Bachelor thesis

Fake news during the Dutch elections: automatic classification and government strategies

Leiden University

Student: Viola Braams (s1953923)

Study: Computer Science and Economics

Supervisors: Olga Gadyatskaya & Suzan Verberne (LIACS)

Date: 25-09-2021

Abstract

This thesis explores the current policy strategies regarding combatting fake news and to what extend automated methods can help in detecting fake news on Twitter. This all with the focus on election times where fake news threatens the basic functioning of the democratic state. A small-scale machine learning study is conducted involving tweets related to the 2021 Dutch elections. To implement the text classification experiment, a collection of tweets was selected and annotated according to pre-defined categories. The aim is to get an understanding to what extent text features can help determine whether a Dutch tweet is true or false. The results show that it is a difficult task to determine whether a tweet is true or false based on the content. Yet, the European approach and their policy strategy describe the importance to invest in new technologies and further research the ways to detect fake news. It is a promising field of study as even with a small dataset, the Support Vector Machine was able to correctly classify 58% of the tweets. Data mining techniques are a useful tool in the detection of fake news on social media and can, with further research and improvements, increasingly play a role in protecting democratic states from the spread of misinformation.

Table of Contents

1. Introduction

In the digitalized world, beliefs, thoughts, and attitudes are influenced and formed by the way people communicate through social media and via the internet. Actors are present who vigorously engage in influence operations on the internet, with the sole goal to "disrupt civil discourse, sew discord and spread disinformation" (Carley, 2020, p. 365). This is a direct threat to democracy, formed in the cyber space. Hence, the computational social cybersecurity discipline emerged to describe, recognize, and predict cyber-mediated shifts in human behavior and take the effects on culture, politics, and human interaction into consideration; and to shape an infrastructure where civilization sustains in this new environment that is illustrated by ever fluctuating circumstances revolving cyber threats (Carley, 2020, p. 366). Social cybersecurity is different from cybersecurity as it focusses on the humans and the ways they are compromised, transformed, and referred to the unimportant (Carley, 2020, p. 367). Disinformation and misinformation are currently dominant topic areas of social cybersecurity, and the phenomenon was aggravated during the coronavirus pandemic and election period (Carley, 2020, p. 368). "Bots, trolls, cyborgs and humans engage with others in cyberspace in ways designed to, and send messages that are constructed to, take advantage of three things: the technology, the mind and emotions, and the world view" (Carley, 2020, p. 372). The Dutch National Coordinator for Security and Counterterrorism also stresses that disinformation is a source of distrust. They mention that due to the online environment the threshold to spread misinformation has lowered and makes it possible to flow into the mainstream channels (Ministerie van Justitie en Veiligheid, 2020).

Social media is used to create and share content by the users and help people connect to each other. Traditional media are getting less popular as the younger generation turns to social media as their prime source of news (Newman, Fletcher, Kalogeropoulos, Levy, & Kleis Nielsen, 2018). This expose is constant and serves them with broad ranges of opinions and information (Flintham, Karner, Creswick, Bachour, Gupta, & Moran, 2018, p. 1). This is also part of a phenomenon that is called the 'post-truth' era, which can be defined as situations where objective facts are less significant in determining public opinion than appeals to sentiment and personal beliefs (Flintham et al., 2018, p.1). The term "truthiness" has been defined as "the quality of stating concepts or facts one wishes or believes to be true, rather than concepts or facts known to be true" (Munger, 2017). In our modern society and the digital age, this sentiment is often seen to be of more importance than real, verifiable content. Information overload is a trend which follows from the fact that the number of platforms for

news consumption has grown. This means that with the increase of the amount of information, consumers have a larger task with determining whether information they face on these platform is correct (Auxier & Vitak, 2019, p. 1). When asked about the level of trust in media, three-quarters of Americans answered, “not very much” or “almost none” (Munger, 2017). This trend is bad for the functioning of democracy, as having access to the right information is of crucial importance. This is intensified by the creation of “echo chambers” and “filter bubbles”, which are the result of the rise of algorithmic filtering of content (Auxier & Vitak, 2019, p. 1). This means that when a user repeatedly follows, “likes”, retweets, or has other digital interactions with conservative, right-winged news, the algorithm will tend to surface more of this same sort of news. This results in a digital news environment which only explores a limited perspective (Auxier & Vitak, 2019, p. 1). Consumers online tend to prefer information sticking to their opinions, disregard dissenting information, and form polarized groups around mutual views (Auxier & Vitak, 2019, p. 3). Furthermore, when the level polarization is high, misinformation rapidly grows (Cinelli, Morales, Galeazzi, Quattrociocchi, & Starnini, 2021, p. 7).

Within democracies, fake news has thrived in contemporary political climates, creating misinformation on social media platforms. Lee (2019) states that misinformation “has served to diminish the credibility of mainstream news networks, dividing the public further, both ideologically and on the mere acceptance of the fact, providing credence to ideological claims of fake news” (p. 16). Online manipulation and disinformation have been used in at least eighteen countries during elections in recent years (Freedom House, 2017). Disinformation erodes trust in institutions and in digital and traditional media. It also harms our democracies by obstructing the ability of citizens to take informed decisions (European Commission, 2018, p. 1). The 2016 United States presidential elections are known for the role fake news played over the whole of the election period (Grinberg, Joseph, Friedland, Swire-Thompson, & Lazer, 2019, p. 1). Social media created platforms for the distribution and consumption of news and therefore also for the distribution and consumption of fake news. This online environment basically adhered to the increasing division of politics. Filter bubbles were active and factually fake news was used as a method for political gain (Lee, 2019, p. 20). The most damaging result of fake news is the negative effects on liberal democracy. Overall, all what a liberal democracy entails, including its political processes, is centered around reliable information (Lee, 2019, p. 20). With the extensive reproduction of fake news, this crucial foundation of reliable knowledge is at risk. Consequently, people are not always able to shape well-thought-out opinions and thus make rational political choices (Lee, 2019,

p. 20). This is of special importance during times of elections, as this is the time the public must choose their representatives within the government. Hence, there is a need to develop means to combat fake news as a way of safeguarding the functioning of democratic processes like elections. However, it is challenging to combat fake news manually, given the scale of the problem and the overflow of news items on large social platforms. Thus, automated methods to detect fake news are required. For this thesis the goal is to investigate automated methods that can help detect fake news on Twitter in relation to the Dutch elections. The first research question is focused on the societal perspective:

- *What are the strategies to combat fake news, especially during the Dutch elections in 2021?*

To answer this question this thesis explores the current policy strategies regarding combatting fake news. Moreover, the second question addresses the automated methods to detect fake news:

- *To what extent can automated methods help detect fake news on Twitter in relation to the Dutch elections?*

To answer these questions, the related work section will start to introduce the topic of fake news and its connection to the elections and the current approach to combat fake news. There will also be an analysis of the research domain regarding automated methods for fake news detection. This leads to a text classification experiment that aims to predict whether a tweet is spreading true or false news. Six topics from a fact-checking website are used that relate to the Dutch elections of 2021. These topics will generate queries to find corresponding tweets via the snscrape tool and will help to put each tweet into one category of news. This is a manual annotation process mainly done by the thesis student. This will result in a text mining experiment where text features will be used to try to get an understanding into the automated detection of fake news.

2. Related work

2.1. Defining fake news

Fake news can be defined as news articles that are purposely and verifiably false and might consequently mislead readers (Alcott & Gentzkow, 2017, p. 213). Cyber security in the media space is important, and fake news has arisen as a cyber-attack mechanism. Fake news is not always spread with purely malicious goals in mind, as they can also be distributed to attract attention and increase advertisement revenue. Financial and ideological motivations are the

two factors that cause the production of fake news (Tandoc, Lim, & Ling, 2018, p. 138). Advertisement revenue is collected through the clicks the articles gather from their mostly striking fake claims. Also, fake news suppliers discredit opposing views through the spread of fake claims and to promote a certain ideology. This paper will use the term *fake news*, which includes news that is false, but also news that is misleading. Misleading news means that correct facts are used to lead to incorrect conclusions. The focus is on the political perspective, as Alcott and Gentzkow (2017) include purposefully fabricated news articles and "many articles that originate on satirical websites but could be misunderstood as factual, especially when viewed in isolation on Twitter or Facebook feeds" (p. 213). A distinction can be made between misinformation and disinformation. Misinformation is "false by definition" and disinformation can be seen as a subset of misinformation, where the intent clearly is to misinform (Guess & Lyons, 2020, p. 11). Misinformation can be of unintentional nature or spread by accident as the person is not aware of its false message (Guess & Lyons, 2020, p. 11). For the purpose of this thesis the term fake news will be used which can mean both misinformation, if the person is not aware of the false claims in their tweets, or disinformation, where the person is intentionally tweeting out false claims.

2.2. Fake news as a weapon in the political arena

The reasons for people to consume news through social media are that it is at a low cost, easy to access and due to its rapid distribution of information (Shu, Sliva, Wang, Tang, & Liu, 2017, p. 22). Through social media, one can easily share news, comment on topics, and discuss them with others. "The proliferation and viral spread of fake news - false information passed off as factual – is a global problem, accelerated by information and communications technology that enables near-instant and easily disguised messaging" (Guadagno & Guttieri, 2021, p. 167). For the 2016 presidential elections, Donald Trump, the candidate for the Republicans, used the term 'fake news' to discredit the traditional media, but also used the label to promote false reports of less reputable, but friendly sources (Guadagno & Guttieri, 2021, p. 168). Foreign powers also use fake news as a tool of psychological warfare, as they are trying to shape politics of another country (Guadagno & Guttieri, 2021, p. 172). According to reports from the intelligence community of the United States, the Russian government has used "an information campaign disseminated by various means, including paid human "trolls" posting provocative or divisive comments, and software or bots to reproduce and spread information" (Guadagno & Guttieri, 2021, p. 174).

Concerning the security of elections, the European approach pronounces the necessity of particular attention to election processes, as they lie in the basis for democratic functioning (European Commission, 2018, p. 11). “Disinformation now forms part of a wider array of tools used to manipulate electoral processes, such as hacking or defacing websites or gaining access to and leaking personal information about politicians” (European Commission, 2018, p. 11). Disinformation has turned into a cyber-attack mechanism. This is critical during election times, where busy and hectic agendas may prevent well-timed detection of disinformation and the proper reaction to stop it from spreading (European Commission, 2018, p. 11).

2.3. Fake news detection on social media

As Shu et al. (2017) state: “to help mitigate the negative effects caused by fake news – both to benefit the public and the news ecosystem – It’s critical that we develop methods to automatically detect fake news on social media” (p. 80). However, this poses new challenges for the research domain of automated detection. Fake news is purposefully written to mislead readers, which makes it hard to detect solely based on news content. “The content of fake news is rather diverse in terms of topics, styles and media platforms, and fake news attempts to distort truth with diverse linguistic styles while simultaneously mocking true news” (Shu et al., 2017, p. 81). Besides, the quality of the data itself for researching this is mostly unstructured, big, and noisy. “Effective methods to differentiate credible users, extract useful post features and exploit network interactions are an open area of research and need further investigations” (Shu et al., 2017, p. 81). The fundamental assumption of these methods is that fake news may involve specific keywords, or groupings of keywords, hence a post with sufficient fake news warnings can be classified (Wu, Morstatter, Carley, & Liu, 2019, p. 84).

Different focuses in the automated detection domain can be distinguished, as for the context of this research, the feature-oriented and model-oriented approaches are the most relevant. “Feature-oriented fake news research aims to determine effective features for detecting fake news from multiple data sources” (Shu et al., 2017, p. 90). Moreover, model-oriented approaches also use features to learn models and concerning fake news detection: “fake news research opens the door to building more effective and practical models for fake news detection” (Shu et al., 2017, p. 91). Most methods focus on obtaining numerous features, then merging these features into supervised classification models, and finally choosing the classifier that performs the best (Shu et al., 2017, p. 91).

Zhang and Ghorbani (2020) conducted recent research into fake news and its characterization, detection, and discussion. They also stress that fake news detection is a rising topic and technical solutions are being sought to determine whether a post is fake or not. Yet, they state that “accurate fake news detection, is still challenging, due to the dynamic nature of the social media, and the complexity and diversity of online communication data” (Zhang & Ghorbani, 2020, p. 3). They showcase that on the aspect of fact-checking, that full automate fake news detection has a long way to go and some popular websites still only depend on manual checks. Nevertheless, they claim that it is crucial to develop automatic detection methods (Zhang & Ghorbani, 2020, p. 8). News content analysis is one way Zhang and Ghorbani (2020) proclaim to help detect fake news. They say that "by extracting useful information from the news content, linguistic and semantic analysis can analyze the associate language patterns, structures and meanings of the news” (Zhang & Ghorbani, 2020, p. 13). From a data mining perspective, they describe supervised machine learning algorithms to be broadly used in previous literature regarding hoaxes, frauds, etc. (Zhang & Ghorbani, 2020, p. 15). For looking at word-level features, “bag-of-words, n-gram, term frequency (TF), term frequency-inverted document frequency (TF-IDF) are the most commonly used linguistic features for natural language processing” (Zhang & Ghorbani, 2020, p. 17). Research into the 2019 Indonesian elections and fake news on Twitter, by Suryadikara (2020), creates a text mining experiment to investigate the possibility to detect fake news. Topics regarding the Indonesian election are selected from fact-checking websites and accordingly the twitter data is gathered and annotated (Suryadikara, 2020, pp. 15-16). The author has used the Multinomial Naive Bayes (MNB), Support Gradient Descent (SGD), and Random Forest (RF) classifiers from the Python Sklearn module (Suryadikara, 2020, p. 15). Concluding the research, “the most prominent text feature to detect and distinguish true news, false news, and misleading news is word n-grams, particularly any composition that includes unigram” (Suryadikara, 2020, p. 56).

3. Method

To answer the research questions, a small-scale machine learning study is conducted involving tweets related to the 2021 Dutch elections. To implement the text classification experiment, a collection of tweets was selected and annotated according to pre-defined categories. The aim is to get an understanding to what extent text features can help determine

whether a Dutch tweet is true or false. This is done through text classification with the Python Sklearn module.

3.1. Twitter

Social network sites (SNSs), like Twitter, have millions of users, who integrate the interaction with these sites into their daily lives (Boyd & Ellison, 2007, p. 210). These social network sites are defined "as web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system" (Boyd & Ellison, 2007, p. 211). Twitter is a popular social network site with 192 million active users per day (Lin, 2021). Here, people can participate in the form of microblogging, where users make status posts, called *tweets*, with a maximum of 280 characters (Go, Huang, & Bhayani, 2009, p. 1). Twitter is an ideal platform for news channels due to its ability to send immediate updates. Moreover, journalists make up 24.6% of verified accounts on this platform (Osman, 2020).

3.2. Data Collection

For the collection of tweets, the snscrape tool[1] is used. This tool is mainly needed to bypass the limitations of the official Twitter API, which largely revolves around the fact that older tweets cannot be found. This tool is a scraper for social networking services, and it scrapes for users, hashtags, or query terms and returns the relevant posts. This enables the targeted search of historical Twitter data for a particular query. The queries that are used for this research are based on topics published by the fact-checking website of Leiden University.[2] These topics all relate to political topics, so statements made by politicians, about politicians, or the organization of political parties. The selected time frame starts on the 30th of December 2020, which is the day when all participating parties participating in the upcoming elections were announced. The end is the 26th of March as this is the day that the determination of the election results took place. I selected six topics regarding the Dutch presidential elections from the nieuwscheckers website. As Table 1 shows, each topic corresponds to a URL from the nieuwscheckers website and using the snscrape tool a number of tweets was collected that matches each topic. The queries used are displayed in appendix A.

[1] https://github.com/JustAnotherArchivist/snscrape
[2] http://nieuwscheckers.nl

Topic	**URL**	**Date**	**Number of Tweets**
Nederlandse staatsschuld relatief lager dan in de jaren na de kredietcrisis	https://nieuwscheckers.nl/nieuwscheckers/factcheck-hoekstra-staatsschuld/	22/01/2021	147
Nieuws over dreigende corona-burnout bij 80 procent van jongeren is ‘totale quatsch’	https://nieuwscheckers.nl/nieuwscheckers/nieuws-over-dreigende-corona-burnout-bij-80-procent-van-jongeren-is-totale-quatsch/	16/02/2021	216
Koopkracht gepensioneerden veel minder hard gedaald dan 50Plus beweert	https://nieuwscheckers.nl/nieuwscheckers/koopkracht-gepensioneerden-veel-minder-hard-gedaald-dan-50plus-beweert/	18/01/2021	121
Grote hoeveelheid stikstof wél schadelijk voor natuur	https://nieuwscheckers.nl/nieuwscheckers/grote-hoeveelheid-stikstof-wel-schadelijk-voor-natuur/	03/02/2021	53

Tijdens tien jaar Rutte zijn 750 duizend niet-westerse migranten naar Nederland gekomen. 470 duizend hebben het land verlaten	https://nieuwscheckers.nl/nieuwscheckers/tijdens-tien-jaar-rutte-zijn-750-duizend-niet-westerse-migranten-naar-nederland-gekomen-470-duizend-hebben-het-land-verlaten/	12/03/2021	405
Huisartsen hielden COVID-patiënten niet uit ziekenhuis vanwege 'Code Zwart', zoals Geert Wilders beweerde	https://nieuwscheckers.nl/nieuwscheckers/huisartsen-hielden-covid-patienten-niet-uit-ziekenhuis-vanwege-code-zwart-zoals-geert-wilders-beweerde/	26/01/2021	349

Table 1 Nieuwscheckers topics

3.3. Annotation strategy

The thesis student has performed the largest task of annotation. This person did not have any political affiliation or does not belong to a political party. One other person has annotated a small set, one topic with approximately 200 Tweets, with the purpose of estimating inter-rater reliability. All topics are linked to one supporting URL from the Leiden University fact-checking website. All tweets have been selected per topic, based on queries that correspond to the information discussed and/or claims made. The column that is used for annotation is "categoryOfNews". Here, the annotator choses the category in which the tweet belongs. The annotator is tasked to put each tweet into a category based on the definitions from the research of Suryadikara (2020):

- **True**: Tweets that relate to the topic and are true or accurate according to supporting URLs.
- **False**: Tweets that relate to the topic and are false or incorrect according to supporting URLs.

- **Misleading**: Tweets that relate to the topic and have correct information according to supporting URLs but cause wrong conclusions.
- **Neutral**: Tweets that relate to the topic but do not mention the precise fact-checked essence of information from the supporting URLs in any form (neither true nor false statements).
- **Other**: Tweets that do not relate to the topic or are not discussed within supporting URLs.

Before examining the tweets, the annotator is asked to read the attached URL concerning the topic of the tweet. This URL will be all the annotator needs to use to verify the tweet and put it into a category. First the annotator checks whether the tweet is a statement or claim mentioned in the description of the supporting URL. The annotator needs to limit their work to the claims or statements that are related to the topic or discussed in the supplied URL. This is done with the aim to stay focused on the topics selected.

An example of each category is given to guide the annotator. The examples are given from the annotation that the thesis student has done concerning the topic "gepensioneerden". The main statement concerning this topic was that over the past 10 years, pensioners have lost about 20 percent of their purchasing power. This was fact-checked, and the conclusion is that this statement is false. A collection of 121 Tweets were gathered concerning this topic and all were put into one of the mentioned categories. Each category will be supported with an example. One example of a false news tweet is:

- @VVD Gepensioneerden optimistisch? 25% koopkracht afgejat door Rutte en zijn trawanten.

This information is False, as the supporting link from the fact-checking website explains that the average loss in purchasing power is not near the 25% mentioned in this Tweet.

The figures below show the distribution of the tweets per category of news.

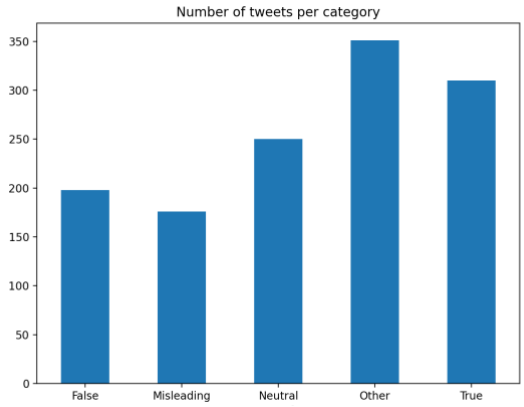

Figure 1 Tweets per Category of News

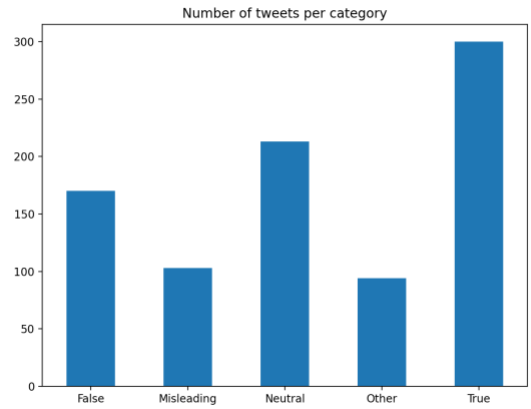

Figure 2 Tweets per Category of News excluding "migratie"

Figures 1 and 2 show the distribution of tweets for each category. Where the “migratie” topic was excluded in the second figure, this shows a less even distribution for the misleading category. Yet, the ‘other’ category has significant more tweets in the first figure, where all topics are included. The migration topic was harder to annotate due to a lot of other subtopics surrounding migration that were active during the election times. Moreover, this news topic was a bit controversial as the fact-checking article stated that the claim that was made was true in nature, but the politician framed it to be very misleading. However, as misleading news is part of the definition of fake news in this research, this news topic will be included in the experiment. Following the annotation process, an agreement table is presented in Table 1.

Agreement Table		Annotator					
		True	**False**	**Misleading**	**Neutral**	**Other**	Total
Thesis student	**True**	26 (7.17)	0	1	1	0	28
	False	0	26 (6.02)	0	1	0	27
	Misleading	0	1	5 (0.50)	0	0	6
	Neutral	5	0	4	28 (12.84)	5	42
	Other	0	0	0	7	11 (2.38)	18
	Total	31	27	10	37	16	121

Table 2 Agreement Table

The inter-rater agreement is important to know the reliability of the data. The metric is Cohen’s Kappa and following the annotation done by the thesis student and annotator the value of the metric is 0.73, which falls into the category of substantial agreement.

3.4. Text classification experiment

To perform the text classification experiment, the datasets of each topic were combined into a pandas data frame. To execute the text classification, the Python scikit-learn module was used for the pre-processing as well as the classification tasks. The pre-processing of the tweets involved some text cleaning, where all symbols and digits were removed to create a generalizable model. Also, all letters were transformed to lowercase. Next, a train-test split

was made where 80 percent of the tweets were used for training, and 20 percent for testing. The built-in function in scikit-learn was used to make sure this split is chosen randomly. Moreover, *TfidfTransformer* (TF-IDF) was used, in combination with the *CountVectoriser* (Pellarolo, 2018). The purpose was to split the text of the tweets up into words (tokens). These words were kept in their true form. Furthermore, the frequency of the words is calculated and assigned a weight proportional to its frequency. According to Suryadikara (2020), word n-grams are the most dominant features applied in text classification. This research uses unigrams for the detection of fake news. All topics are taken together to form one large badge of tweets where the classification task is done. As the topics differ in number of tweets, the decision was made to take them all together to aim for the best prediction. Due to the small dataset, the classification models that were used are: Multinomial Naive Bayes (MNB), Support Vector Machines (SVM), and Random Forest (RF). The implementation was done in scikit-learn and the hyperparameter settings were kept at their default values. The built-in metrics report of scikit-learn is used to compare these classifiers and the precision and recall scores are used to calculate the F1-scores. The overall accuracy is also displayed in the table in the analysis and results section. Additionally, the confusion matrixes are calculated using scikit-learn to display what category was predicted compared to the actual category of the tweet. This will give insight into what categories are often confused. Finally, the feature importance of the best performing model will be calculated. Looking at the top hundred features that play the largest role in the prediction can give an understanding of whether some words are overall important or whether no generalizable conclusions can be made because the words correspond to the dominant news topic.

4. Analysis & Results

First, a content analysis of policy documents is conducted to answer the first research question regarding the current strategies to combat fake news. The following analysis is of the experiment in which three different classifiers are compared on their metric scores and confusion matrixes. Also, a comparison is made between the distinction of the false news and misleading categories and the fake news category, where both false and misleading news are taken together as one. Moreover, we perform an analysis of the feature importance to see whether the most important features correspond to a certain topic or other typical characteristic in the content of the tweets. The results are presented in tables.

4.1. Policy strategies to combat fake news and safeguard elections

Dutch policy, as found on the website of Rijksoverheid[3], states that the Dutch government wants citizens themselves to be able to detect fake news. Yet, the only mention of policymaking directs to the efforts of the European Union and its associated countries that are working together on an action plan against disinformation (Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, 2021). Their online approach stresses the importance of combatting fake news as "disinformation erodes trust in institutions and in digital and traditional media, and harms our democracies by hampering the ability of citizens to take informed decisions" (European Commission, 2018, p. 1). As the European Union is taking action to tackle disinformation online, they define three main drivers that help the spread of disinformation on social media. First, algorithms are used to prioritize the display of certain information, determined by the platform's business model (European Commission, 2018, p. 5). The action plan mentions that "by facilitating the sharing of personalized content among like-minded users, algorithms indirectly heighten polarization and strengthen the effects of disinformation" (European Commission, 2018, p. 5). Second, the digital advertisement business is based around clicks, which leads to the popularity of sensational and viral content (European Commission, 2018, p. 5). The final driver is revolved around technologies such as automated amenities that artificially magnify the spread of disinformation (European Commission, 2018, p. 5). Users themselves also play a significant role in the dissemination of disinformation online; The speed and volume of sharing content online, without prior verification, is ever growing and increases the risk of sharing disinformation (European Commission, 2018, p. 6).

The European approach states that:

> *Given the complexity of the matter and the fast pace of developments in the digital environment, the Commission considers that any policy response should be comprehensive, continuously assess the phenomenon of disinformation, and adjust policy objectives in light of its evolution.* (European Commission, 2018, p. 6)

It is a complex matter, and no single solution can be given. However, the Commission states that inactivity is not a possibility (European Commission, 2018, p. 6). Four objectives guide the action plan to tackle disinformation: transparency, diversity, credibility, and inclusive solutions (European Commission, 2018, p. 6). Important components of this approach are "adequate changes in platforms' conduct, a more accountable information ecosystem,

[3] https://www.rijksoverheid.nl/onderwerpen/desinformatie-nepnieuws/aanpak-desinformatie-en-nepnieuws

enhanced fact-checking capabilities and collective knowledge on disinformation, and the use of new technologies to improve the way information is produced and disseminated online" (European Commission, 2018, p. 7).

Fact-checking has emerged as an important link in the media value chain. It aids the verification of content and assessing the credibility based on facts and evidence (European Commission, 2018, p. 9). "Fact-checkers credibility depends upon their independence and their compliance with strict ethical and transparency rules" (European Commission, 2018, p. 9). However, the fact-checking activities are rising, various other facets of disinformation continue to be insufficiently analyzed and online platforms still give limited access to their data (European Commission, 2018, p. 9). The European approach states that academic researchers and fact-checkers need to put effort into gathering data and analyzing possible new developments regarding indicators for source transparency, recognizing and recording disinformation mechanisms that add to digital magnification, and other activities (European Commission, 2018, p. 9). Moreover, the Commission stresses the importance of new technologies. "Emerging technologies will further change the way information is produced and disseminated, but they also have the potential to play a central role in tackling disinformation over the longer term" (European Commission, 2018, p. 10). They mention Artificial Intelligence (AI) as a crucial factor in verifying, detecting, and labelling disinformation (European Commission, 2018, p. 11).

4.2. Metric Scores

The experiment's aim is to show whether these classifiers can help distinguish the categories of news from the content of the tweets. In the first case, all five selected categories of news are used in the classification models. The results are shown in Table 2. Here, the Support Vector Machine and Random Forest algorithm are both performing the same with an accuracy of 0.50. Yet, looking more closely at the F1 scores per category, the Random Forest algorithm is doing a slightly better job at predicting the false and misleading categories.

Category of News	Precision			Recall			F1-score			Support
	MNB	SVM	RF	MNB	SVM	RF	MNB	SVM	RF	
False	0.50	0.37	0.45	0.03	0.35	0.35	0.05	0.36	0.39	37
Misleading	1.00	0.44	0.77	0.07	0.47	0.33	0.12	0.45	0.47	30
Neutral	0.10	0.28	0.27	0.02	0.28	0.30	0.04	0.28	0.28	47
Other	0.40	0.67	0.60	0.85	0.63	0.66	0.55	0.65	0.63	68
True	0.49	0.56	0.53	0.64	0.59	0.62	0.56	0.58	0.57	76
Accuracy							0.43	**0.50**	**0.50**	258

Table 2 Evaluation of three text classifiers for the classification of tweets

As this thesis uses the term fake news consistently and has explained that both the false and misleading category of news together are a part of fake news, the experiment also takes together these two categories to form a fake news category. As the classification must be done on less categories, consequently the accuracy is expected to be higher when the combined classes are similar (false and misleading news), together with the fact that there are less categories to choose from (four instead of five). So, when looking at the metric scores in Table 3, all classification models perform better on the four categories. Moreover, there is an increase in the ability to predict the news categories as the Support Vector machine algorithm has an overall accuracy of 58% compared to the previous 50% where five categories were used. Yet, the overall accuracy is still low. The fake news category has an accuracy of 54%, where before the false (36%) and misleading (45%) categories had a substantial smaller score.

Category of News	Precision			Recall			F1-score			Support
	MNB	SVM	RF	MNB	SVM	RF	MNB	SVM	RF	
Fake	0.36	0.51	0.42	0.99	0.56	0.60	0.53	0.54	0.50	68
Neutral	1.00	0.43	0.48	0.02	0.37	0.26	0.04	0.40	0.34	54
Other	0.86	0.76	0.70	0.66	0.78	0.77	0.75	0.77	0.74	74
True	0.86	0.53	0.41	0.19	0.53	0.34	0.32	0.53	0.37	62
Accuracy							0.50	**0.58**	0.52	258

Table 3 Metric Scores using Fake News category

These scores show that in terms of overall accuracy, the Support Vector Machine algorithm performs the best over the four categories of news, including fake news. When including all selected categories of news, the Support Vector Machines algorithm and Random Forest algorithm both perform the same.

When looking at the topics individually, “pensioenen” and “stikstof” performed the worst when looking at their accuracy (both around 35%). This makes sense because these topics have the least number of tweets. In addition, “migratie” did the best with an average of 73% accuracy. This is also explained to the fact that this topic had the greatest number of tweets. Here, the importance of a large pool of tweets is shown when aiming to get the highest accuracy possible.

4.3. Confusion Matrix

The confusion matrix of the SVM classification model is shown in the table below and is interesting to look at because it gives into how many tweets from each category are classified correctly or misclassified into other categories.

Confusion Matrix Support Vector Machines		Predicted				
		False	Misleading	Neutral	Other	True
Actual	False	13	9	5	5	5
	Misleading	0	14	8	1	7
	Neutral	9	0	13	11	14
	Other	5	2	9	43	9
	True	8	7	12	4	45

Table 3 Confusion Matrix (SVM)

This confusion matrix shows that for each classification model there are several tweets that are predicted in the wrong category. The most interesting categories to look at are the true versus false confusion. When looking at the resulting table, the Support Vector Machine algorithm performs best concerning the confusion of false news for true news. Yet, the number of true tweets in this train-test sample is much higher than the number of false tweets, which partly can explain this result. A notable result is that for all three classifiers no actual misleading tweets are confused with false tweets. This can make sense because misleading tweets more resemble true news than false news as they include true facts but lead to the

wrong conclusions. Also, the neutral category seems hardest to predict. This was also seen in the lowest scores in the table displayed the metric scores.

4.4. Feature Importance

As the Support Vector Machine shows the highest accuracy, it is interesting to look at the features that are the most determining in the classification task. The *coef_* function from scikit-learn is used to obtain the feature weights. The feature names and their corresponding coefficients are put together and used to get the top hundred most determining features shown in appendix b. The number one word is "rutte" which seems realistic as he is one of the main figures during the 2021 Dutch elections. The other words differ, as some are from certain topics and others are more general words, like "zij". As the migration topic was dominant with 405 tweets there are some specific words, like "massaimigratie", that are important for the classification task. This shows some bias to the topics that are overrepresented compared to other topics. Yet, there is a substantial number of words that do not specifically correspond to one topic.

When looking at categories of news on their own, the feature importance coefficient of the fake news category shows that the most determining words, like "jullie", "nietwesterse", and "hun", do direct to the use in a negative perspective of a distinction between the "we" versus "they" group. This is interesting as fake news concerning the elections is mostly aimed at spreading false statements politicians proclaimed to have made. For some extend, the words with a high coefficient number have some relation to the topics of news chosen. This is due to the limited number of tweets in the dataset. When looking at the true and neutral categories of news, some the most determining words, like "factcheck", "klopt", and "nieuwscheckers", do point in the direction of revering to using fact-checked items as reference for making statements. Again, due to the small dataset, some words are linked to the chosen topics of news. However, looking at both collections of words, this is promising for further research with larger and cleaner datasets to determine key words that can, for example, flag fake news.

5. Discussion

This research investigates the strategies to combat fake news during the election times, especially focused on the Dutch elections. Due to the lack of national policy regarding combatting fake news in the Netherlands, the European Approach is described in this thesis.

The Dutch government's website also refers to this European Approach, therefore this is leading in the overview of the current strategies to combat fake news.

Due to the small dataset, with the limited the number of tweets and the few topics selected, no harsh conclusions can be made concerning the results of the text mining experiment. The accuracy could be different for a Twitter-scale dataset. Also, the application on the Dutch elections and therefore the Dutch language being the focus of analysis have influence on the accuracy of the classification. So, this being done on different languages and topics in different countries can result in different results. However, it does give insight into the capability of automated methods to predict whether a tweet is true or false. These results do show that even with a small dataset, models can be trained to predict the categories of news. So, for future research this is promising. With a larger dataset and further hyperparameter tuning or the inclusion of other classification models, improvements can be made regarding the abilities of the models to predict the category of news.

In comparison with the thesis focused on the Indonesian elections, where the average of the three classification models was used to showcase the results, the same is seen where the scores are better when selecting fewer categories. The F1-scores, the main metric in this thesis to distinguish the best feature, are higher with the selection of three classes instead of five, which is comparable for the results of this thesis. As the average of the models is taken in the Indonesian thesis, no precise comparison can be made for the results per category of news. However, the numbers do show, looking at the unigram words results as these are used in this study, that the recall of misleading is higher than the Indonesian study. Mainly due to the smaller number of tweets categorized as misleading in the Indonesian study, where in this study the number of misleading tweets is more in line with the other categories. This shows that the number of tweets per category does play a role in the ability to correctly predict the category. For further research this can be taken into consideration to focus on gathering equal numbers of tweets per category.

For the purpose of this thesis, I choose to focus on the results gathered from the experiment including the fake news category, as both false and misleading news fall under the definition used. Moreover, the other category was kept in the experiment as excluding it would not give realistic results. It is natural that concerning a news topic and the detection of fake news, a lot of other news or topics are picked up regarding tweets on Twitter. The aim was to find a generalizable model, so with the small dataset the decision made was to keep the other category included in the experiment.

6. Conclusion

The purpose of this thesis was to investigate to what extent automated methods can help detect fake news on Twitter in relation to the Dutch elections. This taken together with the broader question what the current strategies are to combat fake news in election times. When looking at the Dutch policies regarding combatting fake news the government website referred to the European approach and their action plan to tackle online disinformation. Even though it is a complex matter, the European Commission stresses the importance to tackle fake news. Their action plan is guided by four objectives: transparency, diversity, credibility, and inclusive solutions (European Commission, 2018, p. 6). Fact-checking is also labeled as an important tool to the verification of content and check the trustworthiness of sources on social media. Additionally, new technologies need to be explored to help tackling disinformation, including automated methods. The European Commission stresses the importance of combatting fake news during election times, as elections lie in the basis for the democratic functioning of a state. The spread of fake news during election times to influence the public's choice and manipulate electoral processes has turned into a cyber-attack mechanism. Also, even with the limited focus of the exploratory study being done in this thesis, a large amount of false and misleading tweets was discovered when analyzing just the six selected topics. This taken together showcases the importance to look at technologies to combat fake news distribution via social media.

One of the emerging technologies is in the field of data mining. Especially regarding fake news on social media, as in this digitalized world, many people consume their news through online media, and increasingly through social media. Therefore, this is an important platform to look at regarding the automated methods of the detection of fake news. This thesis aims to help understand these automated methods through a text mining experiment. Twitter is used as this is, until this day, still a popular platform where people share and spread their opinions and news. Topics were selected from the period of the Dutch elections of 2021 and tweets were collected that corresponded to these topics. All tweets were manually annotated and put into the selected categories of news. The experiment was conducted using three classification models and the best results were seen when the categories of false and misleading news were merged into the category of fake news. The Support Vector Machines algorithm had an accuracy of only 58%, which was the highest followed by the Random Forest algorithm (52%) and Multinomial Naive Bayes (50%). When looking at the feature importance, some bias to the overrepresented topics is noticeable. However, in the top

hundred most determining words, a variety of words is shown that do not directly link to one of the topics. When focussing on category specific feature importance, some words can be identified with a more positive or negative sense, for example ‘jullie’ or ‘hun’ by fake news and ‘klopt’ or ‘factcheck’ by true news. This indicates that some words tend to direct to false news statements which can be further analyzed to determine whether this can be a method to flag for fake news statements.

Due to the small dataset, no hard conclusions can be drawn from this experiment. Nevertheless, to answer the first research questing regarding the automated methods, there are possibilities for further research to improve the ability of automated methods to detect fake news. This small-scale explorative study did show that it is very difficult to detect whether a tweet is true or false from the content of the tweet. When looking at societal implementation of automated detection, the precision and recall trade-off of a model comes into play. In the case of Twitter and fake news detection, it is expected that in real time a large imbalance will take place in the dataset, when realistically taking all tweets excising in Twitter together, only a small section will be labeled as fake news. Therefore, recall and precision become more important in real life cases in comparison to the accuracy metric in the experiment conducted in this thesis. As fully automated detection is not yet implemented in real life, fact-checkers could progress this implementation with models that obtain a high recall, and therefore flag a lot of cases of fake news, and extent this with a human fact-check follow-up to determine which tweets are actually fake news items. When automated methods develop, the trade-off between recall and precision can be more balanced. For now, it is important for an automated method to find as many cases of fake news as possible as the spread is vigorously, especially during times of elections. Concluding, as technologies improve and more research is being done in this field of data mining and the detection of fake news on social media, automated methods can increasingly help in the process of combatting fake news, with special attention during election times. The elections are one of the main elements of the functioning of a democratic state, so with the increasing occurrence of cyber-attacks directed to the elections, involving the spread of fake news as a cyber-attack mechanism, the priority to tackle this trend is high. Therefore, current strategies regarding combatting fake news are centralized in the European approach where the seriousness of the issue is stressed, and policy is made to prioritize the development of technologies that help the battle against fake news.

7. Bibliography

Auxier, B. E., & Vitak, J. (2019). Factors motivating customization and echo chamber creation within digital news environments. Social media+ society, 5(2), 2056305119847506.

Allcott, H. & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. The Journal of Economic Perspectives, 31(2), 211-235.

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication*, *13*(1), 210-230.

Carley, K. M. (2020). Social cybersecurity: An emerging science. Computational and Mathematical Organization Theory, 26(4), 365-17.

Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. Proceedings of the National Academy of Sciences, 118(9).

European Commission. (2018, April). *Tackling online disinformation: a European Approach*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0236&from=EN

European Commission. (2021, February 23). *Code of Practice on Disinformation*. Shaping Europe's Digital Future - European Commission. https://ec.europa.eu/digital-single-market/en/code-practice-disinformation

Flintham, M., Karner, C., Creswick, H., Bachour, K., Gupta, N., & Moran, S. (2018). Falling for fake news: investigating the consumption of news via social media. In CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systemsdoi:10.1145/3173574.3173950

Freedom House. (2017, November). *Manipulating Social Media to Undermine Democracy*. Freedom on the Net. https://freedomhouse.org/sites/default/files/2020-02/FOTN_2017_Final_compressed.pdf

Go, A., Huang, L., & Bhayani, R. (2009). Twitter sentiment analysis. Entropy, 17, 252.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 US presidential election. Science, 363(6425), 374-378.

Guadagno, R. E., & Guttieri, K. (2021). Fake news and information warfare: An examination of the political and psychological processes from the digital sphere to the real world. In Research Anthology on Fake News, Political Warfare, and Combatting the Spread of Misinformation (pp. 218-242). IGI Global.

Guess, A. M., & Lyons, B. A. (2020). Misinformation, disinformation, and online propaganda. Social media and democracy: the state of the field, prospects for reform, 10-33.

Lee, T. (2019). The global rise of "fake news" and the threat to democratic elections in the USA. Public Administration and Policy.

Lin, Y. (2021, May 17). *10 Twitter Statistics Every Marketer Should Know in 2021 [Infographic]*. Oberlo. https://www.oberlo.com/blog/twitter-statistics

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2021, March 25). *Desinformatie en nepnieuws tegengaan*. Desinformatie en nepnieuws | Rijksoverheid.nl. https://www.rijksoverheid.nl/onderwerpen/desinformatie-nepnieuws/aanpak-desinformatie-en-nepnieuws

Ministerie van Justitie en Veiligheid. (2020, 15 oktober). *Dreigingsbeeld Terrorisme Nederland 53*. Publicatie | Nationaal Coördinator Terrorismebestrijding en Veiligheid.

https://www.nctv.nl/documenten/publicaties/2020/10/15/dreigingsbeeld-terrorisme-nederland-53

Munger, M. (2017, November 15). *Truthiness and the origins of "fake news."* Learn Liberty. https://www.learnliberty.org/blog/truthiness-and-the-origins-of-fake-news/

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A. L., & Kleis Nielsen, R. (2018). Reuters Institute Digital News Report 2018. Reuters Institute for the Study of Journalism. http://media.digitalnewsreport.org/wp-content/uploads/2018/06/digital-news-report-2018.pdf?x89475

Osman, M. (2020, September 3). *Mind-Blowing Twitter Stats and Facts on Our Favorite Network (2021)*. Kinsta. https://kinsta.com/blog/twitter-stats/

Pellarolo, M. (2018, April 16). *Customers' tweets classification - Martín Pellarolo*. Medium. https://medium.com/@martinpella/customers-tweets-classification-41cdca4e2de

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter, 19(1), 22-36.

Suryadikara, R. (2020, August). False News Classification and Dissemination Analysis: The 2019 Indonesian Presidential Election (Master's Thesis in Computer Science). https://theses.liacs.nl/1772

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). "Defining "Fake News" A Typology of Scholarly Definitions". In: *Digital Journalism* 6.2, pp. 137–153.

Vosoughi, S., Roy, D. & Aral, S. (2018). The spread of true and false news online. Science (American Association for the Advancement of Science), 359(6380), 1146-1151.

Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in social media: definition, manipulation, and detection. ACM SIGKDD Explorations Newsletter, 21(2), 80-90.

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, *57*(2), 102025.

Appendix A

Topic	**Queries Snscrape**
Gepensioneerden	“Koopkracht gepensioneerden”
Stikstof	“stikstof schadelijk”
Burnout	“burnout jongeren”
Migratie	“niet-westerse allochtonen”
Staatsschuld	“staatsschuld Nederland”
Code zwart	“code zwart zorg”

Nieuwscheckers Topics and Queries

Appendix B

The top 100 value pairs are:

'rutte': 1.3478957441572095	'dankzij': 0.7415939845345435
'zij': 1.3345811907576743	'das': 0.7414214732386285
'zit': 1.179054445973182	'bovenop': 0.7397994796419644
'massaimmigratie': 1.1634038652446252	'miljoenen': 0.7347632788730419
'burnout': 1.1208358673492098	'beleid': 0.7272954678537379
'failliet': 1.0716272456830305	'hoe': 0.7207344116578996
'kinderen': 1.0649899457862102	'geld': 0.7141592729002724
'markrutte': 1.0041867919190415	'toekomst': 0.7140040740838374
'schoffering': 0.9395694333436214	'denk': 0.7122005432252769
'al': 0.9252467911370389	'beroerte': 0.7043496220514128
'niemand': 0.8904398208449239	'hartstilstand': 0.7043496220514128
'he': 0.8850188881407918	'adnl': 0.7040665948035731
'alle': 0.8819145795676794	'stemzeweg': 0.7002209058360683
'woorden': 0.8505874431796683	'stukje': 0.6906129101828561
'tachtig': 0.8337724408273464	'hugodejonge': 0.688431165589881
'nu': 0.8226793257916489	'idee': 0.6879765798305427
'tien': 0.8223324617901102	'meer': 0.6805033741512164
'laatste': 0.8164244897236955	'ellende': 0.6751746123035264
'leven': 0.8075087481957649	'doorgaat': 0.6707207489708247
'heeft': 0.8053886594669394	'nederlanders': 0.6623367019110141
'krijgt': 0.8021381281193233	'procent': 0.6590653240154137
'kind': 0.8021374653050253	'inmiddels': 0.6582861304699377
'liever': 0.7952099073849204	'huizen': 0.6496184102438373
'druk': 0.7871866978548746	'enige': 0.6436597324425003
'enz': 0.7822660949224292	'gt': 0.6335876746244324
'laat': 0.7751056454519498	'eerste': 0.6264853914173318
'partijen': 0.7636135086993736	'maanden': 0.6209004138098769
'land': 0.759546584619666	'jongeren': 0.6201497044208117
'we': 0.7578169169393406	'dagen': 0.6169849635777626
'lossen': 0.7523263093495135	'bijna': 0.6130180997104192

'oplossing': 0.6128513993662871	'record': 0.5303618672222459
'httpstcowxudjezd': 0.602726353323803	'httpstcoqrzcyqedo': 0.5274656073139564
'huidige': 0.600711967309502	'leverden': 0.5274656073139564
'vanaf': 0.5991712083968491	'woekerpolisnl': 0.5274656073139564
'gek': 0.5870375916433702	'je': 0.5252693673118386
'staan': 0.5845374225390154	'leeft': 0.5250062064407708
'had': 0.5829204715083827	'rand': 0.5231751872251863
'blijfbinnen': 0.571492650220731	'verwijderd': 0.5228240190684631
'httpstconzkbrkrshn': 0.571492650220731	'gewone': 0.5216975407551337
'httpstcozucagvcj': 0.571492650220731	'eindhoven': 0.5214048929435536
'online': 0.5698322513040598	'rondom': 0.5214048929435536
'jouw': 0.5659580686019441	'covidbesmetwie': 0.5206342204986247
'heb': 0.55614405437315	'ouderenzorg': 0.5206342204986247
'gezegd': 0.5560630985729399	'patientde': 0.5206342204986247
'mijn': 0.5551887290502024	'snelste': 0.5206342204986247
'honderd': 0.545232834146995	'vaccinmet': 0.5206342204986247
'spreken': 0.5416805236561215	'verpleging': 0.5206342204986247
'van': 0.535924438737458	'gaan': 0.5202923989598678
'jaar': 0.5315841912399065	'blijven': 0.517194134374824
'echt': 0.530628455974361	'aow': 0.5143599283894978

Top 50 of Fake news category:

'jullie': 1.0613087079850239	'maken': 0.7459412175306666
'nietwesterse': 0.9695627942242496	'als': 0.734969175563984
'hun': 0.9670018124778937	'stikstof': 0.7315355189161137
'grote': 0.9153727421022254	'driekwart': 0.7100728219590815
'tot': 0.8530557099309706	'groenlinks': 0.7090708053898838
'onder': 0.788234657371345	'was': 0.7013674666340789
'we': 0.7814527340928297	'zien': 0.6843097881959151
'iedereen': 0.7570951220235533	'gaan': 0.6722854960296413

'waarvan': 0.6676270708101241	'uit': 0.5372793713779035
'wil': 0.6617666739228331	'ben': 0.5319994959315455
'jaren': 0.6612560559669699	'veel': 0.528856698337129
'bentebecker': 0.6607466087880638	'en': 0.5274431511717533
'worden': 0.6526595400306207	'rutte': 0.524838644438698
'binnengehaald': 0.6458252758288441	'ik': 0.5245519379860487
'schadelijk': 0.6256843115597529	'drie': 0.5202267706511562
'dus': 0.6246945973440805	'van': 0.5160817856361174
'moeten': 0.5945205596792448	'sluiten': 0.5155867135050135
'stress': 0.5906865931776077	'extra': 0.5147609989927135
'reguliere': 0.5885434100307928	'httpstcoiqjquqqmcb': 0.5125938642525816
'cs': 0.5849518902879276	'afgelopen': 0.5085529206967976
'ziekenhuizen': 0.5817686375670641	'nwa': 0.5081669511047114
'liefst': 0.5657353928354082	'wilders': 0.5046135892479859
'erbij': 0.5621756173601072	'zichzelf': 0.4974563004042485
'doorgaan': 0.5481851009727683	'sp': 0.49311693888598224
'randje': 0.5469850212968685	'januari': 0.4891372193474547

Top 50 of True/Neutral news category:

'krijgen': 1.3447167659055987	'miljard': 0.7983195393053625
'factcheck': 1.12749529819348	'jaar': 0.7671501854621103
'kan': 1.0017246320141857	'goed': 0.7560097304134605
'wel': 0.9068392295334744	'deze': 0.7538791601445167
'jaren': 0.9053542680525857	'het': 0.7516957311248158
'betekent': 0.8938104421307905	'zelf': 0.7411909825775408
'nieuwscheckers': 0.8769521759550984	'zijn': 0.733055606238062
'klopt': 0.8623888253813826	'nieuws': 0.729768930073646
'werkenden': 0.8612855399132091	'kijk': 0.7289977275705015
'je': 0.844529889106177	'natuurlijk': 0.7272675069001632
'zorg': 0.8251517410239065	'na': 0.7160751142005247
'weer': 0.8062212342381553	'pensioenen': 0.6966941235280396
	'ben': 0.6896785702412147

'nieuwsuur': 0.6882730129810879
'was': 0.6787804046832114
'europa': 0.6749850910912685
'beweert': 0.664349980362966
'mee': 0.66127327669197
'mijn': 0.6553448325066357
'bericht': 0.6419285467425212
'cijfers': 0.6306186678467416
'stikstofoxiden': 0.6138083489096371
'nl': 0.6078890866622028
'totaal': 0.6047607445701506
'acute': 0.6007335440621333
'gemiddeld': 0.5984215991167828

'medischcontact': 0.5979959222596983
'tegen': 0.5965866262478443
'lage': 0.5904468103382357
'regels': 0.5887220792674184
'pensioen': 0.588655561668971
'sprake': 0.5820489153582015
'overbelaste': 0.5630843755545212
'nh': 0.5591770439154707
'uitgewerkt': 0.5499225586161769
'cijfer': 0.5483894571129606
'vast': 0.5474182515479415
'zeker': 0.5451100250735518

Appendix C

The annotation guidelines are the following:

One person will do the largest task of annotation. This person does not have any political affiliation or belongs to a political party to enable impartiality. The thesis student will annotate a small set (around 100 tweets) with the purpose of estimating inter-rater reliability. All topics are linked to one supporting URL from the Leiden University fact-checking website. All tweets will be selected per topic, based on keywords that correspond to the information discussed and/or claims made. The columns that are used for annotation are "Topic", "URL", "Tweet", "True Tweet", "False Tweet", "Misleading Tweet", "Neutral Tweet", and "Other". The annotator is tasked to put each tweet in a category based on the definitions from the research of Suryadikara (2020):

- True: Tweets that relate to the topic and are true or accurate according to supporting URLs.
- False: Tweets that relate to the topic and are false or incorrect according to supporting URLs.
- Misleading: Tweets that relate to the topic and have correct information according to supporting URLs but cause wrong conclusions.
- Neutral: Tweets that relate to the topic but do not mention the precise fact-checked essence of information from the supporting URLs in any form (neither true nor false statements).
- Other: Tweets that do not relate to the topic or are not discussed within supporting URLs.

The process

1. Topics that relate to the Dutch elections of 2021 are selected from the nieuwscheckers[4] website.
2. The tweets will be retrieved per query consisting of (multiple) keywords that are mentioned in the fact-checked items from the nieuwscheckers[4] website.
3. The annotator will receive the collection of tweets per topic and is asked to put each tweet into one of the described categories.

[4] http://nieuwscheckers.nl

4. The work done by the annotator will be looked over by the student conducting the research. Tweets that raise questions regarding the proper category will be reexamined.
5. The process concludes in a dataset where tweets are categorized per topic and will be ready to use in the research.

Annotator

Before examining the tweets, the annotator is asked to read the attached URL concerning the topic of the tweet. This URL will be all the annotator needs to use to verify the tweet and put it into a category. First the annotator checks whether the tweet is a statement or claim mentioned in the description of the supporting URL. The annotator needs to limit their work to the claims or statements that are related to the topic or discussed in the supplied URL. This with the aim to stay focused on the topics selected.

Examples

An example of each category is given to guide the annotator. The examples are given from the annotation that the thesis student has done concerning the topic “gepensioneerden”. The main statement concerning this topic was that over the past 10 years, pensioners have lost about 20 percent of their purchasing power. This was fact-checked, and the conclusion is that this statement is false. A collection of 121 Tweets were gathered concerning this topic and all were put into one of the mentioned categories. Each category will be supported with an example.

Tweet	Category of News	Explanation
50_Plus_Drenthe: RT @Erik_de_Graaff: Liane den Haan legt het goed uit bij Nieuwsuur. De achteruitgang van koopkracht van gepensioneerden is niet 20%. Maar het verschil met de koopkracht van werkenden is wel zo veel groter geworden. Het is belangrijk dat 50PLUS mee praat over…	True	This Tweet confirms the fact-checked information and the supported statement about workers and pensioners is true and is seen in a graph in the fact-checked article.

@VVD Gepensioneerden optimistisch? 25% koopkracht afgejat door Rutte en zijn trawanten.	False	This information is False, as the fact-checking website explains that the average loss in purchasing power is not near the 25% mentioned in this Tweet.
@D66 Mede dankzij de antidemocraten66 zien gepensioneerden al 13 jaar hun koopkracht verdwijnen.	Misleading	On the one hand is it true that pensioners on average over the past 13 years have handed in a part of their purchasing power. However, this tweet concludes that it is disappearing, which does not follow the facts.
Wij maken ons niet alleen zorgen voor toekomstige pensioenarmoede, maar ook voor armoede uit het verleden. Wist jij dat er nu al honderdduizenden gepensioneerden niet of nauwelijks kunnen rondkomen. Het al 12 jaar niet indexeren is een sluipmoordenaar die de koopkracht uitholt. https://t.co/60FE4BCBty	Neutral	This information is neutral as the supported fact-checked information does not provide claims to mark the tweet as true or false. It does concern the topic but the essence of information by the supported URL is there.
@VCProfessionals heeft een meldpunt geopend voor koopkracht werkenden en gepensioneerden. De loonstrookjes die deze maand binnenkomen zullen in veel gevallen tegenvallen. "We willen zicht krijgen wat dit voor de koopkracht betekent" aldus de VCP. CMHF https://t.co/8JZQ4eV281 https://t.co/Xv1RupaDcI	Other	This tweet contains information that is not discussed in the supported URL.