Exploratory Analysis on Linguistic Patterns of Fake and Real News Related to the COVID-19 Pandemic

Thesis by Rosa Lucassen
MSc in Applied Data Science

Supervisor: Anastasia Giachanou
Second examiner: Ayoub Bagheri

Universiteit Utrecht

Acknowledgements

During the writing of this thesis, I have received an exceptional amount of support. First, I would like to thank my supervisor Anastasia Giachanou for her assistance. I could always count on your extensive feedback, and you always seem to make time for your students. Thank you for guiding my work to a higher level. I also would like to thank my family and friends for supporting me through this period and always being there for me.

Abstract

Since the first COVID-19 case, the world has been simultaneously dealing with a pandemic and the consequences of fake news shared about it. We believe that fake and real news have different linguistic characteristics. Therefore, this study explores the linguistic and semantical differences between fake and real COVID-19 news on social media. We use a dataset collected by Patwa et al. (2020) that contains real and fake social media posts. First, we use LDA topic modelling on the collection to extract topics regarding COVID-19. In total, the model found 21 topics with a high coherence score of 0.725. The number of fake and real news articles per topic shows which topics need more careful attention by the public.

We use VADER to analyse the differences in sentiment polarity between fake and real news. The results show that fake news is statistically significantly more negative than real news overall. Additionally, specific topics (e.g., the topics "Misinformation COVID-19" and "Donald Trump") showed more negativity within fake news than real news. Lastly, we looked at linguistic and emotional differences between real and fake news using the English LIWC dictionary. Using the Mann-Whitney test, we show that overall fake news shows statistically significantly more anger than real news. Within topics, we found further dissimilarities between fake and real news using grouped LIWC categories.

Table of Content

List of tables

List of figures

1. Introduction

In December 2019, in Wuhan, China, the first infected person was confirmed with a highly infectious virus. Since this moment, the world has been going through an ongoing pandemic of the coronavirus disease 2019 (COVID-19). During the pandemic, numerous fake news articles were spread related to the origin, transmission, cure, and vaccination of COVID-19. Therefore, according to Patwa et al. (2020), the world is not only going through a pandemic but also "fighting an infodemic [1]", i.e., a situation in which a lot of false information is spread in a harmful way.

During the pandemic, fake news articles were spread mainly through social media, which has become a popular way to skim news. In particular, the Pew Research Center announced in 2016 that approximately 62% of U.S. adults used social media as their news source that year, while only 49% of U.S. adults used social media for the same reason in 2012[2]. As stated by Lee (2019), fake news has become dominant across all social media platforms primarily because of its high level of social engagement. This is due to the fact that on social networks users can share their opinions and beliefs on any topic to which other users can react and/or comment.

Although many definitions have been proposed, there is still not a determined definition of the concept of fake news. According to Allcot and Gentzkow (2017), the definition of fake news is "a news article that is intentionally and verifiably false." On the other hand, Lazer et al. (2018) define fake news as "fabricated information that mimics news media content in form but not in organizational process or intent." According to them, the spreaders of fake news lack the news media's editorial norms and processes for ensuring the accuracy and credibility of information. In this thesis, we use the definition proposed by Lazer et al. (2018), which states that "fake news is fabricated information which mimics news media content that is intentionally and verifiably false."

Understanding if a news article is fake or real is very challenging. In particular, the average person is not effective at detecting fake news with typical accuracy rates within the 55-58% range (Frank et al., 2004). Especially during a pandemic there exists a lot of insecurity and fear in a society, the judgment of an average person is not reliable. A high percentage of fake news diffusion in a community can reduce the effectiveness of programs, campaigns, and initiatives aimed at people's health, awareness, and well-being (Pulido et al., 2020). The effects of fake news were known before the pandemic. In particular, researchers first revealed the impact of fake news propagation in social media during the 2016 U.S. presidential election. According to Grinberg et al. (2019), "aggregate exposures to political URLs were from fake news sources." Another event that showed the negative consequences of fake news is the pizzagate story (Britt et al., 2018). This incident that happened in 2016 ended with a gun shooting inside a pizzeria as a result of fake news claiming that there was child sex trafficking happening in the pizzeria.

[1] https://dictionary.cambridge.org/dictionary/english/infodemic

[2] https://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/

Automatic fake news detection is complicated because fake news contains fake and real information intended to confuse the reader. Vosoughi et al. (2018) showed that readers' degree of novelty and emotional reactions might be some of the main linguistic differences between fake and real news. They showed that false stories inspire fear, disgust, and surprise in replies, whereas true stories inspired anticipation, sadness, joy, and trust. Zhou et al. (2020) investigated news content at various levels, including lexicon-level, syntax-level, semantic-level, and discourse-level. They found that fake news texts, compared to true news texts, have fewer reported verbs, a more significant proportion of unique verbs, swear words, and showed more extreme emotions for both positive and negative emotions. This means that looking at sentiment and linguistic features within texts can potentially help detect fake news.

Similar to the previous studies, we will explore any linguistic differences between fake and real news. However, unlike prior work that explored those differences mainly on political news, we will focus on information related to COVID-19. We believe that this is interesting to explore because it can show how the public expresses their feelings on social media during unprecedented times. Therefore, the focus of this study is the analysis of linguistic features around COVID-19 related fake and real news answering the following research question: "What are the lexical and semantical differences between COVID-19 related fake and real news topics on social media?" This question will be answered using the following sub-questions:

- [RQ 1] How is COVID-19 related news discussed on social media and is there a topical difference between fake and real news?
    - [RQ 1.1] What are the most common topics related to COVID-19 discussed on social media (within a specific time period)?
    - [RQ 1.2] Which topics get a higher proportion of fake and which of real posts?
- [RQ 2] How polarized are the COVID-19 related social media posts within each topic?
    - [RQ 2.1] Are fake news posts that are related to a specific topic more negative than real news posts?
- [RQ 3] How are the COVID-19 related social media posts within each topic represented with regards to emotions and to other linguistic features?
    - [RQ 3.1] Are fake news posts that are related to a specific topic more emotional than real news posts?
    - [RQ 3.2] Are fake news posts that are related to a specific topic linguistically different than real news posts?

A dataset[3] containing English written social media posts collected in the first year of the pandemic will be used to accomplish this research. First, we will extract the most distinct topics for both fake and real social media posts using a topic modelling algorithm. The Latent Dirichlet Allocation (LDA) method will be described and used for topic extraction. This will give insight into the different topics discussed within both fake and real news [RQ1]. Then we will extract various textual features from the textual content, including sentiment and emotions, which will provide additional information on the linguistic differences between fake

[3] The data used in this research was collected by Patwa et al. (2020)

and real news. This study will use Valence Aware Dictionary and sEntiment Reasoner (VADER) to examine the polarization of sentiment between fake and real news [RQ2]. Next, we will use the Linguistic Inquiry and Word Count (LIWC) to look at differences in linguistic features and emotion [RQ3]. Finally, we have applied a statistical test to see if the differences in characteristics are significant.

The rest of this thesis is organized as follows. Chapter 2 discusses the literature review, which includes a brief overview of fake news analysis and detection. The third chapter presents the methodology. This chapter gives more insight into the methods used for both topic modelling and emotion detection/linguistic feature extraction. Chapter 4 describes the experimental design, including the data collection, the pre-processing of the data, and what decisions were made on the parameter settings. Chapter 5 presents and discusses the results of our research. Lastly, Chapter 6 concludes the study and offers recommendations for future work.

2. Literature review

Since the beginning of the pandemic, numerous papers have been published specifically about COVID-19 and its relationship with fake news. However, the propagation of fake news and other types of fake news, like hoaxes and rumors, is not a recent phenomenon. Various incidents of fake news can be found in the past. For example, such a story as the "Great Moon Hoax" in which the New York Sun published multiple articles about discovering life on the moon (Allcot and Gentzkow, 2017). However, through the years, the way fake news is propagated has changed. One reason is that the internet has dramatically reduced the cost of producing and distributing diverse information and perspectives (Flaxman et al., 2016). In addition, the increasing popularity of social media has allowed users to access and share content fast and without any costs, which has contributed to the propagation of fake news.

Shu et al. (2017) showed that fake news in traditional news media and social media is not identical. The former is based on psychological and social factors, whereas the latter is based on malicious accounts and echo chambers. Echo chambers are homogeneous clusters formed when content-selective exposure is the main reason for content diffusion (Del Vicario et al., 2016). While in traditional news, a person is responsible for writing a news statement, social media accounts can be run by social bots, i.e., accounts controlled by a computer algorithm (Ferrara et al., 2016). Shu et al. (2017) showed that social bots could become malicious entities because they are specifically programmed to manipulate and spread fake news online. Fake news detection based on writing style helps develop models to identify fake news before propagation on social media (Zhou et al., 2020a). However, entities can easily take countermeasures after noticing the fake news was detected (or taken down) by changing their writing style.

With the rise of online manipulation, research about detecting fake news became more popular. Zhou et al. (2020a) researched the detection of fake news at an early stage. They did this by solely relying on news content, which allowed the model to detect when fake news had been published on a news outlet but had not yet been disseminated on social media. With the advancements in deep learning, Kaliyar et al. (2021) proposed FakeBert to detect fake news on social media. This model is based on the bidirectional pre-trained (BERT) approach proposed initially by Devlin et al. (2019). A BERT-model can get the context representation of a sentence with an achieved accuracy of 98.90% which is 4% higher than baseline approaches. Additionally, Ghanem et al. (2021) proposed a method called FakeFlow that models the flow of affective information in fake news articles using a neural architecture. FakeFlow learned this flow by combining topic and affective information extracted from texts and showed better results than baseline methods.

During the pandemic, numerous researchers have started investigating the detection of (early) fake news related explicitly to COVID-19. In a small study by Apuke et al. (2020), altruism was the strongest predictor of fake news sharing associated with COVID-19. They argued that this could contribute to the proliferation of misinformation and fake news around COVID-19. Ceron et al. (2020) introduced a novel Markov-inspired computation method for identifying

topics in tweets during the pandemic. They mainly used this method to compare Twitter accounts of two Brazilian fact-checking outlets to determine similarities and differences in what they shared. Shahi et al. (2021) did an exploratory study on the propagation, authors, and content of misinformation on Twitter around the topic of COVID-19. They used psycho-linguistic analysis to investigate the differences in emotions between information and misinformation about COVID-19. Shahi et al. (2021) exclusively used the LIWC method, whereas this study will also examine the polarity of social media posts. They used a data collection of 1500 tweets, whereas, in this study, we analyze 10700 tweets.

Mutanga et al. (2020) used an LDA-based topic modelling approach to analyse tweets related to COVID-19 in South Africa. They found that some of the most popular topics were President Cyril Ramaphosa (PCR), the number of cases in real-time updates, the COVID-19 vaccine, and the Vodaphone conspiracy. The study exclusively analyzed South African COVID-19 related tweets, whereas this study uses English-written tweets regardless of geographical background[4]. LDA is a well-known topic modelling approach that works effectively on long textual data. However, Xu et al. (2020) solved the feature sparseness of short texts by training LDA on long external texts related to short texts and achieving the inference and extension of short texts' topics based on LDA.

Besides topic analysis, multiple sentiment analysis methods, including lexical-based approaches and machine learning methods, can give valuable insights into textual data. In particular, the function and emotional words people use can provide important psychological cues to their thought processes, emotional states, intentions, and motivations (Tausczik et al., 2010). Tausczik et al. (2010) described LIWC as a transparent text analysis program that can count words in psychologically meaningful categories. Gonçalves et al. (2013) developed a new method that combines multiple approaches and presented them in a Web service called iFeel, which provides an open API that accesses and compares results across different sentiment methods. Finally, Gautam et al. (2014) proposed multiple machine learning techniques with semantic analysis for classifying sentences based on Twitter data. They found that using both Naïve Bayes and the semantic analysis WordNet together led to the best accuracy of 98.9%.

Sentiment analysis is often used for fake news detection and investigations. For example, Ghanem et al. (2018) compared the emotional language of fake news to real news considering both social media posts and news articles. They did this by looking at four false information types: propaganda, hoax, clickbait, and satire. The results of the study emphasized that emotional features are essential when detecting any kind of false information. Giachanou et al. (2019) proposed a system called EmoCred, which incorporated emotions to investigate their effectiveness in credibility detection. They used three methods to generate emotional signals from the text of the claim. The results showed that EmoCred outperforms the LSTM baseline and that incorporating emotional cues into the system significantly enhanced performance in the credibility assessment task.

[4] We are aware that most English-spoken tweets originate from English-speaking countries.

3. Research methodology

In this study, we are interested in exploring the differences in linguistic features between real and fake social media posts. We achieved this by extracting topics and emotions within texts and analyzing the texts using tools from the field of Natural Language Processing (NLP). Liddy (2001) stated that NLP is defined as a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis to achieve human-like language processing for a range of tasks or applications. Using NLP, researchers can investigate texts via traditional machine learning methods or deep learning methods. In this study, we use two (lexicon and semantic) of the four (lexicon, syntax, semantic, and discourse) representation levels that have been suggested by Zhou et al. (2020) for the linguistic analysis of fake and real news regarding COVID-19.

3.1 Topic modelling

First, we analyzed the topics discussed in the social media posts using topic modelling. In general, a topic model is a statistical model that investigates collections of documents by discovering topics. Papadimitriou et al. (2000) were the first to model topics as probability distributions of terms. A corpus was defined as a collection of documents where a document is a probability distribution of a small number of topics. Tarifa et al. (2020) showed that the two primary purposes of topic modelling are selecting meaningful words to represent each topic and having separate topics by maximizing the cluster inter-distance resulting in the most distinct topics.

In this study, we extracted the most dominant topics of fake and real social media posts by analyzing the whole corpus of social media posts. The documents were analyzed using a Bag-Of-Word (BOW) model that captures the frequency of words within documents. LDA uses BOW to find different topics discussed in the data collection. LDA is a flexible generative probabilistic model for collections of discrete data and can be considered as one of the simplest topic modelling methods (Blei et al., 2003). Blei et al. (2012) showed that the most distinguishing characteristic of LDA is that all documents in the collections share the same set of topics. Still, each document exhibits those topics in different proportions. For our study, we used an LDA topic model to extract topics of posts that are related to COVID-19.

One of the challenges of topic modelling is that you need to define the number of topics beforehand. Therefore, we used the coherence score to optimize the number of topics creating the most interpretable results. Topic coherence scores are measured by calculating the degree of semantic similarity between high-scoring words in a topic (Stevens et al., 2012). In this study, we used the c_v metric. The c_v metric is based on four parts (Syed et al., 2017). First, the segmentation of the data into word pairs followed by the calculation of word or word pair probabilities. Then, a confirmation measure is calculated that quantifies how strongly a word set supports another word set. Lastly, the individual confirmation measures are aggregated into an overall coherence score. For this study, we optimized hyperparameters to get the optimal number of topics. Details on that are further presented in Section 4.3.

3.2 Linguistic analysis

Secondly, we did a psycho-linguistic analysis at a semantic level on the social media posts, including sentiment analysis. Sentiment analysis is the computational study of opinions, sentiments, and expressed emotions in texts (Liu, 2010). In the most common scenario, sentiment analysis assigns one of the following labels to a text: positive, negative, or neutral. There are different ways to calculate the sentiment analysis summarized in various surveys (Giachanou and Crestani, 2016). One of the state-of-the-art ways to calculate sentiment polarity is using word frequencies, for example, calculated by a BOW model, to lexicons that fall into specific psycho-linguistic categories (Zhou et al., 2020). Our study focuses explicitly on sentiment and emotions found within the extracted topics to investigate whether there is any difference in the distribution between fake and real news or not.

For the sentiment analysis, we decided to use VADER (Hutto et al., 2014). VADER is a simple rule-based model for general sentiment analysis designed explicitly for sentiment in microblog-like contexts like social media posts. VADER is sensitive to polarity, i.e., if a text shows positive or negative polarity and intensity of the polarity by emphasizing punctuation and capitalization. VADER was implemented in Python using NLTK's[5] sentiment intensity analyzer. This method returns a positive, neutral, negative, and compound sentiment score per document. These first three scores are ratios for proportions of text that fall into each category and should all add up to 1 per document. The compound score is a normalized, weighted composite score often used as the measurement score in sentiment analysis.

For the additional linguistic analysis, we used LIWC. The core of the LIWC program is a dictionary that contains words belonging to more than 80 linguistic, psychological, and topical categories. LIWC uses four types of categories[6] to process linguistics, including standard linguistic dimensions (pronouns, swear words, etc.), psychological processes (cognitive, social, etc.), personal concerns (work, money, etc.), and spoken categories (assent, fillers, and nonfluencies). The LIWC approach calculates the percentage of words that fall into these specific categories. This method was first described by Pennebaker et al. (2001) and has been translated to numerous languages, including French and Dutch (Piolat et al., 2011; Boot et al., 2017). Additionally, it has been used for a variety of studies, like a study by González-Ibáñez et al. (2011) about the identification of sarcasm on Twitter.

In this study, we used several categories of the LIWC dictionary to analyze the linguistic features of social media posts. These categories include positive emotion, negative emotion, anger, anxiety, sadness, death, causal words, tentative words, swear words, nonfluencies[7], and fillers. The words per category were counted using the Counter method and LIWC dictionary in Python.

---

[5] NLTK is a platform for building Python programs for computational linguistics and Natural Language Programming. Information attained from: https://www.nltk.org

[6] https://lit.eecs.umich.edu/geoliwc/liwc_dictionary.html

[7] Nonfluencies are words that lack fluency, for example, "um" and "er."

3.3 Method evaluation

During the study, we have considered other methods for topic modelling and sentiment analysis. For topic modelling, these included Non-Negative Matrix Factorization (NMF) proposed by Lee et al. (1999), Hierarchical Dirichlet Process (HDP) presented by Teh et al. (2006), and a different implementation of LDA based on the Scikit-learn library[8] in Python. The NMF method was easy to implement when we determined the number of topics before the study. However, the code for defining the optimal number of topics was more complicated than the traditional LDA method. HDP was easy to implement, but the coherence scores were not realistic. LDA using Gensim was chosen instead of HDP as it is a more standard method and has been widely used in studies. Lastly, LDA using Scikit-learn does not provide the coherence score as a linguistic interpretability measure. The coherence score is a straightforward method for optimizing the number of topics and was a preferred method in this study. Therefore, LDA using Gensim was the most suitable to use for this study.

Regarding sentiment analysis, we considered the SentiWordNet method for emotion analysis. Esuli et al. (2006) describe SentiWordNet as a lexical resource that uses WordNet synsets, an interface that looks up words in the WordNet. These synset instances are associated with numerical scores representing the polarity of terms within the documents. A SentiWordNet algorithm can be implemented in Python using NLTK, which counts each document's positive and negative words. However, we preferred to use VADER because it also gives a normalized compound score making the comparison between documents more accessible.

Additionally, we also considered the combined method proposed by Gonçalves et al. (2013) using the iFeel API. The biggest strength of this method is that it calculates the polarity of texts up to 200 characters for 16 different methods, including VADER and SentiWordNet. Unfortunately, however, the API was not functional because the developers have stopped updating the shared Docker file containing the code for the iFeel API.

[8] https://scikit-learn.org/stable/

4. Experimental design

This section discusses the data collection and pre-processing procedures, including a description of what the data looked like before and what changes we made during this study. Lastly, we show the optimization process of the LDA topic modelling parameter settings.

4.1 Data collection

Multiple datasets have been created to facilitate research in the field of fake news detection. In 2017 a new benchmark dataset was created by Wang (2017) called LIAR using PolitiFact data[9] and a novel, hybrid convolutional neural network to integrate metadata with text. With this dataset, it is possible to evaluate automatic fact-checking, rumor detection, and political NLP research. Shu et al. (2020) created a comprehensive repository called FakeNewsNet, which contains information on news content, social context, and spatiotemporal data. Shu et al. (2020) claimed that this repository could facilitate directions such as fake news detection, mitigation, and malicious account detection. Finally, Shahi et al. (2020) created a multilingual cross-domain fact-checking news dataset specifically designed for COVID-19 related news. The creators of this dataset manually annotated articles into 11 different fake news categories and built a classifier to detect fake news.

For this study, we needed a textual data collection to explore the linguistic features of fake and real news. Furthermore, because of the time limitations of the research, we needed data already labelled as fake or real. A dataset created by Patwa et al. (2020) was suitable for this study. The dataset contains social media posts written in English, mainly from Twitter, related to COVID-19 news labelled as either fake or real. Tweets that Patwa et al. (2020) classified as real are from verified sources and give helpful information on COVID-19. In contrast, fake documents include tweets, posts, and articles that make claims and speculations about COVID-19, which were not valid (Patwa et al., 2020). They used fourteen different sources in the curation of the real news data, including the World Health Organization (WHO), Centers for Disease Control and Prevention (CDC), Covid India Seva, Indian Council of Medical Research (ICMR), and more. In addition, Patwa et al. (2020) used websites including PolitiFact, Snopes, and Boomlive to collect fake news, besides other popular media content like Facebook posts, tweets, news pieces, Instagram posts, public statements, press releases, etc.

Patwa et al. (2020) divided the data into three sets: a training set, a validation set, and a testing set. They used 60% percent of the data for training purposes, 20% percent for testing and 20% for validation of the algorithm. Validation and testing datasets are primarily used when creating a new algorithm where the former optimizes parameters, and the latter is used to calculate the new algorithm's performance. After training and validating a new algorithm, researchers can use the test set to calculate the accuracy and precision. Since our study is exploratory, we used the whole dataset in our experiments on topic modelling and linguistic detection analysis. Table 1 shows some examples of the data collection.

[9] PolitiFact is a fact-checking website that rates the accuracy of claims by elected officials and others on its “truth-o-meter.”

To the best of our knowledge, there are no ethical concerns with this data. The data were publicly available and do not contain user information. The data is balanced, with 52.34% and 47.66% of the data classified as real and fake news, respectively. Table 2 shows the distribution across the sets[10].

ID	*Tweet*	*Label*
1	Chinese converting to Islam after realizing that no Muslim was affected by #Coronavirus #COVID19 in the country	Fake
2	COVID-19 Is Caused by A Bacterium, Not Virus and Can Be Treated with Aspirin	Fake
3	6/10 Sky's @EdConwaySky explains the latest #COVID19 data and government announcement. Get more on the #coronavirus data here👇 https://t.co/jvGZlSbFjH https://t.co/PygSKXesBg	Real

Table 1: Example data structure of the collection

Split	*Real*	*Fake*	*Total*
Training	3360	3060	6420
Validation	1120	1020	2140
Testing	1120	1020	2140
Total	5600	5100	10700

Table 2: Data distribution of the collection

4.2 Data pre-processing

Before an exploratory analysis, the data must be cleaned and pre-processed. For this study, multiple pre-processing steps were taken. First, we split all sentences into single words. Because the dataset contains tweets, at-signs (@) and retweet abbreviations ("rt") were removed. We also removed plus-signs, dollar-signs, and stand-alone numbers/characters from the data. Lastly, URLs were removed from the tweets since they are not linguistically relevant.

All words and characters were changed to lower case so that words like COVID-19/covid-19 were processed as the same word. English stop words were removed, using the Spacy[11] stop words library. Additionally, the words "covid," "corona," "covid19", "covid-19," and "coronavirus" were removed to make the topics more distinct. For topic modelling, punctuation and emoticons are removed from the dataset because social media posts often include both. Lastly, the processed data is lemmatized to prevent topics from containing words with the same stem, such as test, tests, testing, tested, etc., multiple times. We followed the same pre-processing steps for the sentiment analysis with the difference that we decided to keep emoticons, punctuation, and upper-casing. This is because VADER uses punctuation and

[10] Table 2 is provided in the paper written by Patwa et al. (2020).

[11] Spacy is a free, open-source library for Natural Language Processing in Python. The stop words list can be found on the original Spacy developers' GitHub page: https://github.com/explosion/spaCy/blob/master/spacy/lang/en/stop_words.py

upper-casing as an indication of sentiment polarity intensity. All code used for this study can be found on GitHub[12].

4.3 Parameter settings

Regarding topic modelling, we have done parameter optimization to find the optimal number of topics with the highest coherence score. The experimental design consisted of changing the alpha parameter (α), which represents document-topic density, and the n most used words from each topic (top_n) that are considered while computing the coherence scores. The higher α is, the more topics are contained in one document. In The LDA Mallet wrapper of Python, the default value of α is 50. However, often in practice the α parameter is set below 1. Since the dataset is relatively small, we tried the following α values $\{0.01, 0.05, 0.1, 0.2\}$. Every trial was done twice, once using a top_n cardinality of $\{5, 10, 15\}$ and once using a top_n cardinality of $\{5, 10, 15, 20\}$. Using a cardinality of $\{5, 10, 15\}$ means that the coherence scores per topic were first calculated using a top_n of 5, 10, and 15 and then averaged to give the final coherence scores per topic.

Additionally, we used the beta (β) default value from the LDA Mallet wrapper. The β parameter represents topic-word density. If β is high, then topics consist of most of the words within the corpus and if β is low, the topics consist of few words. The default value of β in LDA Mallet wrapper is 0.01. Every trial was calculated using 400 iterations. The number of topics was optimized by calculating the coherence score in the range between 1 and 40 topics. Table 3 shows the coherence scores for different parameters as calculated during the parameter optimization phase. Overall, we see that using a top_n cardinality of $\{5, 10, 15, 20\}$ results in lower coherence scores for every trial. In addition, we see the trials result in high coherence scores when compared to other work. For example, Syed et al. (2017) did an experiment on 480 LDA-models and found coherence score between 0.4 and 0.6. In addition, Gangavarapu et al. (2019) used a coherence based LDA-topic model and found a maximum coherence score of 0.52.

Trial	top_n *cardinality*	*Alpha*	*Optimal number of topics*	*Coherence score*
1.1	{5,10,15}	0.01	33	0.695
1.2	{5,10,15, 20}	0.01	24	0.667
2.1	{5,10,15}	0.05	25	0.715
2.2	{5,10,15,20}	0.05	27	0.691
3.1	{5,10,15}	0.1	30	0.712
3.2	{5,10,15, 20}	0.1	39	0.680
4.1	**{5, 10, 15}**	**0.2**	**21**	**0.725**
4.2	{5,10,15, 20}	0.2	5	0.689

Table 3: Experimental design on parameters

[12] https://github.com/RosaLucassen/ThesisADS

Figure 1 demonstrates trial 4.1 and shows that the optimal number of topics in this specific data collection is **21**. The final topic model was calculated with **1000** iterations using an LDA Gensim model with a Mallet wrapper, and the fine-tuned hyperparameters.

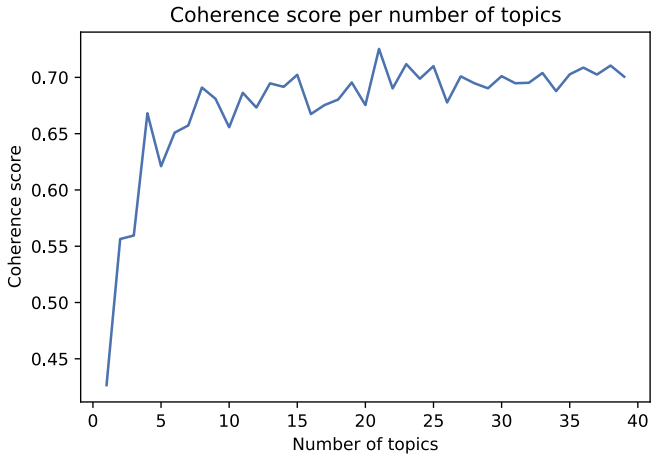

Figure 1: Coherence score per number of topics for trial 4.1

5. Results

In this section, we present the results of the study. First, the topic modelling results will be presented in Chapter 5.1. Then, in Chapter 5.2, the results of the sentiment analysis using VADER will be presented. Chapter 5.3 will present the linguistic and emotion analysis. Lastly, Chapter 5.4 includes the discussion on the results.

5.1 Topic modelling

To answer [RQ 1]: "How is COVID-19 related news discussed on social media and is there a topical difference between fake and real news?" we extracted topics from the collection. As discussed in Chapter 3, LDA uses the frequency of words in a corpus to extract the topics. Figure 2 shows the twenty most common words and their word count found in the corpus after the preprocessing. The data was also visually analyzed on the most common words for the real and fake social media posts. Figure 3 shows the twenty most common terms for fake (left) and real (right) news, respectively.

Interestingly, the most common words for real news appear to be updating social media users on COVID-19 news, these include words like case, test, report, number, and total. Whereas the fake social media posts most commonly use the terms "say" and "claim"[13]. This indicates that fake news is potentially more focused on opinions and facts without clear evidence. We also see that fake news often mentions vaccines, which could be harmful if the public receive false information about this topic. The word frequencies of the entire dataset, including both fake and real news, were used to optimize the number of topics by applying the model described in Chapter 4.3.

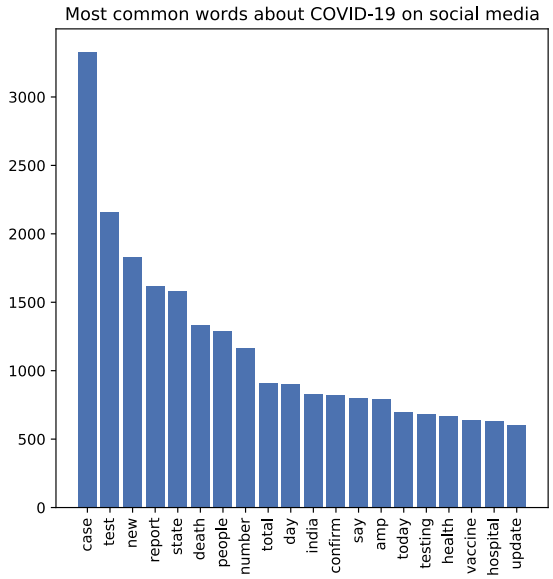

Figure 2: Most common words about COVID-19 on social media

[13] To claim something means to "say that something is true or is a fact, although you cannot prove it and other people might not believe it". Information attained from: https://dictionary.cambridge.org/dictionary/english/claim

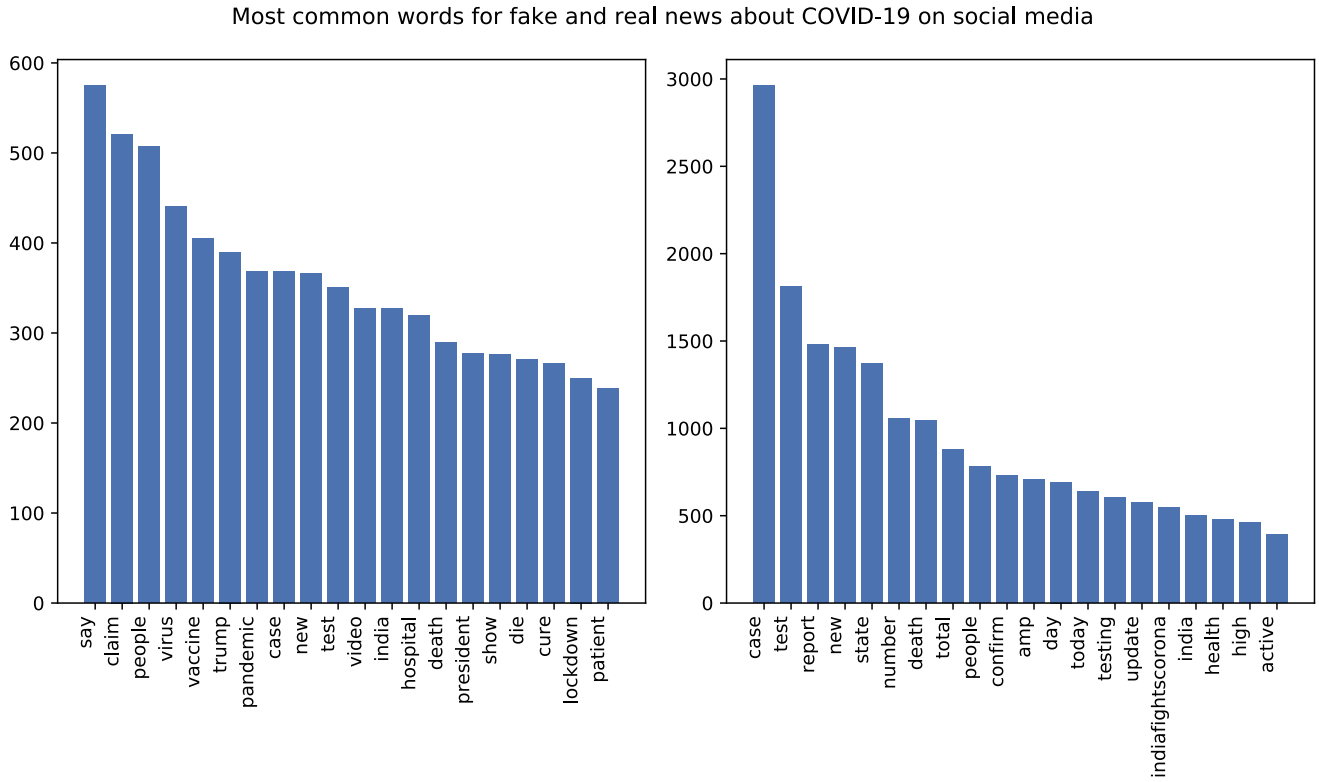

Figure 3: Most common words for fake COVID-19 news (left) and real COVID-19 news (right)

Table 4 shows the 21 topics together with the top representative words. Additionally, the table shows the manually annotated topic descriptions. This table shows the answer to [RQ 1.1]: "What are the most common topics related to COVID-19 discussed on social media (within a specific time period)?" We found multiple topics that address daily or regional updates about COVID-19 (e.g., topics USA State Daily Reports, Nigeria Daily Reports, Indian State Case Updates, etc.). Additionally, people on social media seem to talk about new precautions (e.g., topic Precautions COVID-19), restrictions, and the virus's impact (e.g., topic Restrictions/Impact COVID-19). There are also topics about COVID-19 statistics, news about testing and vaccinations, and more general information regarding COVID-19 (e.g., topics Statistics COVID-19 and General Information COVID-19). Lastly, we found topics on conspiracies, misinformation, clickbait, claims, and fact-checking within the social media posts (e.g., topics Conspiracy, Clickbait COVID-19, Misinformation COVID-19, etc.).

Topic	*Words*	*Topic Description*
1	test complete testing laboratory yesterday number isolation lab people date	Statistics COVID-19
2	bill vaccine gates people pandemic bill_gates year health government china	Conspiracy
3	virus study patient flu test vaccine sars people chloroquine hydroxychloroquine	Medical/Recovery COVID-19
4	covid_19 mohfw_india covid___19 india coronavirusindia case rate coronavirusupdate recovery covidupdate	India Hashtags COVID-19
5	case confirm death nigeria covid19nigeria discharge report case_covid19nigeria discharged discharge_death	Nigeria Daily Reports
6	case hospital people community auckland facility confirm today link positive	Auckland/India Daily Reports
7	state report test case death number today state_report update daily	USA State Daily Reports
8	health amp high pmoindia drharshvardhan pib_india state risk patient ashwinikchoubey	Indian Health Officials
9	cure virus vaccine kill water china wuhan chinese drink scientist	Clickbait COVID-19
10	mask wear spread amp face people protect learn hand wear_mask	Precautions COVID-19
11	case number report total new_zealand zealand confirm health total_number confirm_case	New Zealand Reports COVID-19
12	people lockdown health restriction home late government local stay pandemic	Restrictions/Impact COVID-19
13	trump president donald donaldtrump pandemic donald_trump virus president_trump people die	Donald Trump
14	case confirm state total death pradesh active confirm_case cure maharashtra	Indian State Case Updates
15	video claim show hospital post doctor facebook people die patient	Misinformation COVID-19
16	minister news lockdown pm home boris people johnson test prime	Prime Minister and Politicians
17	vaccine amp country trial health testing support drtedros facility access	Vaccine And Testing News
18	indiafightscorona india test case recovery lakh day rate high total	India Daily Updates
19	case contact amp day patient health school symptom test trace	General Information COVID-19
20	death case report week cdc increase late states show covidview	Regional Updates
21	claim fact check trump false fact_check misinformation coronavirusfact president factcheck	Claims/Factcheck COVID-19

Table 4: Topics extracted from the collection together with the words assigned and the manual topic descript

The next step was to investigate how much fake and real news were contained in each topic. Figure 4 shows the number of fake and real social media posts for every topic. This figure answers [RQ 1.2]: "Which topics get a higher proportion of fake and which of real posts?". The topics named "Statistics COVID-19", "India Hashtags COVID-19", "USA State Daily Reports," "New Zealand Reports COVID-19," and "India Daily Updates" contain mostly real news. Whereas the topics of "Conspiracy," "Clickbait COVID-19", "Donald Trump," "Misinformation COVID-19," and "Claims/Factcheck COVID-19" contain primarily fake news. With this information, the government can make the public aware of fake news by addressing which topics are often untrustworthy when discussed on social media. However, an interesting observation is that most topics contain both fake and real information. For example, the topic "Medical/Recovery COVID-19" includes mostly fake news and some real news. These topics have more potential to confuse the public since it is more challenging to discern whether information is real or fake. Therefore, the government should give more attention and guidance to the public regarding recognizing fake news on those topics.

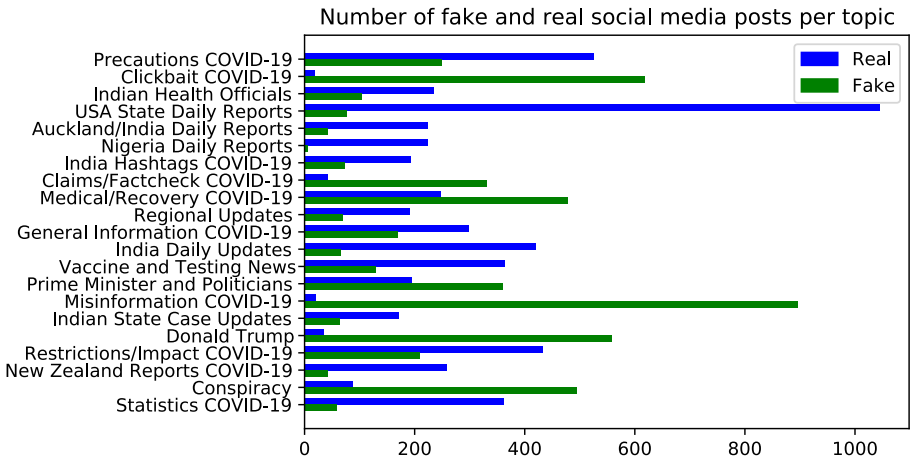

Figure 4: Number of fake and real social media posts per topic

5.2 Sentiment analysis

Before the sentiment analysis, the data had to be processed so that every document was assigned a central topic. This was done by looking at the topic document distribution and choosing the topic with the highest probability for every document. The average normalized compound score per topic shows the polarity of each topic. We found that the two most positive topics are "India Hashtags COVID-19" and "Vaccine and Testing News" with a compound score of 0.263 and 0.207, respectively. The two most negative topics are "Donald Trump" and "Medical/Recovery COVID-19" with compound scores of -0.118 and -0.062, respectively.

We used two methods to answer [RQ 2]: "How polarized are the COVID-19 related social media posts within each topic?" For the first method, we used the compound score generated by VADER. In particular, we have calculated the difference in compound scores between the fake and real documents for each topic. We describe the following two formulas, where $\bar{c}_t^{real}$ is the average compound score per topic t for the real data and $\bar{c}_t^{fake}$ is the average compound score per topic t for the fake data. $c_{d,t}$ is the compound score per document d of topic t. The compound score ranges between -1 and 1 and shows how negatively or positively polarized a document is. The total number of documents per topic is annotated with N_t. Lastly, Δ_t then shows the difference in averaged sentiment polarity per topic.

$$\bar{c}_t^{real} = \frac{\sum_{d=1}^{N_t} c_{d,t}^{real}}{N_t}$$

$$\bar{c}_t^{fake} = \frac{\sum_{d=1}^{N_t} c_{d,t}^{fake}}{N_t}$$

$$\Delta_t = \bar{c}_t^{real} - \bar{c}_t^{fake}$$

Table 5 shows the results of this approach. From the results, we see that considering both the fake and real compound scores, most of the topics predominantly have a positive sentiment on average. The topic that is the most negative, on average, when considering both the real and fake data, is the topic "Prime Minister and Politicians." Conversely, on average, we see that the topic "Claims/Factcheck COVID-19" shows the most positive sentiment.

However, as was shown in Figure 4, the proportion of fake and real news per topic is not balanced. Therefore, we used a statistical test to determine if the polarization is statistically different for fake and real news within each topic. The observations in the fake and real news datasets are independent of each other, suggesting using a t-test. A t-test assumes that the data are normally distributed and that the variances are homoscedastic, i.e., that the variances of the two groups are similar. The normality assumption was checked in Python and was not satisfied for this data. We used boxplots to test the homoscedasticity assumption. Appendix A shows examples of boxplots that do not satisfy the assumption.

Topic Description	*Difference*
Prime Minister and Politicians	−0.052
Statistics COVID-19	−0.046
New Zealand Reports COVID-19	−0.031
General Information COVID-19	−0.030
Donald Trump	−0.028
India Daily Updates	0.012
Nigeria Daily Reports	0.025
Clickbait COVID-19	0.033
Medical/Recovery COVID-19	0.035
Auckland/India Daily Reports	0.035
Conspiracy	0.076
Misinformation COVID-19	0.101
Restrictions/Impact COVID-19	0.108
Indian State Case Updates	0.118
Indian Health Officials	0.119
Regional Updates	0.154
USA State Daily Reports	0.167
India Hashtags COVID-19	0.179
Vaccine And Testing News	0.182
Precautions COVID-19	0.200
Claims/Factcheck COVID-19	0.229

Table 5: Compound score difference per topic

Therefore, we used an alternative and non-parametric method called Mann-Whitney, designed by Mann et al. (1947) to test whether one of two random variables is stochastically larger than the other. Because the Mann-Whitney test is non-parametric, the assumption that both corpora should have at least 20 values should be satisfied. Out of the 21 topics, 18 topics satisfied the assumption and were therefore tested for statistical differences. The topics that did not meet the assumption were "Nigeria Daily Reports," "Clickbait COVID-19," and "Misinformation COVID-19". Additionally, this method assumes that the distributions of the two corpora are similar, without the assumption of being normally distributed. We assumed this is the case because the compound scores always lie between −1 and 1 and often peak around the middle.

To answer [RQ 2.1]: "Are fake news posts that are related to a specific topic more negative than real news posts?" we have used the one-sided Mann-Whitney test. The null hypothesis of this test states that the median of the fake compound score is equal to the median of the real compound score for each topic. The alternative hypothesis of this test states that the median of the fake compound score is less than the median of the real compound score for each topic. Table 6 shows the rejected and not rejected topics by the null hypothesis at a p-value of 0.05. The results show that within all rejected topics, fake social media posts are statistically more negative than real ones.

Additionally, the one-sided Mann-Whitney test was used on the whole corpus, testing the null hypothesis that the median of the fake compound score is equal to the median of the real

compound score, not taking topics into account. The alternative hypothesis is that the median of the fake compound score is less than the median of the real compound score not taking topics into account. With a p-value of $1.8483e^{-62}$, we rejected this hypothesis. This proves that fake social media posts are statistically more negative than real social media posts on the subject of COVID-19.

Rejected by H_0	*Not rejected by* H_0
India Hashtags COVID-19	Statistics COVID-19
USA State Daily Reports	Conspiracy
Indian Health Officials	Medical/Recovery COVID-19
Precautions COVID-19	Auckland/India Daily Reports
Restrictions/Impact COVID-19	New Zealand Reports COVID-19
Indian State Case Updates	Donald Trump
Vaccine And Testing News	Prime Minister and Politicians
Regional Updates	India Daily Updates
Claims/Factcheck COVID-19	General Information COVID-19

Table 6: VADER Mann-Whitney results

5.3 Linguistic and emotional analysis

For the linguistic and emotion analysis, we used LIWC, and we counted the number of words of different LIWC categories in each document. Because LIWC uses a simple counter method, we normalized the counts according to the length of the document. Additionally, every document was assigned a central topic by looking at the document topic distribution and choosing the topic with the highest probability. Finally, we used the Mann-Whitney test to look at the emotional differences between fake and real social media posts for the whole dataset and within topics with a p-value of 0.05.

To answer [RQ 3]: "How are the COVID-19 related social media posts within each topic represented with regards to emotions and other linguistic features?" we use the null hypothesis that overall fake news and real news are linguistically equal. Every LIWC category, which includes positive emotion, negative emotion, anger, anxiety, sadness, death, causal words, tentative words, swear words, nonfluencies, and fillers, is tested separately per topic. We did not compute the categories filler and swear words because the sample sizes were too small. Out of all the categories computed by the Mann-Whitney test, anger was the only category rejected by the null hypothesis when looking at the entire dataset. The Mann-Whitney test did not reject the other computed categories. That means that overall fake social media posts show significantly more anger than real social media posts. All other LIWC categories show no statistically significant difference in linguistic characteristics between fake and real social media posts using isolated categories. Additionally, the results show that without normalizing the counts, the top three angry topics are "Clickbait COVID-19", "Misinformation COVID-19", and "Restrictions/Impact COVID-19". Interestingly, the first two topics are highly dominated by fake news.

We also tried to group some of the LIWC categories to see if the results differed. First, we investigated [RQ 3.1]: "Are fake news posts that are related to a specific topic more emotional than real news posts?". For this, we did a one-sided Mann-Whitney test where the alternative hypothesis is that fake news is more emotional than real news on the grouped categories {nonfluencies, filler words, swear words, anxiety, anger}. The null hypothesis for this test, stating that fake news and real news are equally emotive within these categories, was rejected using a p-value of 0.05. These results show that fake news uses more spoken language that expresses anxiety and anger.

Additionally, we looked at [RQ 3.2]: "Are fake news posts that are related to a specific topic linguistically different than real news posts?" To answer this question, we did a one-sided Mann-Whitney test where the alternative hypothesis states that real news uses more cognitive processes than fake news on the grouped categories {causal, tentative, certain, and assentive words}. The null hypothesis for this test was rejected using a p-value of 0.05. These results indicate that real news explains news more thoroughly and honestly, stating when news is proven and stating when news is preliminary and unconfirmed.

The LIWC scores for the data within a single topic were sparse, and therefore, some categories did not meet the assumption of the Mann-Whitney test within a specific topic. We rejected the null hypothesis in three cases. In particular, it was rejected for the topic "Statistics COVID-19" and positive emotion, for the topic "Medical/Recovery COVID-19" and death, and the topic "Restrictions/Impact COVID-19" and anger. This means that the rejected categories are expressed significantly more in fake than real social media news for those topics. Table 7 shows all the tested topic category combinations. In particular, it shows the topic category combinations that were rejected by the null hypothesis ($*$), not rejected by the null hypothesis ($-$), and not computed due to the small sample size (x). Note that not all topics and categories are shown in the table due to the assumption that sample sizes must be bigger than 20 data points.

Topic	*Positive emotion*	*Negative emotion*	*Causal words*	*Tentative words*	*Certainty*	*Anxiety*	*Anger*	*Sad*	*Death*
Statistics COVID-19	*	x	x	x	–	x	x	x	x
Conspiracy	–	–	–	x	x	x	x	x	x
Medical/Recovery COVID-19	–	–	–	–	x	–	x	–	*
India Hashtags COVID-19	–	x	x	x	x	x	x	x	x
U.S.A. State Daily Reports	x	–	x	x	x	x	x	x	–
Indian Health Officials	–	–	x	x	x	x	x	x	–
Precautions COVID-19	–	–	–	x	x	x	x	x	x
Restrictions/Impact COVID-19	–	–	–	–	–	x	*	x	–
Prime Minister and Politicians	–	–	–	x	–	–	x	–	x
Vaccine and Testing News	–	–	–	x	x	x	x	x	x
India Daily Updates	–	–	x	x	–	x	x	x	x
General Information COVID-19	–	–	–	x	–	x	x	x	x
Regional Updates	x	–	x	x	x	x	x	x	–

Table 7: LIWC Mann-Whitney results

5.4 Discussion on results

Yin et al. (2020) researched topics on social media during the pandemic, performed LDA topic modelling, and found **70** topics to be the most coherent, with a coherence score of **0.39**, this can be considered a low coherence score. In general, a coherence score above a **0.50** indicates that a model is interpretable enough. In our study, we did an extensive experimental design to optimize the coherence score, making sure the final topics were interpretable and coherent. In particular, with the optimized parameters we managed to reach a coherence score of **0.725**.

Essam et al. (2021) researched how Arabs perceived the COVID-19 pandemic by looking at themes on Twitter and found 16 topics discussed in their data collection. The researchers manually created and annotated their topics. A few topics showed similarities with this study, including case updates, conspiracy, and prevention of COVID-19. Missier et al. (2016) investigated topics found in tweets regarding Dengue epidemics. Using LDA, they found four main topics. Similar to this study, the topics included mostly general news, news about campaigns and prevention, and tweets about the illness itself. However, unlike this study, they also found a topic devoted to jokes about the epidemic. Lansley and Longley (2016) analyzed topics discussed in tweets from London, England. We see that these topics include more general information, like sports, socializing and routine activities. In addition, they found a topic about optimism, kindness, and positivity, and a topic about wishes and gratitude. Our study did not find similar topics because all tweets were constrained to be about COVID-19.

To the best of our knowledge, there is no study that analyzes the sentiment differences between fake and real news regarding topics extracted from COVID-19 news. Charquero-Ballester et al. (2021) analyzed misinformation spread on COVID-19 online. They concluded that misinformation about conspiracies, virus characteristics, and statistics had stronger negative emotional valence than misinformation related to cures, prevention, treatment, vaccines, and political measures. Zaeem et al. (2020) investigated the sentiment differences in fake and real news using a variety of sentiment analysis tools. Similar to our study, they found a statistically significant relationship between negative sentiment and fake news. Dissimilar to our study, positive sentiment and real news also showed a statistically significant relationship in their study. This could be the case because news about a pandemic is generally more negative and addresses topics like illness and death.

Shahi et al. (2021) used LIWC and a Mann-Whitney test as a proxy for measuring emotions in tweets regarding COVID-19. They found that both positive and negative sentiment and anger and sadness were significantly more prevalent in COVID-19 misinformation tweets than COVID-19 tweets in general. Additionally, they also found misinformation tweets show less certainty and are less tentative than general COVID-19 tweets. Similar to them, we found that fake news shows less certainty and contains less tentative words. We also found that fake news contains less assentive and causal words (when looking at the overall data).

6. Conclusion and future work

This study explored the linguistic differences between fake and real social media news regarding COVID-19. First, we applied topic modelling on a COVID-19 collection, and found that there were 21 topics discussed. We found that all topics were a mix of fake and real news. The topic "USA State Daily Reports" was the most overrepresented by real news. On the other hand, the topic "Misinformation COVID-19" was the most overrepresented by fake news. When looking at the semantics and polarization within the collection, we found that overall fake news showed statistically significantly more negative polarization than real news. Within the topics that met the Mann-Whitney assumptions, half were significantly more negative in the fake news than in the real news. The null hypothesis did not reject the other half of the topics, which showed that fake news was not significantly more negative than real news for those topics.

When looking at linguistical features and emotion, we found that fake news showed statistically significantly more anger than real social media news for the entire data collection. After grouping specific categories, we also observed two additional findings. Real news significantly used more certain, causal, tentative, and assentive words. Fake news showed significantly more anger, anxiety, swearing, nonfluencies, and filler words when grouped. Within topics, three topic category combinations showed to be significantly more emotional in fake news than in real news. These included positive emotion in the topic "Statistics COVID-19", death in the topic "Medical/Recovery COVID-19", and anger in the topic "Restrictions/Impact COVID-19".

One limitation of the study is that the dataset used includes social media posts from 2020 and only portrays the start of the COVID-19 pandemic. Therefore, we think it is interesting to repeat the study on a more extensive dataset. For example, it would be interesting to see if the emotions portrayed on social media have changed during the second year of this pandemic. In the future, we would also like to do a more in-depth study on why specific topics showed more negative polarity than others. Finally, we think it is worth researching the linguistic features and emotions in a multi-lingual dataset that includes data from different countries and exploring the linguistic differences per country. Lastly, we plan to investigate other sentiment analysis methods. By extracting the features of these models, we could use the results in a machine learning algorithm to improve the effectiveness of fake news detection in the future.

References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives, 31*(2), 211-36.

Apuke, O. D., & Omar, B. (2021). Fake news and COVID-19: Modelling the predictors of fake news sharing among social media users. *Telematics and Informatics*, 56, 101475.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022.

Boot, P., Zijlstra, H., & Geenen, R. (2017). The dutch translation of the linguistic inquiry and word count (Liwc) 2007 dictionary. *Dutch Journal of Applied Linguistics, 6*(1), 65–76.

Britt, M. A., Rouet, J.-F., Blaum, D., & Millis, K. (2019). A reasoned approach to dealing with fake news. *Policy Insights from the Behavioral and Brain Sciences*, *6*(1), 94–101.

Ceron, W., de-Lima-Santos, M.-F., & Quiles, M. G. (2021). Fake news agenda in the era of COVID-19: Identifying trends through fact-checking content. *Online Social Networks and Media, 21*, 100116.

Charquero-Ballester, M., Walter, J., Nissen, I. A., & Bechmann, A. (2021). Different types of covid-19 misinformation have different emotional valence on twitter. *SSRN Electronic Journal, 3776140*.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences, 113*(3), 554–559.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.

Essam, B. A., & Abdo, M. S. (2021). How do arab tweeters perceive the covid-19 pandemic? *Journal of Psycholinguistic Research*, *50*(3), 507–521.

Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, 417-422.

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104.

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, *80*(S1), 298–320.

Frank, M. G., Feeley, T. H., Paolantonio, N., & Servoss, T. J. (2004). Individual and small group accuracy in judging truthful and deceptive communication. *Group Decision and Negotiation*, *13*(1), 45–59.

Gangavarapu, T., Krishnan, G. S., & Kamath, S. (2019, November). Coherence-based modeling of clinical concepts inferred from heterogeneous clinical notes for icu patient risk stratification. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, 1012-1022.

Gautam, G., & Yadav, D. (2014). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. *Proceedings of the 2014 7th International Conference on Contemporary Computing (IC3)*, 437–442. IEEE

Ghanem, B., Ponzetto, S. P., Rosso, P., & Rangel, F. (2021). Fakeflow: Fake news detection by modeling the flow of affective information. *arXiv:2101.09810*.

Ghanem, B., Rosso, P., & Rangel, F. (2020). An emotional analysis of false information in social media and news articles. *ACM Transactions on Internet Technology (TOIT), 20*(2), 1–18.

Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR), 49*(2), 1-41.

Giachanou, A., Rosso, P., & Crestani, F. (2019). Leveraging emotional signals for credibility detection. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 877–880.

Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013). Comparing and combining sentiment analysis methods. *Proceedings of the 1st ACM Conference on Online Social Networks - COSN '13,* 27–38.

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*, 581–586.

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

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 8, No. 1).

Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, *80*(8), 11765–11788.

Lansley, G., & Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, *58*, 85–96.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096.

Lee, T. (2019). The global rise of "fake news" and the threat to democratic elections in the USA. *Public Administration and Policy*, *22*(1), 15–24.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

Liddy, E.D. (2001). Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing, 2*(2010), 627-666.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics, 18*(1), 50–60.

Missier, P., Romanovsky, A., Miu, T., Pal, A., Daniilakis, M., Garcia, A., Cedrim, D., & da Silva Sousa, L. (2016). Tracking dengue epidemics using twitter content classification and topic modelling. *In international conference on Web engineering,* 80–92. Springer, Cham.

Mutanga, M. B., & Abayomi, A. (2020). Tweeting on COVID-19 pandemic in South Africa: LDA-based topic modelling approach. *African Journal of Science, Technology, Innovation and Development*, 1–10.

Papadimitriou, C. H., Raghavan, P., Tamaki, H., & Vempala, S. (2000). Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences, 61*(2), 217–235.

Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M. S., Ekbal, A., Das, A., & Chakraborty, T. (2021). Fighting an infodemic: Covid-19 fake news dataset. *arXiv:2011.03327*.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001.

Piolat, A., Booth, R. J., Chung, C. K., Davids, M., & Pennebaker, J. W. (2011). La version française du dictionnaire pour le LIWC: Modalités de construction et exemples d'utilisation. *Psychologie Française, 56*(3), 145–159.

Pulido, C., Ruiz-Eugenio, L., Redondo-Sama, G., & Villarejo-Carballido, B. (2020). A new application of social impact in social media for overcoming fake news in health. *International Journal of Environmental Research and Public Health*, *17*(7), 2430.

Shahi, G. K., Dirkson, A., & Majchrzak, T. A. (2021). An exploratory study of COVID-19 misinformation on Twitter. *Online Social Networks and Media*, *22*, 100104.

Shahi, G. K., & Nandini, D. (2020). Fakecovid—A multilingual cross-domain fact check news dataset for covid-19. *arXiv:2006.11343*.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2019). Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media. *arXiv:1809.01286]*.

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.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 952–961.

Syed, S., & Spruit, M. (2017). Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. *Proceedings of the 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 165–174.

Tarifa, A., Hedhili, A., & Chaari, W. L. (2020). A filtering process to enhance topic detection and labelling. *Procedia Computer Science, 176*, 695–705.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association, 101*(476), 1566–1581.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151.

Wang, W. Y. (2017). 'Liar, liar pants on fire': A new benchmark dataset for fake news detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 422–426.

Xu, L., Xue, Z., & Huang, H. (2020). Short text semantic feature extension and classification based on LDA. *IOP Conference Series: Materials Science and Engineering, 715*, 012110.

Yin, H., Yang, S., & Li, J. (2020). Detecting topic and sentiment dynamics due to covid-19 pandemic using social media. arXiv:2007.02304.

Zaeem, R. N., Li, C., & Barber, K. S. (2020). On sentiment of online fake news. *Proceedings of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 760–767.

Zhou, X., Jain, A., Phoha, V. V., & Zafarani, R. (2020a). Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice, 1*(2), 1–25.

Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys, 53*(5), 1–40.

Appendix A: Assumption t-test