/2304681-kabinet-wil-boeren-uitkopen-en-op-aantal-wegen-minder-hard-rijden

NOS. (2020). *Burgemeester Zwolle wil landelijke regels voor boerenprotesten: 'Grens overschreden'*. Retrieved October 1, 2021, from NOS: https://nos.nl/artikel/2360505-burgemeester-zwolle-wil-landelijke-regels-voor-boerenprotesten-grens-overschreden

NOS. (2021). *Nieuw boerenprotest, landbouw in bepaalde gebieden verder in het nauw.* Retrieved October 1, 2021, from NOS: https://nos.nl/artikel/2388278-nieuw-boerenprotest-landbouw-in-bepaalde-gebieden-verder-in-het-nauw

Omar, M., On, B.-W., Lee, I., & Choi, G. S. (2015). LDA topics: Representation and evaluation. *Journal of Information Science, 41*(5), 662–675. doi:https://doi.org/10.1177/0165551515587839

Omnicore Agency. (2021). *Twitter by the Numbers: Stats, Demographics & Fun Facts.* Retrieved November 21, 2021, from Omnicore Agency: https://www.omnicoreagency.com/twitter-statistics/

Omroep West. (2020). *Twee aanhoudingen tijdens boerenprotest in Den Haag, boeren rijden toeterend langs Huis ten Bosch.* Retrieved October 1, 2021, from Omroep West: https://nos.nl/artikel/2360505-burgemeester-zwolle-wil-landelijke-regels-voor-boerenprotesten-grens-overschreden

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)* (pp. 79–86). Association for Computational Linguistics. doi:https://doi.org/10.3115/1118693.1118704

Paul, M. (2009). Cross-collection topic models: Automatically comparing and contrasting text. *Urbana, 51*, 61801.

Raad van State. (2019). *PAS mag niet als toestemmingsbasis voor activiteiten worden gebruikt.* Retrieved August 26, 2021, from Raad van State: https://www.raadvanstate.nl/@115651/pas-mag/

Rao, P., & Taboada, M. (2021). Gender Bias in the News: A Scalable Topic Modelling and Visualization Framework. *Frontiers in artificial intelligence, 4*, 664737. doi:https://doi.org/10.3389/frai.2021

Řehůřek, R. (n.d.). *How to Compare LDA Models.* Retrieved September 7, 2021, from Radim Rehurek: https://radimrehurek.com/gensim_3.8.3/auto_examples/howtos/run_compare_lda.html#sphx-glr-auto-examples-howtos-run-compare-lda-py

Řehůřek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3*(2).

Rijksoverheid. (2021). *Aanpak Stikstof Landbouw*. Retrieved October 26, 2021, from Aanpak Stikstof: https://www.aanpakstikstof.nl/themas/landbouw on 24th August 2021

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association, 111*(515), 988–1003. doi:https://doi.org/10.1080/01621459.2016.1141684

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S., & Rand, D. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Scienc, 58*(4), 1064-1082. doi:https://doi.org/10.1111/ajps.12103

Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin III, F., Lambin, E., & Foley, J. (2009). A safe operating space for humanity. *Nature, 461*, 472-475. doi:https://doi.org/10.1038/461472a

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the eighth ACM international conference on web search and data mining (WSDM '15)* (pp. 399-408). New York, NY, USA: ACM.

Rothfels, J., & Tibshirani, J. (2010). Unsupervised sentiment classification of English movie reviews using automatic selection of positive and negative sentiment items. *CS224N-Final Project, 43*(2), 52-56.

RTL Nieuws. (2019). *Boerenprotest op 18 december: 'Genoeg gepraat'*. Retrieved September 10, 2021, from RTL Nieuws: https://www.rtlnieuws.nl/nieuws/nederland/artikel/4950316/boerenprotest-boerenacties-stikstofbeleid-supermarkt

RTL Nieuws. (2019). *Snap je ook niets meer van de stikstofcrisis? Dit is hoe het zit*. Retrieved November 20, 2021, from RTL Nieuws: https://www.rtlnieuws.nl/economie/artikel/4906081/stikstof-snap-niet-meer-dit-hoe-het-zit-crisis-probleem-oplossing

RTL Nieuws. (2020). *Boeren met trekkers toch op snelweg: 'Never give up'*. Retrieved October 1, 2021, from RTL Nieuws: https://www.rtlnieuws.nl/nieuws/nederland/artikel/5027231/boerenprotest-tractor-trekker-snelweg

Schelfaut, S. (2019a). *Stikstofplan ingetrokken na felle acties van boeren*. Retrieved October 1, 2021, from Algemeen Dagblad: https://www.ad.nl/binnenland/stikstofplan-ingetrokken-na-felle-acties-van-boeren~a0815b9d/

Schelfaut, S. (2019b, December 16). *Rutte: Holocaust-vergelijking indringend besproken tijdens boerenoverleg Catshuis*. Retrieved from AD: https://www.ad.nl/politiek/rutte-holocaust-vergelijking-indringend-besproken-tijdens-boerenoverleg-catshuis~add5d7b4/

Scheufele, D. A. (2006). Framing as a theory of media effects. *Journal of Communication, 49*(1), 103-122. doi:https://doi.org/10.1111/j.1460-2466.1999.tb02784.x

Schöne, J., Parkinson, B., & Goldenberg, A. (2021). Negativity Spreads More than Positivity on Twitter after both Positive and Negative Political Situations. doi:https://doi.org/10.31234/osf.io/x9e7u

Schreuder, A. (2019). *'Stikstofcrisis was niet nodig, Nederland is te voorzichtig'*. Retrieved November 21, 2021, from NRC: https://www.nrc.nl/nieuws/2019/11/14/crisis-met-stikstof-was-niet-nodig-a3980368?utm_source=NRC&utm_medium=banner&utm_campaign=Paywall&utm_content=paywall-mei-2019

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin, 86*(2), 420.

Spohr, D. (2017). Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review, 34*(3), 150–160. doi:https://doi.org/10.1177/0266382117722446

Statista. (2021a). *Population of the Netherlands in 2020, by age*. Retrieved November 21, 2021, from Statista: https://www.statista.com/statistics/519754/population-of-the-netherlands-by-age/

Statista. (2021b). *Share of respondents using Twitter in the Netherlands in 2017 and 2018, by age group*. Retrieved November 21, 2021, from Statista: https://www.statista.com/statistics/828876/twitter-penetration-rate-in-the-netherlands-by-age-group/

Steinskog, A. O., Therkelsen, J. F., & Gambäck, B. (2017). Twitter Topic Modeling by Tweet Aggregation. *Proceedings of the 21st Nordic Conference of Computational Linguistics*, (pp. 77-86).

Szmigiera, M. (2021). *The most spoken languages worldwide in 2021(by speakers in millions)*. Retrieved September 7, 2021, from Statista: https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/

Thambinathan, V., & Kinsella, E. A. (2021). Decolonizing Methodologies in Qualitative Research: Creating Spaces for Transformative Praxis. *International Journal of Qualitative Methods, 20*. doi:https://doi.org/10.1177/16094069211014766

Thelwall, M. (2018). Gender bias in sentiment analysis. *Online Information Review, 42*(1), 45-57.

Thelwall, M., Buckley, K., Paltoglou, G. C., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology, 61*(12), 2544–2558.

Traniello, J., & Bakker, T. (2015). Minimizing observer bias in behavioral research: blinded methods reporting requirements for Behavioral Ecology and Sociobiology.

Behavioral Ecology and Sociobiology, 69, 1573–1574. doi:https://doi.org/10.1007/s00265-015-2001-2

Trilling, D., & Boumans, J. (2018). Automatische inhoudsanalyse van Nederlandstalige data: Een overzicht en onderzoeksagenda. *Tijdschrift voor Communicatiewetenschap, 46*(1), 5-24.

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM, 10*(1), 178-185.

Turney, P. D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, (pp. 417-424). Philadelphia. doi:https://doi.org/10.3115/1073083.1073153.

Twitter Inc. (2021). *Permanent suspension of @realDonaldTrump*. Retrieved September 2, 2021, from Twitter Inc.: https://blog.twitter.com/en_us/topics/company/2020/suspension

Twitter. (n.d.). *Tweet Object*. Retrieved November 21, 2021, from Twitter Developer Platform: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet

Van Aelst, P., & Walgrave, S. (2011). Minimal or Massive? The Political Agenda-Setting Power of the Mass Media According to Different Methods. *The International Journal of Press/Politics, 16*(3), 295–313. doi:https://doi.org/10.1177/1940161211406727

Van Winzum, L. (2021). (Master thesis) Sensatiezoekers - Een vergelijkend computationeel onderzoek naar het sensationeel taalgebruik in de koppen van De Telegraaf en NRC Handelsblad. Universiteit Utrecht.

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems, 94*, 101582. doi:https://doi.org/10.1016/j.is.2020.101582

Vazquez, M., Hickey, C., Krishnakumar, P., & Boschma, J. (2020). *Donald Trump's presidency by the numbers*. Retrieved September 2, 2021, from CNN International: https://edition.cnn.com/2020/12/18/politics/trump-presidency-by-the-numbers/index.html

Visser, J. (2020). *Overspeelt Farmers Defence Force zijn hand in woord en daad?* Retrieved November 22, 2021, from Metro: https://www.metronieuws.nl/in-het-nieuws/2020/02/overspeelt-farmers-defence-force-zijn-hand-in-woord-en-daad/

Waikhom, L., & Goswami, R. S. (2019). Fake News Detection Using Machine Learning. *SSRN Electronic Journal*, 1076–2787. doi:https://doi.org/10.2139/ssrn.3462938

Wang, S., Schraagen, M., Tjong Kim Sang, E., & Dastani, M. (2020). Dutch General Public Reaction on Governmental Covid-19 Measures and Announcements in Twitter Data. Retrieved from https://arxiv.org/abs/2006.07283

Wijnants, R. (2020). *Minister Schouten staakt werkbezoek Zeeland wegens dreiging boeren.* Retrieved September 8, 2021, from Joop - BNNVARA: https://joop.bnnvara.nl/nieuws/minister-schouten-staakt-werkbezoek-zeeland-wegens-dreiging-boeren

Wikipedia. (n.d.). *Boerenprotesten Nederland 2019–heden.* Retrieved October 1, 2021, from Wikipedia: https://nl.wikipedia.org/wiki/Boerenprotesten_Nederland_2019%E2%80%93heden#18_december_2019

Winterman, P. (2019). *Voorman boeren moet lachen om grafkist met Jesse Klavers naam erop.* Retrieved August 26, 2021, from AD: https://www.ad.nl/politiek/voorman-boeren-moet-lachen-om-grafkist-met-jesse-klavers-naam-erop~a83f7cac/

Woodward, J. L. (1934). Quantitative Newspaper Analysis as a Technique of Opinion Research. *Social Forces, 12*(4), 526–537. doi:https://doi.org/10.2307/2569712

Zhai, C., & Massung, S. (2016). *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining.* Association for Computing Machinery and Morgan & Claypool. doi:https://doi.org/10.1145/2915031.2915033

Zhang, W., Xu, M., & Jiang, Q. (2018). Opinion Mining and Sentiment Analysis in Social Media: Challenges and Applications. *HCI in Business, Government, and Organizations : 5th International Conference, HCIBGO 2018 Held as Part of HCI International 2018 Las Vegas, NV, USA, July 15–20, 2018 Proceedings. 10923*. Cham: Springer. doi:https://doi.org/10.1007/978-3-319-91716-0_43

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. *ECIR'11* (pp. 338–349). Berlin, Heidelberg: Springer Berlin Heidelberg.