University of Padova

Department of Mathematics

Master Thesis in Data Science

The Influence Of Misinformation On Twitter During COVID-19

Supervisor
Professor Tomaso Erseghe
University of Padova

Master Candidate
Diem Ngoc Le

Academic Year
2022-2023

This thesis is dedicated to my Mother and Father, who always support and encourage me to pursue my dreams. I also dedicate this work and give special thanks to Professor Tomaso Erseghe for his inspiring and sincere guidance throughout the process. And finally, I would like to express my gratitude to the University of Padova for giving me a great opportunity to study and research in a professional and dynamic environment.

Abstract

There has been a significant impact on people's behavior due to misinformation, especially during the Coronavirus Disease 2019, which directly affects general health awareness. Previous studies have shown that individuals are susceptible to easy claims and conspiracies without appropriate evidence, and once these inauthentic claims are given momentum, they are hard to dissuade [1]. Moreover, some research also showed that misinformation campaigns can be more widespread and damaging because of the advent of social media. Understanding the danger of misinformation, a lot of algorithms and methods have been proposed to tackle this problem. The aim of this study is to detect misinformation related to COVID-19 on Twitter by using Transformers. Particularly, instead of classifying tweets that contain misinformation or not, the BERT model is used to identify tweets into three categories: tweets that contain misinformation, neutral tweets, and tweets that debunk misinformation. Since there are a variety of small topics about COVID-19, we chose to focus more on the four macro topics which are "Cure," "Bill Gates," "5G," and "Vaccines." In general, there was a significant difference in the temporal behavior and linguistic properties of false tweets compared to debunked tweets. While false tweets seemed to be written spontaneously and arbitrarily, debunked tweets instead were written more carefully with the provision of scientific evidence to expose the falseness of misinformation. Furthermore, false tweets seemed to tell and encourage people to do something whereas debunked tweets aimed to tell and inform people about the misinformation.

Contents

Listing of figures

Listing of tables

Listing of acronyms

COVID-19 ... Coronavirus disease 2019

WHO ..... World Health Organization

SVM ..... Support Vector Machine

LSTM ..... Long Short-Term Memory

NLP ..... Natural Language Processing

BERT ..... Bidirectional Encoder Representations from Transformers

1 Introduction

In 2020, the world faced an outbreak of the Coronavirus disease 2019 (COVID-19) for the first time, causing heavy damage in many fields and taking many people's lives. Moreover, the development of technology and social media has brought many benefits in keeping people safe, informed, and connected during the pandemic and has facilitated the spread of fake news and misinformation. On 15 February 2020, World Health Organization (WHO) Director-General Tedros Adhanom Ghebreyesus declared, "We are not just fighting an epidemic; we are fighting an infodemic." [5]. Based on WHO, an *infodemic*, a portmanteau of "information" and "epidemic", is a level of information overload affecting people's ability to find trustworthy sources and reliable guidance when necessary. In today's world, thousands of sources display news on social media. People need honest and trustworthy sources of information about the world around them because fake news of various types plays an essential role in misleading the community. There has been an allegation that misinformation is being used to spread COVID-19 [6]. Similar trends were seen during other epidemics, such as the recent Ebola, yellow fever, and Zika outbreaks [7]. *Misinformation* is defined as false or inaccurate information that is deliberately created and intentionally propagated. Misinformation during a pandemic can negatively affect human health. A study published in the American Journal of Tropical Medicine and Hygiene [8] estimated that from January to March 2020, 5,800 people were hospitalized (with 800 deaths) due to a rumor that claimed to drink highly-concentrated

alcohol-based cleaning products would cure COVID-19. In addition, much of this misinformation is based on conspiracy theories, and these elements are incorporated into seemingly mainstream discussions. Various aspects of the disease have been inaccurately reported, including how the virus originated, its cause, its treatment, and its mechanism of spread. Several narratives have emerged claiming that the virus is caused by 5G cellular technology [9] or Bill Gates uses the virus to enslave humanity through a global vaccination program [10]. Misinformation can spread quickly and influence people's behavior, potentially resulting in them taking more risks. These factors increase the severity of the pandemic, causing more harm and endangering the reach and sustainability of the global health system. For this reason, it is necessary to verify the information to maintain its integrity.

Detection is the most crucial step in tackling misinformation on social media. It is a challenging task since misinformation is often manufactured to deceive. In December 2021, over 10,000 unique articles regarding the pandemic were fact-checked by the International Fact-Checking Network (IFCN) at the Poynter Institute [11], the #CoronaVirusFacts unites more than 92 fact-checking organizations around the world in publishing, sharing, and translating facts surrounding the COVID-19 pandemic. Implementing fact-checking tools is critical in building acceptance and trust in society [12]. Fact-checking is validated by professionals will only make a difference if they are disseminated on the internet broader and faster than the fake news [13], and a consistent and adequate fact-checking process is vital to combating misinformation.

Social media is an ever-expanding environment. It provides a fascinating lens to see how people share their lives, interact with others, and gather information about themselves and their health [14]. Due to that reason, the aim of this study is to detect fake news on social media, in particular, on Twitter, by using Transformers. Specifically, instead of classifying tweets that contain fake news or not, the BERT model is used to identify tweets into three categories: tweets that contain fake news, neutral tweets, and tweets that debunk fake news. Furthermore, linguistic aspects of tweets were also studied using BERTAgent and LIWC to better understand the differences between the three types of tweets.

The thesis is structured as follows: Chapter 2 reviews recent research on fake news detection on social media. BERT and BERTopic are described in Chapter 3. In Chapter 4, the data collection and pre-processing steps are shown, followed by

the models comparison in Chapter 5. The results of the proposed method will be discussed in Chapter 6. Finally, Chapter 7 is dedicated to the conclusion of this thesis.

2

Related Research

The proliferation of internet news in the early 2000s produced a new set of concerns, including the possibility that an excess of the diversity of ideas would make it easier for like-minded citizens to construct "echo chambers" or "filter bubbles" where they would be sheltered from opposing viewpoints [15]. In recent years, social media has changed the world rapidly through the dramatic increase in the number of users. Particularly, the percentage of US adults who use social media rose from 5% in 2005 to 79% in 2019 [16]. Consequently, it is possible to share content between users without the involvement of third-party filtering, fact-checking, or editorial judgment. Richard Bowyer, Senior Lecturer in Journalism at the University of Derby, said that: "Nowadays everyone is an editor and everyone can publish news - especially on social media." With the rapid development of technology, misinformation has become more prevalent among us. His quote explains the danger it poses and how it adversely affects the world of journalism. Following the 2016 US presidential election, it was claimed that fake news played a role in President Trump's election [15]. Furthermore, the outbreak of the coronavirus pandemic has brought another wave of misinformation. Many articles advising how to treat Coronavirus have been shared across the globe, posing a threat to lives. UNESCO is leading effects to counter the flood of fake news during the pandemic, as fears are mounting

that it is putting lives in danger[1]. Despite the prevalence of fake news, there is currently no agreement on terminology across communities for false and inaccurate information. We define "misinformation" broadly as circulating information that is false [17]. Moreover, this term is commonly used to refer specifically to when false information is shared accidentally whereas "disinformation" is used to refer to false information shared deliberately [18]. No claims were made in this study concerning the intent of information providers, whether accidental or deliberate. As a result, regardless of purpose, we categorize false information pragmatically. Therefore, the term "misinformation" was used in this study because it is inclusive and not as politicized and polarised as "fake news" as recommended in [19].

In recent years, fake news detection on social media has emerged as a promising field of research and attracting considerable attention. since it can adversely affect both individuals and society. The paper "Fake News Detection on Social Media: A Data Mining Perspective" [20] pointed out several reasons that make fake news detection uniquely challenging. Firstly, fake news is intentionally written to mislead readers, which makes it nontrivial to detect simply based on news content. Moreover, exploiting this auxiliary information leads to another critical challenge regarding data quality. This study gave us a comprehensive review of detecting fake news on social media, including fake news characterizations on psychology and social theories, and existing algorithms from a data mining perspective, including feature extraction and model construction. They also further discussed the datasets, evaluation metrics, and promising future directions in fake news detection research and expand the field to other applications. An exploratory study was conducted by [7] in order to gain early insights about the propagation, authors, and content of misinformation on Twitter around the topic of COVID-19. Two datasets were used for the study. In particular, 1500 tweets were collected between January and mid-July 2020 based on the fact-checked claims related to COVID-19 by over 92 professional fact-checking organizations of which 1274 were labeled as false and 226 were labeled as partially false claims. Meanwhile, the second dataset consists of COVID-19 tweets collected from publicly available corpus TweetsCOV19 and in-house crawling from May-July 2020. Exploratory analysis of author accounts revealed that the verified Twitter handles (including Organization/Celebrity) are also involved in either creating (new tweets) or spreading (retweet) the misinformation. Additionally, they

[1]https://news.un.org/en/story/2020/04/1061592

found that false claims propagate faster than partially false claims and the authors use less tentative language and appear to be more driven by concerns of potential harm to others. [21] proposed a method to classify a tweet as real or fake based on basic features like the tweet hashtags, URLs included, sentiment, the popularity of the tweet (obtained by sentiment analysis), and other features mentioned in the paper. Real-time tweets relating to the pandemic were collected using a tool called Twint by using keywords like coronavirus, covid19, coronavirusPandemic, …. Next, the tweets were manually labeled as fake or real based on the URLs, their source, and other extra information provided in the tweets. The final corrected dataset contained 768 fake tweets and 749 real tweets. Multiple machine learning and deep learning algorithms such as Linear Regression, Decision Tree, Support Vector Machine (SVM), and Long short-term memory are used for comparison, and according to the result, they found that the Random Forest classifier had the best performance compared to other machine learning algorithms used in their project with an accuracy of 84.54% and F1-score of 0.842.

When it comes to detecting fake news, deep learning systems have emerged as a promising solution, mainly due to their ability to provide high precision and accuracy in comparison to traditional machine learning methods. Deep learning has the advantage of being able to acquire hidden representations from less complex inputs, whereas, traditional machine learning approaches produce high-dimensional representations of linguistic information, resulting in the curse of dimensionality, which makes it difficult to achieve prominent results in fake news detection. Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are two widely utilized ideal models for deep learning in cutting-edge artificial neural networks. [22] proposed a hybrid deep learning model that combines CNN and RNN for fake news classification, and the model was successfully validated on two fake news datasets ISO and FA-KES. An optimized CNN model called OPCNN-FAKE [23] based on machine learning and deep learning was presented using grid-search for optimization and n-gram with the TF-IDF feature extraction approach. This method outperformed other models, including regular models like Random Forest, Decision Tree, SVM, Logistic Regression, Naive Bayes, K-Nearest Neighbors, RNN, and LSTM, with a 95.26% accuracy rate when tested on four datasets. The authors of [24] used a neural network based on a deep learning architecture by combining LSTM and CNN with an SVM classifier optimized by stochastic gradient descent (SVM-SGD)

using Pearson's correlation coefficient to detect fake news and untrustworthy people on social networks, achieving 90% accuracy in detecting fake news and 92% accuracy in identifying fake accounts on Twitter. Moreover, [25] used the combination of CNN and RNN, as well as the GloVe embedding technique, to extract features that are critical in determining if an article is fake or real, which achieved an improved accuracy compared to previous models. The authors of [26], instead, modeled the propagation path of news stories as a multivariate time series and build a time series classifier with CNN and RNN to detect fake news. Experimental results show that their proposed model can detect fake news with an accuracy ranging from 84% to 92% among three datasets in 5 minutes, significantly faster than state-of-the-art baselines. Recent deep learning models, such as BERT (Bidirectional Encoder Representation from Transformers), have made significant contributions to NLP. The paper [27] presented a hybrid model named BerConvoNet based on the concatenation of CNN and BERT for fake news classification which adopts a multi-scale feature learning from news articles. Additionally, a self-ensemble SCIBERT (Scientific BERT) based model [28] that utilizes domain-specific word embeddings is proposed to detect health misinformation in news. Furthermore, the study in [29] proposed a hybrid deep model based on behavior information (HMBI) which is the first time to apply BERT, Transformer, and CNN synthetically for fake news detection. The experimental analysis on real-world data shows that the detection accuracy of HMBI is increased by 10.41%, surpassing 50% for the first time.

Because of the pervasive presence of misinformation and its variants in media content, current (especially online) journalism focuses not only on the gathering, processing, and dissemination of information but also on the verification of previously presented knowledge, i.e. the identification of untruths and falsehoods. The given method of detecting false facts is called "debunking" [30]. In scientific research, it is crucial to have a profound understanding of the complex cognitive and perceptual processes of individuals when attempting to debunk misconceptions. This is because it is important to understand how people acquire and process information, how their pre-existing knowledge is affected, and how their personal beliefs and values may impact their logical reasoning abilities. Therefore, the objective of debunking is not solely focused on people's beliefs, but rather on the way in which they interpret and process information [30]. Many in-lab experiments have been conducted to investigate the effects of fact-checking on human behavior, but

the results show significantly diverse behavior in different settings, and fact-check articles posted on social media are likely to get more exposure when shared by a friend instead of strangers [31]. Besides, one of the most concerning notions for science communicators, fact-checkers, and advocates of truth is the "backfire effect." Based on [32], they identified that a backfire effect occurs when an evidence-based correction is presented to an individual and they report believing even more in the very misconception the correction is aiming to rectify. However, the results from a study [33] provide unequivocal support for the benefits of correcting misinformation and suggest that backfire effects are driven substantially more by measurement error or inconsistencies in beliefs rather than the psychological mechanisms proposed to explain them. In addition, many studies [34, 35, 36] showed that while fact-checking helps in correcting misinformation, it only works for some individuals, such as those with higher levels of cognitive ability. To others, exposure to correction might have a boomerang effect and instead, reinforce belief in the original yet incorrect information [35, 37]. However, debunking false information related to autism interventions has been shown to effectively decrease the endorsement of treatments lacking empirical evidence, such as dieting [38]. The presentation of court-ordered corrective advertisements from the tobacco industry regarding the correlation between smoking and illness has the potential to enhance knowledge and diminish misunderstandings about smoking [39]. Additionally, a video that discredits various misconceptions about vaccination has proven to be successful in decreasing influential misunderstandings, such as the erroneous notion that vaccines cause autism or diminish the efficacy of the natural immune system [40]. Consistent meta-analyses have demonstrated that interventions involving fact-checking and debunking can be effective [41, 42], even in the context of addressing health misinformation on social media [43].

Until now, there have been many studies related to the correction of misinformation in terms of semantic or user behaviors. However, there is a lack of research on how to automatically differentiate between misinformation and its refutation on social media platforms, particularly Twitter. Despite ongoing discussions regarding the efficacy of correcting misinformation, it still remains crucial to assist social media users in staying informed about misinformation and thus raising public awareness.

3

BERT and BERTopic

This chapter is divided into two sections in which the BERT model will be discussed in Section 3.1, and Section 3.2 for BERTopic.

3.1 BERT

3.1.1 Transformer

Transformer was first introduced by Vaswani and his team in Google Brain in 2017 with their famous research paper "Attention Is All You Need" [44]. Transformer is a type of artificial neural network architecture used to solve the problem of the transformation of input sequences into output sequences in deep learning applications with an encoder-decoder architecture based on attention layers. It is the current state-of-the-art technique in the field of Natural Language Processing (NLP).

In translation, simply translating word by word in the order of appearance may result in an output that sometimes a native speaker would consider ungrammatical. Thus, the main concern in sequence transduction is learning representations for both the input and output sequences in a robust manner to ensure that no distortions are introduced since mistranslating an important message is terrible. Taking a simple example, *"The cat ran away when the dog chased it down the street"*. Normally, this sentence is very easy to comprehend for a person, but there will be some difficulties

if you think of processing this sequentially. Once you get to the *"it"* part, how do you know what it refers to? In this case, you have to store some states to identify that the key protagonist in this sentence is the *"cat."* Then, you need to find some ways to relate the *"it"* to the *"cat"* as you continue reading the sentence. Practically, the sentence could be any number of words in length and the problem comes since pre-BERT models could only prioritize the importance of words that were most recently processed. As they continued to move along the sentence, the importance or relevance of previous words started to diminish. This is known as the "Vanishing Gradient" problem. Fortunately, Transformers can address this problem by using the mechanism called *"Attention."*

In Psychology, attention is the cognitive process of selectively concentrating on one or a few things while ignoring others. The attention mechanism, also known as attention models, is an attempt to implement the same action of selectively concentrating on a few relevant things while ignoring others in deep neural networks since a neural network is considered to be an effort to mimic human brain actions in a simplified manner. Therefore, in general, attention models are deep learning techniques used to provide an additional focus on a specific component, it relates to focusing on something in particular and noting its specific importance. The model typically focuses on one component within the network's architecture that is responsible for managing and quantifying the interdependent relationships within input elements, called self-attention, or between input and output elements, called general attention. The aim of attention models is to reduce larger, more complicated tasks into smaller, more manageable areas of attention to understand and process sequentially. Attention models evaluate inputs to identify the most critical components and assign each of them a weight. For example, if using an attention model to translate a sentence from one language to another, the model would select the most important words and assign them a higher weight. Similarly, it assigns the less significant words a lower value. This helps achieve a more accurate output prediction.

Transformers have two main blocks, the *encoder* and the *decoder*, each with a *self-attention* mechanism. As indicated above, self-attention refers to the ability of a transformer model to attend to different parts of the same input sequence when making predictions. The encoder basically processes the input text, looks for important parts, and creates an embedding for each word based on relevance to

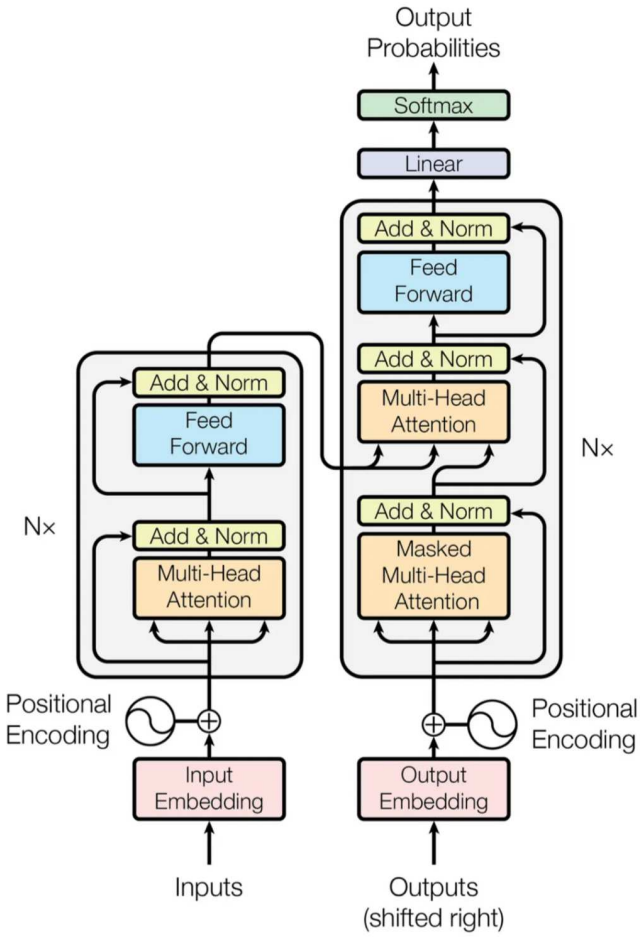


Figure 3.1: Model architecture of Transformer.

other words in the sentence. Whereas, the decoder takes the output of the encoder, which is an embedding, and then turns that embedding back into a text output, which means the translated version of the input text. Figure 3.1 visualizes the model architecture of the Transformer [44]. The encoder is on the left and the decoder is on the right. Encoder and decoder are composed of modules that can be stacked on top of each other multiple times, as described by $N\times$ in Figure 3.1. In addition, the modules consist mainly of Multi-Head Attention and Feed Forward layers.

Computers do not understand words naturally, and it only works on numbers, vectors, or matrices. Therefore, in the beginning, the input and output sentences are first embedded into an n-dimensional space since we cannot use strings directly.

This space is called Embedding Space, and every word, according to its meaning, is mapped and assigned a particular value. However, there is another problem that every word in different sentences has different meanings. Hence, Positional Encoders come to solve the issue. Since we have no recurrent networks that can remember how sequences are fed into a model, we need to somehow give every word in our sentence a relative position since a sequence depends on the order of its elements. These positions are added to each word's embedded representation (n-dimensional vector) and ready to be transmitted to the encoder. The goal of the attention layer is to capture the contextual relationships existing between different words in the input sentence. Multiple attention vectors, usually called Multi-Head Attention Blocks generate an attention vector for each word. For the next step, a feed-forward neural network is applied to every attention vector to transform them into a format that is expected by the next multi-head attention layer in the decoder.

Imagine that if we want to train a translator from English to Italian, we need to give an English sentence along with its translated Italian version for the model to learn during the training phase. Therefore, our English sentences pass through the encoder block, and Italian sentences pass through the decoder block, instead. The decoder block, as shown in Figure 3.1, consists of three main layers: Masked Multi-Head Attention, Multi-Head Attention, and a Feed-forward network. Similar to the encoder part, at the beginning, there are the embedding layer and the positional encoding part. After that, it will pass through the self-attention block, where attention vectors are generated to represent how much each word is related to every word in the Italian sentences. The Masked Multi-head Attention Block is used to hide (or mask) the next Italian word so that it can predict the next word itself using previous results without knowing the actual translated word. Then, the resulting attention vectors from the previous layer and the vectors from the encoder block are passed into another Multi-head Attention Block. This block is also called as the encoder-decoder attention block that aims to map English and Italian words and determines their relationships. The output of this block is attention vectors for every word in the English and Italian sentences, which can be proceeded to a feed-forward unit or a linear layer. The output is then passed through a softmax layer that transforms the input into a probability distribution, which is human interpretable, and the resulting word is produced with the highest probability after translation.

3.1.2 BERT

Bidirectional Encoder Representations from Transformers (BERT) is a Machine Learning model based on Transformers for Natural Language Processing (NLP). This model was created and published in 2018 by Jacob Devlin and his colleagues from Google AI Language [3]. Moreover, BERT achieved groundbreaking results in more than 11 natural language understanding tasks. People interact with NLP (and likely BERT) almost every single day, as BERT is useful for a wide range of language tasks [45], such as:

- **Sentiment Analysis:** Can determine how positive or negative a movie's reviews are.
- **Question answering:** Helps chatbots answer your questions.
- **Text prediction:** Predicts your text when writing an email on Gmail.
- **Text generation:** Can write an article about any topic with just a few sentence inputs.
- **Summarization:** Can quickly summarize long legal contracts.

Given the fact that any specific NLP technique aims to understand human language as it is spoken naturally. A huge amount of data is essential to better recognize patterns in language and identify relationships between words and phrases. Therefore, one of the biggest challenges in NLP is the need for more training data. In order to perform well, deep learning-based NLP models require much more significant amounts of data since they see major improvements when trained on millions, or billions, of annotated training examples. Due to that reason, it is very critical for Machine Learning models like BERT to use *pre-trained* models. Basically, deep learning models (such as Transformers) which are trained on a large dataset to perform specific NLP tasks are called pre-trained models. The idea of pre-trained models is not new in deep learning and has been practiced for many years in image recognition. When trained on a large corpus, pre-trained models can learn the universal language representations, which can be fine-tuned for downstream NLP tasks and therefore, can avoid training a new model from scratch, which is a huge benefit.

In practice, an enormous dataset of 3.3B words has contributed to the success of BERT over the years. Particularly, BERT was specifically trained on English Wikipedia (2,500M words) and Google's BooksCorpus (800M words). Training on a large dataset takes a long time. BERT's training was made possible thanks to the novel Transformer architecture and sped up by using TPUs (Tensor Processing Units - Google's custom circuit built specifically for large Machine Learning models).

Basically, Transformer includes two separate mechanisms which are an encoder that reads the text input and a decoder that produces a prediction for the task. Since BERT's goal is to generate a language model, only the encoder mechanism is necessary. The original Transformer architecture needs to translate text so it uses the attention mechanism in two separate ways: one is to encode the source language, and the other was to decode the encoded embedding back into the destination language. However, BERT uses the encoder part of the Transformer since its goal is to create a model that performs a number of different NLP tasks and using the encoder enables BERT to encode the semantic and syntatic information in the embedding, which is needed for a wide range of tasks. Historically, language models could only read text input sequentially, either left-to-right or right-to-left, but could not do both at the same time [45]. However, the BERT model outperforms since it is designed to read in both directions at once thanks to the advantage of the Transformer encoder. Therefore, it is considered bidirectional, though it would be more accurate to say that it is non-directional. This characteristic allows the model to learn the context of a word based on all of its surrounding which means both the left and right of the word.

Furthermore, when training language models, there is a challenge of defining a prediction goal. Many models predict the next word in a sequence (e.g. "The child came home from ..."), a directional approach that inherently limits context learning. To overcome this challenge, BERT uses two training strategies which are Masked Language Modeling and Next Sentence Prediction.

Masked Language Model (MLM)

The objective of the Masked Language Model training is to hide a word in a sentence and then make the program predict what word has been hidden (masked) based on the hidden word's context provided by the other, non-masked, words in the

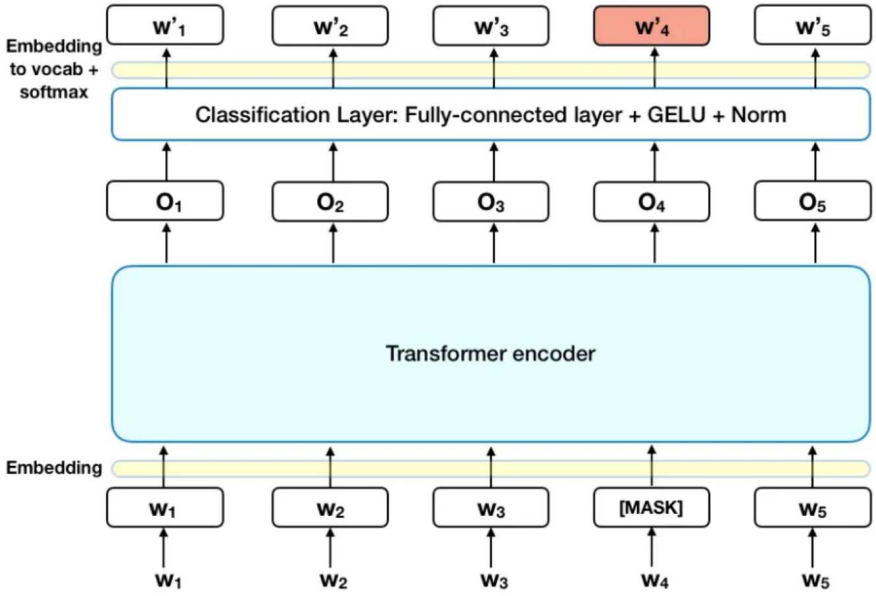


Figure 3.2: Masked Language Model [2].

sequence. A random 15% of tokenized words are hidden during training and BERT's job is to correctly predict the hidden words. In technical terms, the prediction of the output words requires:

- Adding a classification layer on top of the encoder output.
- Multiplying the output vectors by the embedding matrix, transforming them into the vocabulary dimension.
- Calculating the probability of each word in the vocabulary with softmax.

The BERT loss function takes into consideration only the prediction of the masked values and ignores the prediction of the non-masked words. As a consequence, the model converges slower than directional models, a characteristic that is offset by its increased context awareness.

However, in practice, the BERT implementation is slightly more elaborate and does not replace all of the 15% masked words. Particularly, as discussed in an article in Towards Data Science website [2], the author indicated that among 15% of the tokens, 80% are truly replaced with a "[MASK]" token, while 10% with a random word, and 10% use the original word. The intuition that led the authors to pick this approach is as follows:

- If we used [MASK] 100% of the time the model would not necessarily produce good token representations for non-masked words. The non-masked tokens were still used for context, but the model was optimized for predicting masked words.
- If we used [MASK] 90% of the time and random words 10% of the time, this would teach the model that the observed word is never correct.
- If we used [MASK] 90% of the time and kept the same word 10% of the time, then the model could just trivially copy the non-contextual embedding.

No ablation was done on the ratios of this approach, and it may have worked better with different ratios. In addition, the model performance was not tested with simply masking 100% of the selected tokens.

Next Sentence Prediction (NSP)

The objective of Next Sentence Prediction training is to have the program predict whether two given sentences have a logical, sequential connection or whether their relationship is simply random. In training, 50% of correct sentence pairs are mixed in with 50% random sentence pairs to help BERT increase next sentence prediction accuracy. Assuming that the random sentence will be disconnected from the first sentence. In order to help the model distinguish between the two sentences in training, the input is processed in the following way before entering the model:

- A `CLS` token is inserted at the beginning of the first sentence and a `SEP` token is inserted at the end of each sentence.
- A sentence embedding indicating Sentence A or Sentence B is added to each token. Sentence embeddings are similar in concept to token embeddings with a vocabulary of 2.
- A positional embedding is added to each token to indicate its position in the sequence. The concept and implementation of positional embedding are presented in the Transformer paper.

To predict if the second sentence is indeed connected to the first, the following steps are performed:

- The entire input sequence goes through the Transformer model.

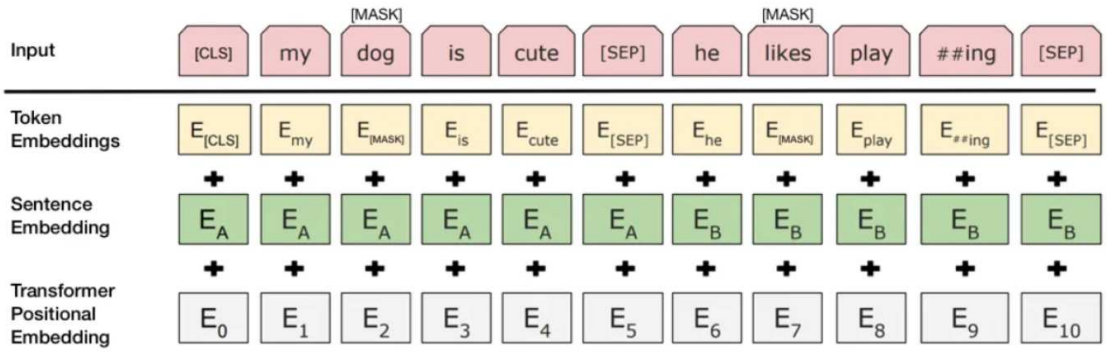


Figure 3.3: Next Sentence Prediction [3].

- The output of the `CLS` token is transformed into a 2x1 shaped vector, using a simple classification layer (learned matrices of weights and biases).
- Calculating the probability of IsNextSequence with softmax.

When training the BERT model, Masked Language Models and Next Sentence Prediction are trained together, with the goal of minimizing the combined loss function of the two strategies.

3.2 BERTopic

Topic models have proven to be a powerful unsupervised tool to uncover common themes and the underlying narrative in text. Given that most of the information we generate and exchange as human beings has a textual nature. It usually comes from some sources such as news articles, social media posts, messages, emails, and conversations. Data Science has been dealing with the problem of automatically extracting value from these sources without (or with limited) prior knowledge for a very long time. This is the reason why topic modeling is considered as an unsupervised Machine Learning problem, in which unsupervised means that the algorithm learns patterns in the absence of tags or labels. There exist some conventional and well-known models, such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF), these models describe a document as a bag-of-words and model each document as a mixture of latent topics. However, the limitation of these models is that through bag-of-words representations, they disregard semantic relationships among words. As these representations do not account for the

context of words in a sentence, the bag-of-words input may fail to accurately represent documents. Text embedding techniques have rapidly become popular in the natural language processing field to overcome this issue. The semantic properties of these vector representations allow the meaning of texts to be encoded in such a way that similar texts are close in vector space [46]. In this study, we chose to use transformer-based models such as BERT as they have shown amazing results in various NLP tasks over the last few years. Furthermore, pre-trained models are especially helpful as they are supposed to contain more accurate representations of words and sentences. This model is called BERTopic, which was introduced by Maarten Grootendorst in 2022 [46].

BERTopic is a topic modeling Python library that combines transformer embeddings and clustering model algorithms to identify topics in NLP in which it extracts coherent topic representation through the development of a class-based variation of TF-IDF. In general, BERTopic generates documents embedding with pre-trained transformer-based language models, clusters these embeddings, and finally, creates topic representations with the class-based TF-IDF procedure.

3.2.1 Documents Embedding

Embeddings are an important part of text interpretation in order to make textual data understandable to machine learning and NLP models. Recalling that *Word embedding* is a representation of a word in multidimensional space such that words with similar meanings have similar embedding. It means that each word is mapped to the vector of real numbers that represent the word. In addition, *Document embedding* is usually computed from the word embeddings in two steps. First, each word in the document is embedded with the word embedding, then word embeddings are aggregated. The most common type of aggregation is the average over each dimension. Therefore, they are basically the vector representation of the documents. So that, at the very first step, we convert the documents to numerical data.

BERTopic supports many embedding models that can be used to embed the documents and words, such as Sentence-Transformers, Hugging Face Transformers, Flair, Spacy, Gensim, and USE (Universal Sentence Encoder). As the default, BERTopic uses Sentence-BERT (SBERT) framework to get document embeddings from a set

of documents. This framework allows users to convert sentences and paragraphs to dense vector representations using pre-trained language models. Principally, the default sentence transformer model is `all-MiniLM-L6-v2` which is a popular high-performing model that creates 384-dimensional sentence embeddings. Therefore, the quality of clustering in BERTopic will increase as new and improved language models are developed. This allows BERTopic to continuously grow with the current state-of-the-art in embedding techniques [46]. Moreover, BERTopic can also perform the topic modeling of over 50 languages.

3.2.2 Documents Clustering

In order to have a good performance, we need to guarantee that documents with similar topics are clustered together such that we are able to find the topics within these clusters. However, since working with high-dimensional data usually refers to the Curse of Dimensionality in which the increase of data dimensions implies to an exponential increase in computational efforts required for its processing and/or analysis. Despite clustering approaches exist for overcoming this problem, a more simpler approach is to reduce the dimensionality of embeddings. Therefore, in the next step, after building the embeddings, BERTopic compresses them into a lower-dimensional space before running a clustering model since the embedding vectors usually have very high dimensions. In terms of dimension reduction techniques, PCA and t-SNE are well-known methods which have been using more frequently when dealing with this issue. However, PCA method works by preserving larger distances using mean squared error so that the global structure of the data is usually maintained. Meanwhile, t-SNE technique tries to preserve the local structure of the data by minimizing the Kullback-Leibler divergence between the two distributions with respect to the locations of the points in the map. In order to capture the best of both methods, BERTopic uses UMAP (Uniform Manifold Approximation and Production) to perform dimension reduction. For each datapoint, UMAP searches through other points and identifies the k-th nearest neighbors where k is controlled by the `n_neighbors` parameter. By increasing `n_neighbors` we can preserve more global structures, whereas a lower `n_neighbors` better preserves local structures. Therefore, UMAP mantains distinguishable features that are not preserved by PCA and a better global structure than t-SNE.

Clustering is a Machine Learning technique that involves the grouping of data points. Theoretically, data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features. Clustering methods can be broken into flat or hierarchical and centroid or density-based techniques. Basically, flat or hierarchical focuses simply on whether there is (or is not) a hierarchy in the clustering method. Flat clustering requires a prior understanding of the clusters as we have to set the resolution parameter such as `k` in K-means and `eps` in DBSCAN (Density-Based Spatial Clustering of Applications with Noise), whereas, hierarchical clustering let the machine decide how many clusters to create based on its own algorithms. For the other split between centroid-based or density-based, the clustering based on proximity to a centroid or clustering based on the density of points. The centroid-based is ideal for "spherical" clusters, whereas density-based clustering can handle more irregular shapes and identify outliers. BERTopic uses HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) which is a hierarchical, density-based method to cluster the reduced embeddings. This algorithm also works quite well with UMAP since UMAP maintains a lot of local structures even in lower-dimensional space. Furthermore, HDBSCAN algorithm uses a soft-clustering approach allowing noise to be modeled as outliers. This prevents unrelated documents to be assigned to any cluster and is expected to improve topic representations [46].

3.2.3 Topic Representation

After assigning each document in the corpus into a cluster, the next step is to get the topic representation using a class-based TF-IDF called c-TF-IDF where the top words with the highest c-TF-IDF scores are selected to represent each topic. Basically, TF-IDF is a measure for representing the importance of words between documents by computing the frequency of a word in a given document and also measure how prevalent the word is in the entire corpus. When you apply TF-IDF as usual on a set of documents, what you are doing is comparing the importance of words between documents. The classic TF-IDF procedure combines two statistics, term frequency, and inverse document frequency [47]

$$W_{t,d} = tf_{t,d} \cdot \log(\tfrac{N}{df_t}),$$

where the term frequency $tf_{t,d}$ models the frequency of term t in document d. The inverse document frequency $\frac{N}{df_t}$ measures how much information a term provides to a document and is calculated by taking the logarithm of the number of documents in a corpus N divided by the total number of documents that contain t.

As described in [46], the author generalized this method to clusters of documents by treating all documents in a single cluster as a single document and then performing TF-IDF. The result would be the important scores for words within a cluster instead of individual documents

$$W_{t,c} = tf_{t,c} \cdot \log(1 + \frac{A}{df_t}),$$

where the term frequency $tf_{t,c}$ models the frequency of term t in a class c in which the class c is the collection of documents concatenated into a single document for each cluster. Then, the inverse document frequency $\frac{N}{df_t}$ is replaced by the inverse class frequency $\frac{A}{df_t}$ to measure how much information a term provides to a class. It is calculated by taking the logarithm of the average number of words per class A divided by the frequency of term t across all classes and adding one to the division within the logarithm to guarantee only positive values for the output. After that, in order to create a topic representation, the top 20 words per topic based on their c-TF-IDF scores are chosen. The more important words are within a cluster, the more representative they are of that topic.

4 Dataset

In this chapter, we describe the steps involved in the data collection and the method for the preprocessing of the tweets for analysis. We used a total of five datasets for our study. The first dataset consists of false claims and debunking explanations associated with COVID-19; the other four are tweets related to four misinformation topics which were downloaded by Twitter API with keywords linked to these topics. This chapter is divided into three sections: Data collection, Preprocessing process, and Training dataset in Section 4.1, Section 4.2, and Section 4.3, respectively.

4.1 Data collection

4.1.1 Poynter database

With the International Fact-Checking Network (IFCN) agreement, we can access the CoronaVirusFacts database [11] to extract the topics of misinformation related to COVID-19. The dataset contains 10448 titles to fact-check in which titles are classified as false, accounting for over 82% as shown in Figure 4.1. Moreover, this study aims to focus on misinformation and to simplify, we only extract false titles, which are 8660 samples for further study; Figure 4.2 shows the Poynter dataset in which false titles belong to "What did you fact-check?" column. In addition, Figure 4.3 visualizes the frequency of misinformation per week from January 2020

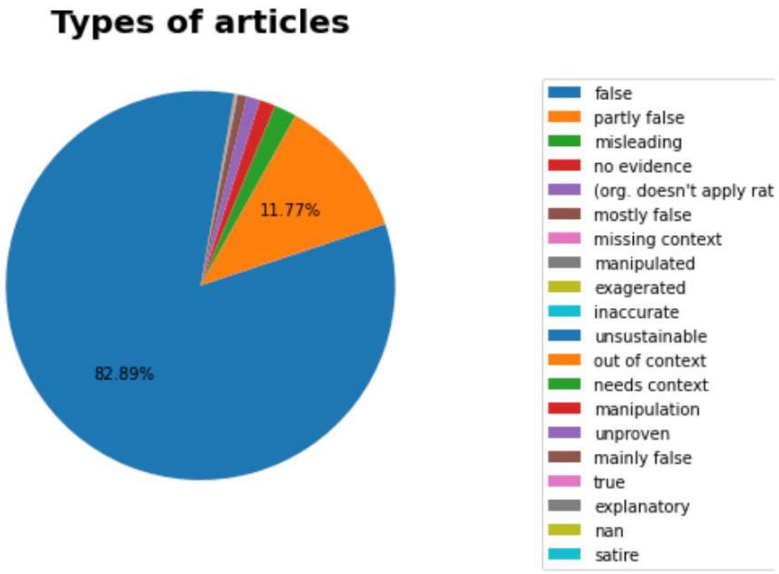


Figure 4.1: Percentage of article types in the Poynter dataset.

	When did you see the claim?	Countries	Organization	What did you fact-check?	Who said/posted it?	Link to the original piece	Language of your fact-check	Final rating	Explanation	Category
0	1/14/2020	[Philippines]	[rappler]	A chain message circulated on Tuesday, Jan. 14...	[chain message]	https://www.facebook.com/rowena.molina.142/pos...	English	False	The Department of Health (DOH) and Healthway M...	[spread]
1	1/18/2020	[Mexico]	[animal político]	Stores and supermarkets in Veracruz (Mexico) w...	[whatsapp]	NaN	Spanish	False	As of Mar. 18, stores had not said they would ...	[authorities]
2	1/21/2020	[France, United States]	[afp]	The coronavirus was created in a lab and paten...	[facebook, website]	https://perma.cc/79T7-RFKC	English	False	The patents circulating are linked to other co...	[conspiracy theory]
3	1/22/2020	[China]	[afp]	Saline solutions can kill the new coronavirus.	[many social media platforms]	https://perma.cc/M893-58WA	English	False	The WHO and many doctors deny this claim.	[cures]
4	1/22/2020	[Sri Lanka]	[afp]	You have to wear a disposable mask with the bl...	[facebook]	https://perma.cc/KJ9V-3DDL	English	False	The only way to wear a mask is with the blue f...	[other]

Figure 4.2: Example of Poynter dataset.

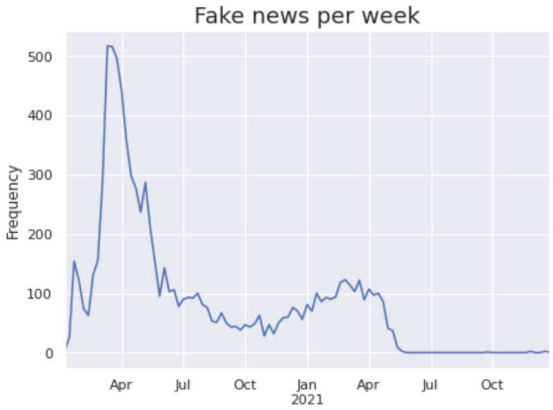


Figure 4.3: Frequency of misinformation per week in the Poynter dataset.

to December 2021. It was around April 2020 that misinformation reached its peak. Therefore, we chose a one-year period starting on July 2020 to evaluate their influence on social media and how they were debunked since it takes time for false information to circulate online before it can be debunked.

Figure 4.4 visualizes topics extracted from fake titles in the Poynter dataset by using BERTopic. It can be seen from the graph that the main topics are directly related to "5G technology" (Topic 14), "Vaccines" (Topic 8 and 13), "Bill Gates" (Topic 8), and "COVID-19's treatment - Cure" (Topic 4, 5, 10, 16, and 18). Therefore, in this study, we chose to focus on these four macro topics to distinguish between three types of tweets which are false, neutral, and debunked during COVID-19.

4.1.2 Tweets

Twitter data is widely used in the world of Natural Language Processing. The Twitter API is a set of programmatic endpoints that can be used to understand or build the conversation on Twitter. An API, short for Application Programming Interface, is a way for two or more computer programs to communicate with each other. This API allows us to find, retrieve, engage with, or create various resources like tweets, users, spaces, ... As mentioned, this study used the datasets downloaded by the Twitter API with the default English language and excluded retweets. In order to evaluate the influence of misinformation and how they were debunked on Twitter, we decided to focus on the four macro topics indicated in Section 4.1.1. To find keywords for downloading tweets, we visualized WordCloud plots for each macro topic of misinformation in Figures 4.5.

For the first topic, "Bill Gates," based on Figure 4.5, we can see that besides the keyword "Bill Gates," there are also some other helpful words like "vaccine," "population," "depopulate," and "Melinda." Since we have a Vaccine topic, we chose the keywords "Bill Gates" and "depopulation" to download tweets related to this misinformation. Next, for the topic related to COVID-19's treatment, "Cure," due to the high frequency of the words "hydroxychloroquine" and "chloroquine" appeared in Figure 4.5 together with "cure," "covid," and "coronavirus." For this topic, we decided to use the keywords "cure" and "hydroxychloroquine" for downloading tweets in this misinformation. For the topic directly related to vaccines,

Macro Topic	Keywords for Twitter API	# Tweets per day	# Tweets collected
Cure	(cure ∨ hydroxychloroquine) & covid19	100	28703
Bill Gates	bill gates & depopulation	100	
	bill gates & covid19	60	18772
5G	5g & (coronavirus ∨ (vaccines & covid19))	100	14754
Vaccines	vaccines & kill & covid19	100	
	vaccines & covid19	60	24868

Table 4.1: List of keywords used for tweets collection in the period July 1, 2020-June 30, 2021.

in order to focus on COVID-19 vaccines and also target the community against vaccines, we chose the keywords "vaccines" and "kill" to download tweets related to this misinformation. Finally, for the topic related to 5G, given that there were many rumors about the relationship between the fifth-generation mobile network and coronavirus, which can be seen in Figure 4.5 that these two keywords have the highest frequency together with "WhatsApp" and "covid." Therefore, the keywords for this misinformation are "5G" and "coronavirus."

The specific keywords used for downloading tweets were identified by inspecting the wordcloud plots of Figure 4.5, and led to the keyword choice summarized in Table 4.1. A maximum of 100 tweets per day were downloaded from Twitter (depending on their availability) by using the search in Table 4.1. For "Bill Gates" and "Vaccines" some additional 60 tweets per day were downloaded by using a weaker search, as the total number of collected tweets was in this case limited. In this way we guaranteed a rich and evenly distributed search along the observation period.

4.2 Preprocessing of tweets dataset

In order to have a better performance, preprocessing data plays a key role in Machine Learning. For each tweet, we define a function to clean the date and time format. Notably, the original data's default date and time format is ISO 8601, which represents date and time by starting with the year, followed by the month, day, hour, minutes, seconds, and milliseconds. To simplify the further analysis, only the day, month, and year were kept. Moreover, tweets were lowercase and removed URLs (abbreviation for Uniform Resource Locator). Then, all mentioned accounts were removed, and only the hashtag symbol (#) was deleted for the hashtags. We kept the words of these hashtags because Twitter users often use them

to convey essential information. Also, in specific sentences, they can be used as substitutes for ordinary words, as shown in Figure 4.6, so removing them would lose essential words. Hereafter, stop words, which are the most common words in any language (such as articles, prepositions, pronouns, and conjunctions) and do not add much information to the text, were removed from tweets. However, the negation word "not" is considered to be a stop word in Natural Language Toolkit (NLTK), and due to the sensitivity of topics related to COVID-19, removing "not" may completely change the meaning of the sentence, for example, this sentence "5G is not causing coronavirus" change to "5G is causing coronavirus" makes the original meaning of the sentence changed entirely. Consequently, all stop words in the sentence were eliminated except the word "not". Then, cleaned tweets were tokenized using NLTK. Additionally, tokenized words were joined again for topic extraction, and we will talk more about this in Chapter 6. Besides, hashtags, mentioned accounts, and retweet/ like/ reply counts were extracted for the exploratory data analysis. Figure 4.7 shows an example of a final dataset after preprocessing.

4.3 Training dataset

As mentioned in Chapter 2, until now, there is no available tweets dataset in which tweets are divided into three categories false, neutral, and debunked. Due to that limitation, in other to have data for training a classification model, a portion of tweets in the four macro topics were manually labeled. Figures 4.8, 4.9, and 4.10 show examples of tweets classified as false, neutral, and debunked, respectively. In particular, 600 tweets were classified as neutral, and for false and debunked categories, only 228 and 200 were labeled, respectively. Besides neutral tweets, the number of tweets in the other two groups are small. Therefore, we extracted and filtered the information from the Poynter database corresponding to the topics indicated in Section 4.1 with false titles as false claims and explanations as debunked claims and then combining with the labeled tweets before having the final training dataset. Finally, the training dataset has 1800 claims, with 600 claims for each type of tweet, as summarised in Table 4.2.

For convenience, tweets were labeled as 0 for false, 1 for neutral, and 2 for debunked. Figure 4.11 shows the first five samples in the training dataset.

Cure	False	209	**5G**	False	140
	Neutral	115		Neutral	176
	Debunked	183		Debunked	114
Bill Gates	False	118	**Vaccines**	False	133
	Neutral	180		Neutral	129
	Debunked	117		Debunked	186

Table 4.2: Data in the training dataset, organised by class (false/neutral/debunked)

Figure 4.4: Topics extraction based on the Poynter database.

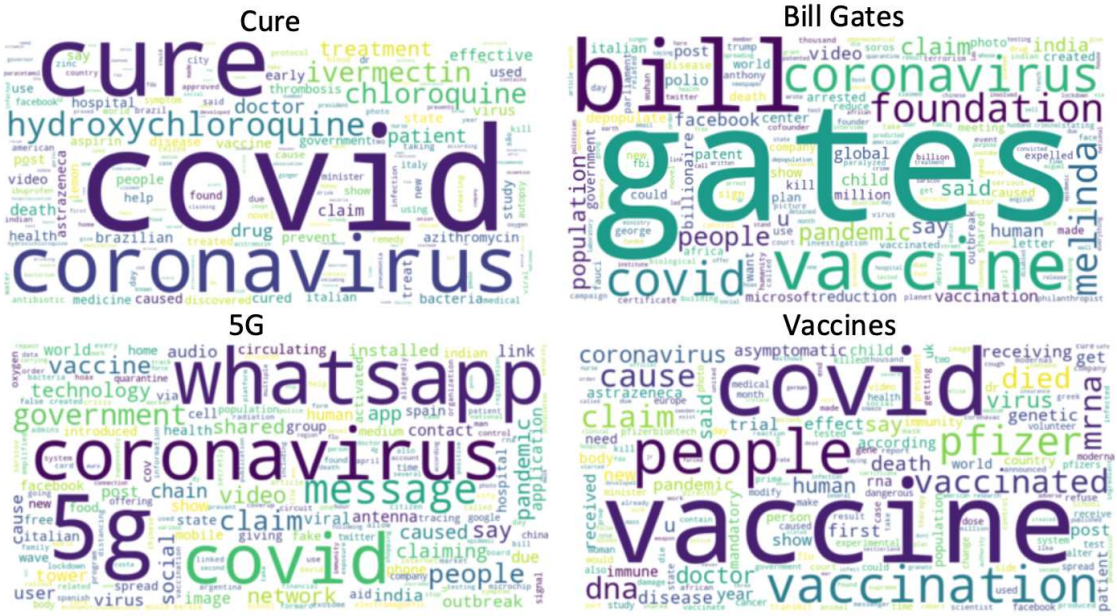


Figure 4.5: WordCloud for misinformation in the four macro topics.

Maria C #EnoughisEnough
@Maria4CarmsEast

2 facts for those who think #5G causes #coronavirus. There is only 1 country in the world with full 5G cover - Monaco.
It also has 96 cases & 4 deaths.
So how can anyone believe that it's 5G?
#ConspiracyTheory

7:29 PM · May 16, 2020

24 Retweets **3** Quote Tweets **42** Likes

Figure 4.6: Example of a tweet used hashtags to express their thought.

	text	date	clean_text	text_tokenized	text_without_stopwords	hashtags	mentions	retweet_count	like_count	reply_count
0	"No, 5G isn't causing coronavirus 👑" https://...	2020-07-01	"no, 5g is not causing coronavirus 👑"	[5g, not, causing, coronavirus]	5g not causing coronavirus	[]	[]	0	0	0
1	Ep99 #coronavirus #latteart OMG it's #worldufo...	2020-07-01	ep99 coronavirus latteart omg it is worldufoda...	[ep99, coronavirus, latteart, omg, worldufoday...	ep99 coronavirus latteart omg worldufoday toda...	[latteart, covid19, coronavirus, worldufoday, ...	[]	0	0	0
2	Coronavirus: 5G and microchip conspiracies aro...	2020-07-01	coronavirus: 5g and microchip conspiracies aro...	[coronavirus, 5g, microchip, conspiracies, aro...	coronavirus 5g microchip conspiracies around w...	[]	[]	0	0	0
3	#Peston Get The Picture yet ? #coronavirus #5G...	2020-07-01	peston get the picture yet ? coronavirus 5g	[peston, get, picture, yet, coronavirus, 5g]	peston get picture yet coronavirus 5g	[5g, coronavirus, peston]	[]	1	0	0
4	When will MISTER Joreny Corfym comment on the ...	2020-07-01	when will mister joreny corfym comment on the ...	[mister, joreny, corfym, comment, connection, ...	mister joreny corfym comment connection corona...	[]	[]	6	56	1

Figure 4.7: Example of a dataset after cleaning.

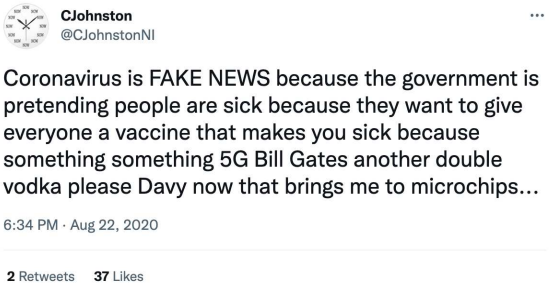


Figure 4.8: Example of tweet classified as false.

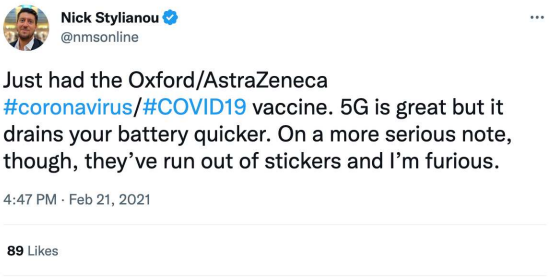


Figure 4.9: Example of tweet classified as neutral.

Figure 4.10: Example of tweet classified as debunked.

	claim	label	class
0	obama, bill gates, and hillary duff run antif...	false	0
1	united we stand!! divided we fall!! american p...	neutral	1
2	there's no covid19 in this world ma bru.. th...	false	0
3	so how bill gates knew about the coronavirus i...	false	0
4	he is not for freedom of speech he is for the ...	false	0

Figure 4.11: Example of tweets in training dataset.

5 Models Comparison

In this chapter, we introduce the BERT model architecture used for classifying tweets. Moreover, to reinforce the final decision of choosing the BERT model, we evaluate the performance of this model with some other machine learning algorithms such as Stochastic Gradient Descent Classifier (SGD), Multinomial Naive Bayes (Multinomial NB), Random Forest, Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN).

5.1 BERT model for Text Classification

The BERT model for text classification used in this study was inspired by Nicolo Cosimo Albanese [4]. This model requires the preprocessing steps as visualized in Figure 5.1. Firstly, we need to add two special tokens which are `[CLS]` and `[SEP]` at the beginning and at the end of the sentence, respectively. Then, each sentence is transformed in order to have the same length. This is achieved by *padding*, which means adding values of convenience to shorter sequences for matching the desired length, and longer sequences are truncated. The maximum sequence length allowed is 512 tokens. Generally, the padding `[PAD]` tokens have ID 0, the `[CLS]` and `[SEP]` tokens have IDs 101 and 102, correspondingly. Finally, an attention mask is created with a list of 1 and/or 0 indicating whether the model should consider the

Figure 5.1: Preprocessing steps for BERT [4].

tokens when learning their contextual representation. We expect `[PAD]` tokens to have value 0. Moreover, since the `max_length` was set as 252 in order to capture as much information as possible, longer sentences will be truncated, while shorter sentences will be populated with `[PAD]` tokens (ID 0) until they reach the desired length.

Then, we split the dataset into the train (80%) and validation (20%) sets and wrap them around a `torch.utils.data.DataLoader` object. With its intuitive syntax, `DataLoader` provides an iterable over the given dataset.

In order to achieve better performance, we fine-tune the model based on the recommendations from the BERT paper [3]. As indicated in the paper, the optimal hyperparameter values are task-specific, but the following range of possible values works well across all tasks:

1. Batch size: 16, 32
2. Learning rate (Adam): 5e-5, 3e-5, 2e-5
3. Number of epochs: 2, 3, 4

After evaluating the model with these hyperparameter values, the final model which has batch size 32, learning rate 5e-5, and 4 epochs outperforms other combinations with 88.9% accuracy on our validation set.

5.2 Models comparison

As indicated above, we compare the performance of the BERT model with the other five machine learning models which are Stochastic Gradient Descent, Multinomial Naive Bayes, Random Forest, Long Short-term memory, and Convolutional Neural Network. Furthermore, for the purpose of making this study more objective, not only the training dataset is used but also another dataset simply called the Poynter$^+$ dataset in order to distinguish from the original Poynter dataset. In this Poynter$^+$ dataset, we use 1200 claims consisting of fake titles and the respective explanations which directly related to the four macro topics, and we labeled them as false and debunked labels correspondingly. For the neutral labels, we just simply use the neutral tweets of the training dataset. Therefore, in the end, this Poynter$^+$ dataset also has 1800 samples with 600 samples for each class which is similar to the training dataset.

Figures 5.2 and 5.3 visualize the confusion matrices and evaluation metrics comparison among six models. In general, the performance of six models with the Poynter$^+$ dataset is better than the training dataset. Moreover, in both datasets, the BERT model has always outperformed the other five models. Particularly, the neutral group was easier to recognize in the Poynter$^+$ dataset than the false and debunked ones. Especially, the BERT model did very well in classifying neutral tweets among the false claims, and also the debunked claims among the neutral tweets in the Poynter$^+$ dataset which can be seen in the confusion matrix with 0% in these two cells. However, Multinomial Naive Bayes had the worst performance compared to the other models, particularly in classifying the false group with only 65% of correct labeling. Similarly, the Random Forest algorithm did not classify very well the debunked group since only 55.83% of the debunked samples were correctly labeled. Besides, the Stochastic Gradient Descent, LSTM, and CNN performed quite well in the range of 80% to over 86% of correctness. Meanwhile, with the training dataset, debunked tweets seemed easier to be distinguished rather than the other two types of tweets. It can be seen in Figure 5.2 that among the total of six models, except the Random Forest, the correct classification of the debunked tweets in the other five algorithms is very high and it is always the highest value compared to the other two groups of tweets. Since debunked tweets may have a completely different way of conveying the message than false tweets and neutral tweets. The BERT

Figure 5.2: Confusion matrices comparison.

Study	Metrics	Machine Learning Models					
		SGD	**Multinomial NB**	**Random Forest**	**LSTM**	**CNN**	**BERT**
Poynter+ Dataset	Accuracy	86%	73%	75%	83%	84%	96%
Training Dataset		76%	68%	71%	72%	72%	89%
Poynter+ Dataset	Precision	False: 82% Neutral: 92% Debunked: 85%	False: 68% Neutral: 93% Debunked: 61%	False: 64% Neutral: 89% Debunked: 76%	False: 78% Neutral: 89% Debunked: 82%	False: 79% Neutral: 87% Debunked: 85%	False: 94% Neutral: 99% Debunked: 95%
Training Dataset		False: 78% Neutral: 74% Debunked: 76%	False: 65% Neutral: 73% Debunked: 65%	False: 66% Neutral: 67% Debunked: 81%	False: 64% Neutral: 64% Debunked: 89%	False: 69% Neutral: 65% Debunked: 85%	False: 86% Neutral: 87% Debunked: 94%
Poynter+ Dataset	Recall	False: 86% Neutral: 86% Debunked: 87%	False: 65% Neutral: 82% Debunked: 71%	False: 84% Neutral: 84% Debunked: 56%	False: 82% Neutral: 84% Debunked: 83%	False: 80% Neutral: 87% Debunked: 84%	False: 95% Neutral: 99% Debunked: 94%
Training Dataset		False: 63% Neutral: 74% Debunked: 89%	False: 47% Neutral: 76% Debunked: 81%	False: 66% Neutral: 78% Debunked: 68%	False: 63% Neutral: 72% Debunked: 79%	False: 65% Neutral: 75% Debunked: 77%	False: 82% Neutral: 87% Debunked: 97%
Poynter+ Dataset	F1-score	False: 84% Neutral: 89% Debunked: 86%	False: 67% Neutral: 87% Debunked: 66%	False: 72% Neutral: 87% Debunked: 64%	False: 80% Neutral: 87% Debunked: 83%	False: 80% Neutral: 87% Debunked: 85%	False: 95% Neutral: 99% Debunked: 95%
Training Dataset		False: 70% Neutral: 74% Debunked: 82%	False: 54% Neutral: 74% Debunked: 72%	False: 66% Neutral: 72% Debunked: 74%	False: 64% Neutral: 68% Debunked: 84%	False: 67% Neutral: 69% Debunked: 81%	False: 84% Neutral: 87% Debunked: 96%

Figure 5.3: Evaluation metrics comparison.

model classified very well the debunked tweets with over 97% of the samples were correctly labeled. However, the other five models did not distinguish effectively the false tweets, with at most 65.83% of correct labeling. In terms of evaluation metrics, Figure 5.3 shows the comparison between accuracy, precision, recall, and F1-score among six models and between the two datasets. Similar to the confusion matrices shown above, the BERT model has outperformed the other five models and better results are seen in the Poynter$^+$ dataset. The Poynter$^+$ dataset consists of two different types of data which are neutral tweets and non-tweet types for the false and debunked groups, therefore, it is easier for the model especially BERT to recognize the pattern of neutral class with the others, as we can notice that the model has 99% of F1-score for the neutral group. Consequently, when applying this model to the practical data, it will be likely to predict the "neutral" class, so we do not actually have a good metric. Simultaneously, the training dataset only consists of a portion of non-tweet types for the false and debunked groups which are from the Poynter$^+$ dataset. Despite the performance is not very well compared to the Poynter$^+$ dataset but the model can learn better the pattern of tweets when categorizing them into three classes. Furthermore, as indicated above, the debunked tweets somehow are more unique in comparison with the false and neutral tweets, since the BERT model returned a significantly high score for the debunked group in the training dataset, at 96% of the F1-score.

In conclusion, after considering the performance of six models among two datasets, the BERT model which was trained with the training dataset as defined in Section 5.1 is used for the classification of tweets in the four macro topics which are "Cure," "Bill Gates," "5G," and "Vaccines."

6 Results

In this Chapter, we discuss the results of the proposed methods. Particularly, Section 6.1 shows the results from the BERT classification model while the word cloud plots about texts and hashtags used in tweets will be discussed in Section 6.2. Temporal behaviors and topic document visualization are analyzed in Sections 6.3 and 6.4, respectively. Finally, the semantic aspects of tweets will be considered in Section 6.5.

6.1 BERT classification results

Figure 6.1 visualized the percentage of three types of tweets in misinformation related to the four macro topics of COVID-19 after using the BERT classification model defined in Section 5.1.

In general, there was a bit of similarity in the performance between datasets of the four macro topics. In particular, neutral tweets dominated in each of the four datasets with at least 68.2% to at most 85%, followed by false tweets. Moreover, the percentages of false tweets on the topics of Bill Gates and Vaccines were quite prominent compared to the ones for Cure and 5G. By contrast, debunked tweets on the topics of Cure and 5G accounted for a higher proportion in comparison with the topics of Bill Gates and Vaccines. The number of debunked tweets, however, made

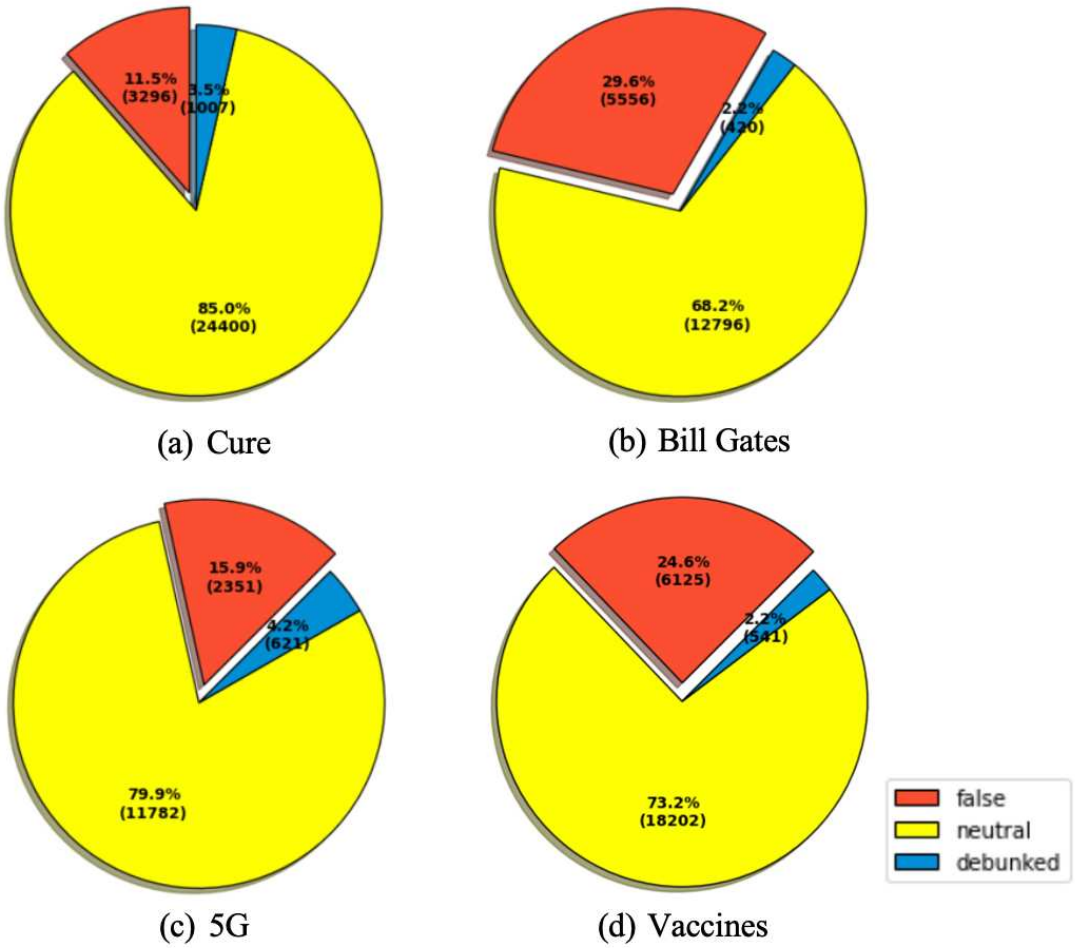


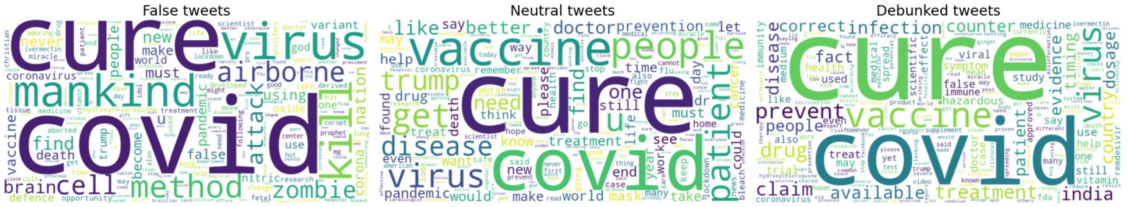
Figure 6.1: Percentage of false, neutral, and debunked tweets in the four macro topics.

up a relatively small percentage of the total number of tweets, which was ranging from 2.2% to 4.2%.

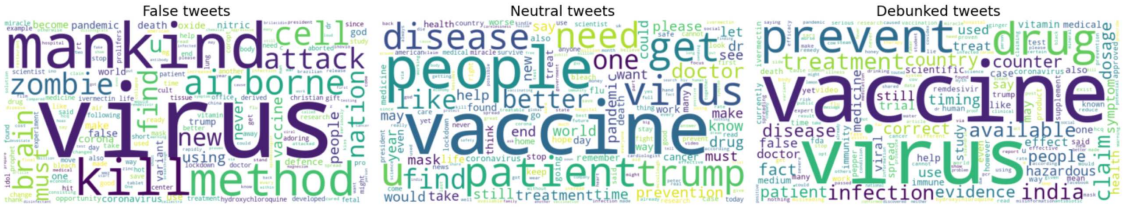
6.2 Word Clouds

Cure

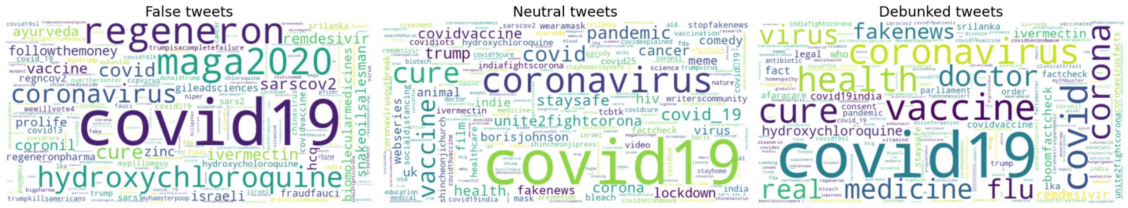
Figure 6.2a visualized the word cloud plot of tweets with misinformation related to the Cure topic. Due to the fact that at the current time, COVID-19 has not had a



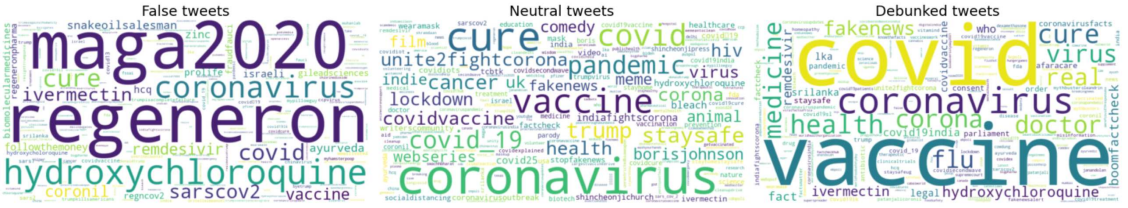
(a) Words used in tweets.



(b) Words used in tweets without "covid," and "cure."



(c) Hashtags used in tweets.



(d) Hashtags used in tweets without #covid19.

Figure 6.2: WordCloud plots of tweets related to Cure.

cure yet and most misinformation on social media tries to target some treatments and promote everyone using them as a cure for COVID-19. It can be seen in the Figure that the two words "cure" and "covid" had the highest frequency among other words. Therefore, in Figure 6.2b these two words were removed so that we can see other words' frequency without being dominated too much by the main keywords. It can be seen that no word related to a specific treatment is superior to the others. Furthermore, in the false group, a huge portion of people used the word "virus" in their tweets together with the word "kill" in order to refer to misinformation that something can kill the COVID-19 virus. Some pronouns were used mostly in neutral tweets compared to false and debunked ones like "people" and "patient". Moreover, people usually discussed "Trump" and talked about many general things like "disease", "mask", "pandemic", and "cancer", together with many verbs that had a high frequency such as "would", "find", and "get" in neutral tweets while in false tweets they preferred using "must" which is a strong verb, and in debunked tweets, they used the verb "prevent" more than other verbs. Moreover, in debunked tweets, people tended to use the word "claim" when referring to a piece of misinformation in which "claim" is defined as a statement that something is true although it has not been proved and other people may not agree with or believe it[1]. Additionally, some words related to debunking also were used such as "evidence", "fact", "false", "study", and "scientific."

Hashtags used in tweets are visualized in Figure 6.2c. After removing the hashtags #covid19 as shown in Figure 6.2d, we can see that all tweets in three categories used #cure and #vaccine. In false tweets, a huge portion of hashtags used are #maga2020[2] and #regeneron which is the name of a biotechnology company. It can be noticed that in false and debunked tweets, they appeared the hashtags #hydroxychloroquine and #ivermectin, where Hydroxychloroquine is a disease-modifying anti-rheumatic drug (DMARD) that is used to treat malaria. Ivermectin is instead an FDA-approved[3] antiparasitic drug used to treat several neglected tropical diseases, including onchocerciasis, helminthiases, and scabies. For the neutral group,

[1]Oxford Learner's Dictionaries

[2]"maga" means "Make America Great Again"

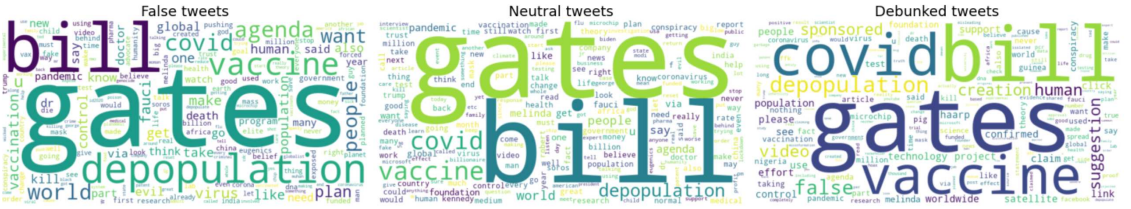
[3]FDA (Food and Drug Administration) approval of a drug means that data on the drug's effects have been reviewed by CDER (Center for Drug Evaluation and Research), and the drug is determined to provide benefits that outweigh its known and potential risks for the intended population

people tended to use hashtags like #unite2fightcorona, #staysafe, and a small portion of tweets used #factcheck and #fakenews, whereas, mostly debunked tweets used #fakenews.

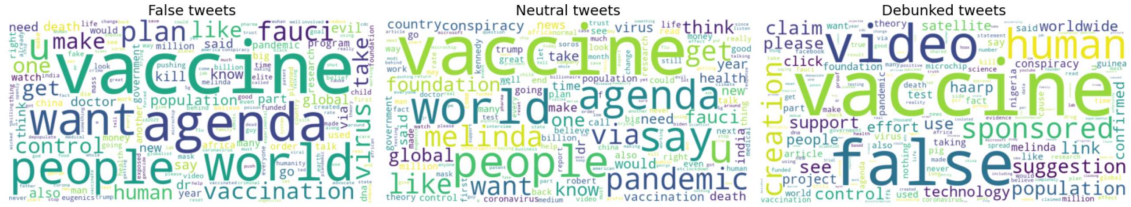
Bill Gates

For tweets in misinformation related to Bill Gates, since we used keywords "bill gates" and "depopulation" for downloading the tweets, most of the tweets had these keywords in the content, which is shown in Figure 6.3a. After removing these keywords, in Figure 6.3b, we can see that the word "vaccine" appeared with the highest frequency in all three groups because misinformation related to Bill Gates mostly talked about vaccines. Moreover, the keywords followed by some words in both false and neutral tweets were "agenda," "world," and "people." Additionally, "claim" and "false" appeared again in debunked tweets together with "support" and "sponsored" which also had a high frequency.

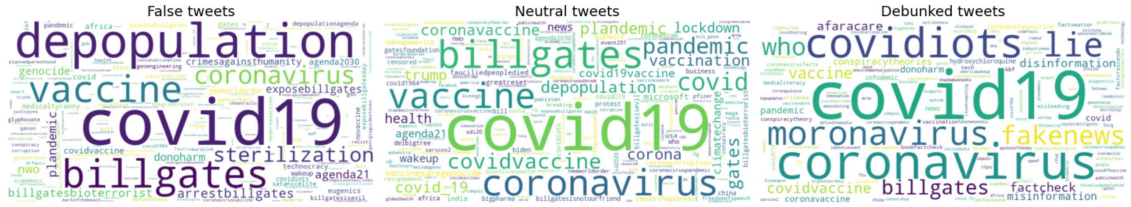
Regarding hashtags, Figure 6.3c visualized the hashtags used in tweets related to the misinformation of Bill Gates. After removing the hashtags #covid19, Figure 6.3d showed the frequency of hashtags used in tweets belonging to three categories. In particular, for false tweets, people tended to use long hashtags compared to other tweets, such as #arrestbillgates, #billgatesbioterrorist, #crimeagainsthumanity, and #exposebillgates. Meanwhile, for debunked tweets, it can be noticed that people used these two hashtags #covidiots and #moronavirus together with other hashtags like #fakenews, #factcheck, #disinformation, #misinformation, #WHO and a small portion of people using #hydroxychloroquine. To explain the meaning of #covidiots and #moronavirus, based on the VOA Learning English Website[4], the word "covidiot" which combines COVID-19 and idiot, indicates a person who ignores health and social distancing rules for preventing the spread of the virus. The word "moronavirus" has a similar meaning to covidiot; it combines the word "coronavirus" with the word "moron," another insulting word.



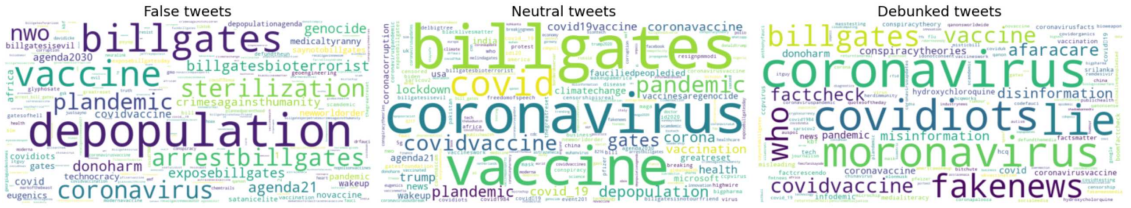
(a) Words used in tweets.



(b) Words used in tweets without “bill gates,” “billgates,” “depopulation,” and “covid”.

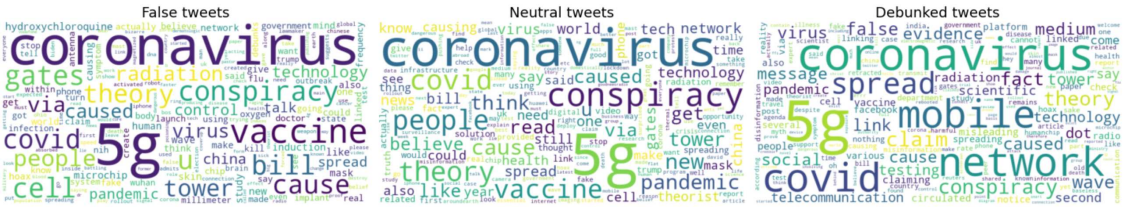


(c) Hashtags used in tweets.

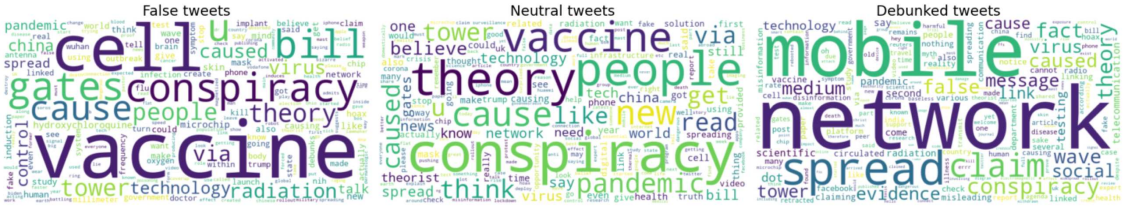


(d) Hashtags used in tweets without #covid19.

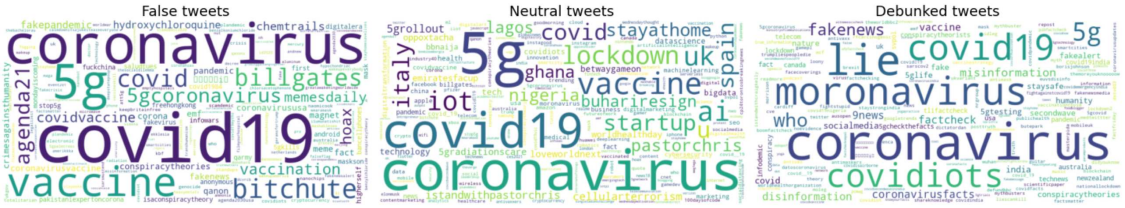
Figure 6.3: WordCloud plots of tweets related to Bill Gates.



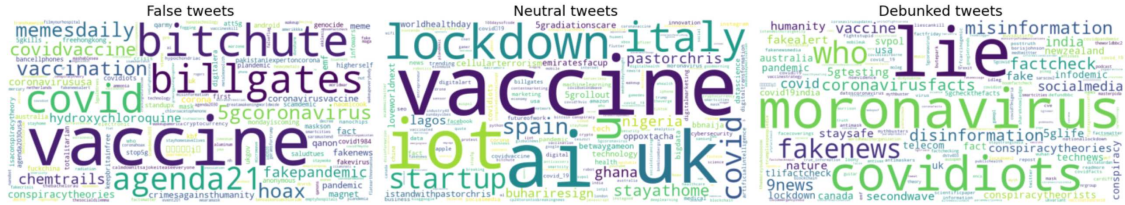
(a) Words used in tweets.



(b) Words used in tweets without “coronavirus,” “covid,” and “5g.”



(c) Hashtags used in tweets.



(d) Hashtags used in tweets without #covid19, #coronavirus and #5g.

Figure 6.4: WordCloud plots of tweets related to 5G.

5G

In 5G, as indicated above, this misinformation mostly talked about the connection between 5G and coronavirus as shown by the high frequency of these two words used in tweets. After removing these two words, in Figure 6.4b, we can see that some words like "cell," "vaccines," "bill gates", and "conspiracy" make up the majority of false tweets. While in debunked tweets, the two words "mobile" and "network" were used a lot together with "spread", "claim", "wave", "evidence", and "fact".

In terms of hashtags used, from the result shown in Figure 6.4c, #covid19, #coronavirus, and #5g were removed. Then, in Figure 6.4d, #vaccine is mostly used in false and neutral tweets. Furthermore, people tended to use #billgates, and #bitchute[5]. Moreover, they also used long hashtags such as #crimesagainsthumnanity, #fakepandemic, #5gcoronavirus, #isaconsipiracytheory. About neutral groups, people used hashtags like #lockdown, #italy, #AI, #UK, and #startup in their tweets. In debunked tweets, once again, they used #moronavirus and #covidiots together with other hashtags like #fakenews, #WHO, #factcheck, #misinformation, #coronavirusfacts, #disinformation.

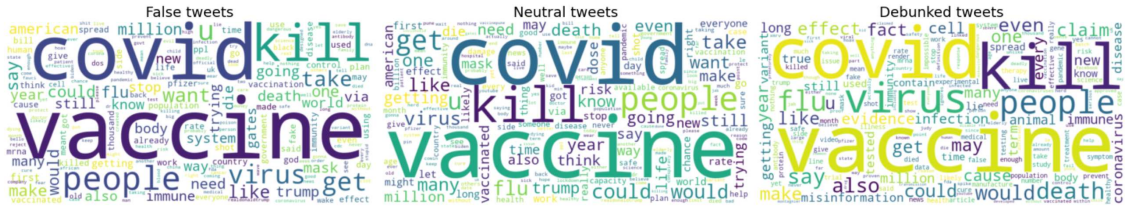
Vaccines

For the last topic of misinformation related to Vaccines, Figure 6.5a visualizes words used in tweets, and after removing keywords which are "covid," "vaccine," "kill," and "people" we have Figure 6.5b. In this Figure, the word "virus" appeared in both false and debunked tweets. Moreover, a large portion of verbs was used in false tweets such as "get," "take," "like," "want," and "say". While in the debunked group, it can be easily noticed that, in addition to the keywords indicated above, tweets also had words like "death," "flu," "say," "would," "fact," and "claim".

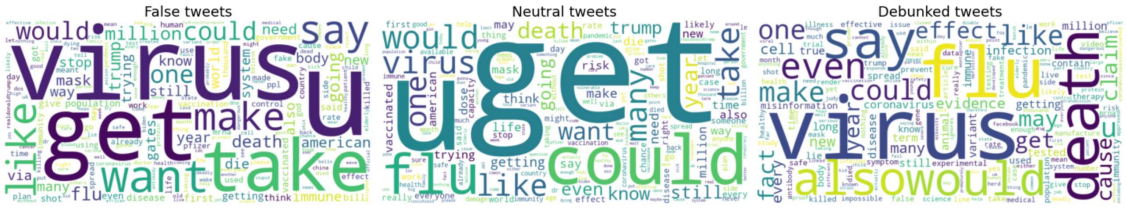
For the hashtags used in tweets, after eliminating some general hashtags like #covid19, #covid, and #coronavirus, Figure 6.5d showed some notable points between the three groups. In false tweets, we can easily notice the appearance of #billgates which shows the connection between the topic of Bill Gates with

[4]https://learningenglish.voanews.com/a/quarantini-moronavirus-covid-10-wordplay-brings-humor-to-these-times/5441194.html

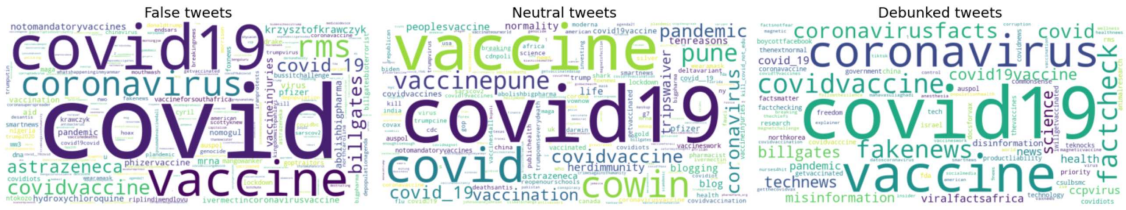
[5]BitChute is an alt-tech video hosting service launched by Ray Vahey in January 2017. It describes itself as offering freedom of speech, while the service is known for hosting far-right individuals, conspiracy theorists, and hate speech.



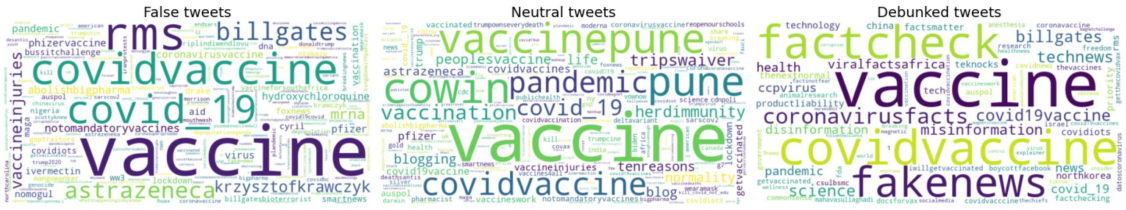
(a) Words used in tweets.



(b) Words used in tweets without “covid,” “vaccine,” “kill,” and “people.”



(c) Hashtags used in tweets.



(d) Hashtags used in tweets without #covid19, #coronavirus and #covid.

Figure 6.5: WordCloud plots of tweets related to Vaccines.

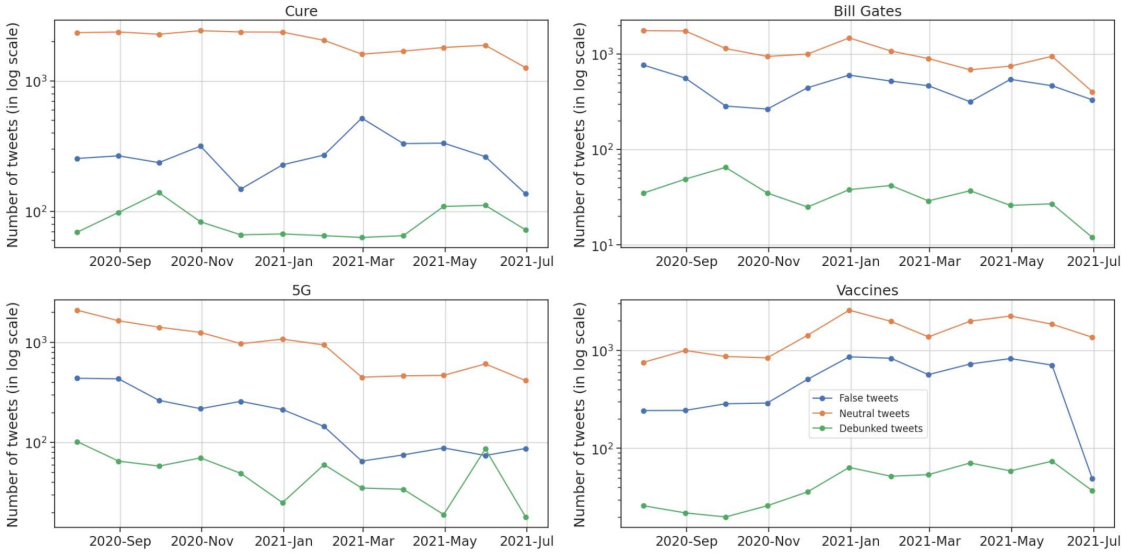


Figure 6.6: Number of tweets per month (in log scale).

this topic. Moreover, long hashtags such as #notomandatoryvaccines, #vaccineinjuries, #astrazeneca, #krzysztofkrawczyk, #abolishbigpharma, and #vaccineforsouthafrica were also seen in this group. Whereas, for tweets in the debunked group, they mostly used some hashtags like #factcheck, #fakenews, #coronavirusfacts, #misinfomation. In neutral tweets, some hashtags were used related to India like #vaccinepune, #cowin, #pune, and some other general hashtags related to COVID-19 like #covidvaccine, #pandemic, #vaccination.

6.3 Temporal Behavior

In this section, tweets were analyzed as a time-ordered sequence of observations. Particularly, some metrics are extracted to evaluate: the number of tweets, retweet count, like count, and reply count. Since the number of neutral tweets was significantly higher than false and debunked tweets which made the graph skewed, the Logarithmic scale (log scale) was used for all of the graphs in this Section to deal with this problem.

Number of tweets

The graph 6.6 showed the number of tweets changes over the period in all four misinformation macro topics. This graph has three lines which were colored orange, blue, and green to indicate neutral, false, and debunked tweets, respectively. Because the number of neutral tweets was higher than the other two groups, the orange lines in four small graphs in Figure 6.6 were always above the blue and green lines, followed by the blue lines for false tweets. In general, there was a contrast between false tweets and debunked tweets, as once the number of debunked tweets increased, the number of false tweets decreased, respectively, and vice versa. Particularly in the macro topic of Cure, at the beginning of the period, in July 2020, the number of false tweets was higher than the number of debunked tweets, but then it slightly decreased while debunked tweets peaked at the same period in October 2020. Meanwhile, the number of neutral tweets mostly remained the same. Subsequently, both false and debunked groups declined to one of the lowest points at the end of 2020 and then started going up to reach a peak in March 2021 for the number of false tweets, while the number of debunked tweets dropped to the lowest point in the same period. At the end of the concerned time, the number of tweets in all three categories went down.

For tweets in misinformation related to Bill Gates, it can be seen in Figure 6.6 that, the number of tweets in three groups tended to decrease after one year. While the number of false tweets fluctuated more than the other two groups, the number of neutral tweets decreased and the number of debunked tweets changed insignificantly compared to the other two groups throughout the one-year period. In October 2020, the number of debunked tweets reached a peak, whereas the number of false tweets was the second-lowest score in the same period.

Similarly, for the number of tweets in misinformation related to "5G", the number of tweets in three categories tends to decline. As seen in Figure 6.6, we can easily notice the contrast behavior between the false and debunked groups. Starting at a peak in July 2020, the number of false tweets reached its lowest point in March 2021 and slightly decreased again until the end of the period. Meanwhile, at the starting point, having the lowest number of tweets among the three categories, the debunked group also dropped to its lowest point in May 2021 and suddenly jumped to the peak again after 11 months in June 2021. Then, it sharply collapsed at the

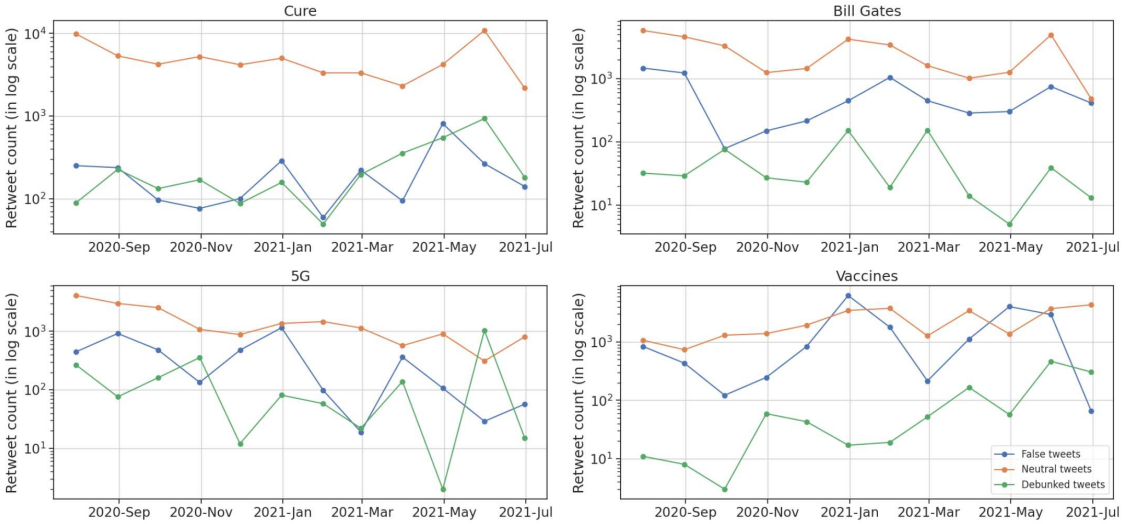


Figure 6.7: Retweet per month (in log scale).

end of the period.

In contrast, the number of tweets with misinformation related to vaccines significantly increased over the first half period from July 2020 to January 2021. Then, it was likely to be stable until Jun 2021 and suddenly dropped to its lowest point at the end of the period. Whereas, the number of debunked and neutral tweets tended to increase after a year even though the change was not so significant.

Retweet count

In this part, we extracted the number of times a tweet was retweeted and visualized the change over time. In general, there were some similarities with the behavior of the number of tweets, which is shown in Figure 6.7. Particularly on the topic of Cure, the number of retweet counts in neutral tweets was significantly higher than in the other two groups. Moreover, the number of retweets in the false group slightly decreased while the debunked groups tended to increase at the end of the period. The number of retweet counts for false tweets peaked in May 2021, and then the number of retweet counts for debunked tweets reached its highest point after one month, in June 2021.

In the topic of Bill Gates, it can be easily noticed that starting with the highest amount of retweet count throughout the period, the false group suddenly decreased to its lowest point after three months and then started increasing slightly again. By

contrast, despite the retweet count in the debunked group being the lowest among the three groups, it had the most variation. Particularly, the number of retweet counts in the debunked group peaked twice in January and March 2021. Moreover, there was a sudden decline in February 2021 for the debunked group, whereas the retweet count for the false group tended to increase this month. Throughout the remaining period, the number of retweet counts in the three groups had a tendency to decrease.

Next, in the topic of 5G, from Figures 6.7, we can see that the lines for false tweets and debunked tweets are opposing zigzags. Notably, in May 2021, the retweet count for debunked tweets reached its lowest point and only after one month, it had the highest number among three groups meanwhile, the retweet count for false tweets reached its second lowest point in the same period. Furthermore, at the end of the period, the debunked group suddenly dropped while the other two groups slightly increased.

Finally, for the topic of misinformation related to Vaccines, based on Figure 6.7, the number of retweet counts for neutral and debunked groups increased while the one for the false group decreased after a year. Particularly, the number of retweet counts for false tweets peaked in January 2021 and then decreased to its lowest point at the end of the period. Meanwhile, the debunked group fluctuated considerably over time but we can still notice the opposites between false and debunked groups.

Like count

In this part, we extracted the number of likes a tweet had and visualized the change over time. In Figure 6.8, we can notice some similarities between like count and retweet count plots, which showed that somehow a person who liked a tweet tended to retweet that tweet also. Moreover, since the number of neutral tweets accounted for the most, the orange lines were almost above the lines of the other two groups throughout the period except at some points, and there were also some fluctuations in the performance of tweets in all three groups. In the first topic related to Cure, throughout the period, the difference in the behavior between false and debunked tweets was not so significant. Notably, starting from March 2021, the like count for debunked tweets increased and peaked in June 2021, while during the same period, false tweets peaked one month before and started falling after that. Meanwhile, in

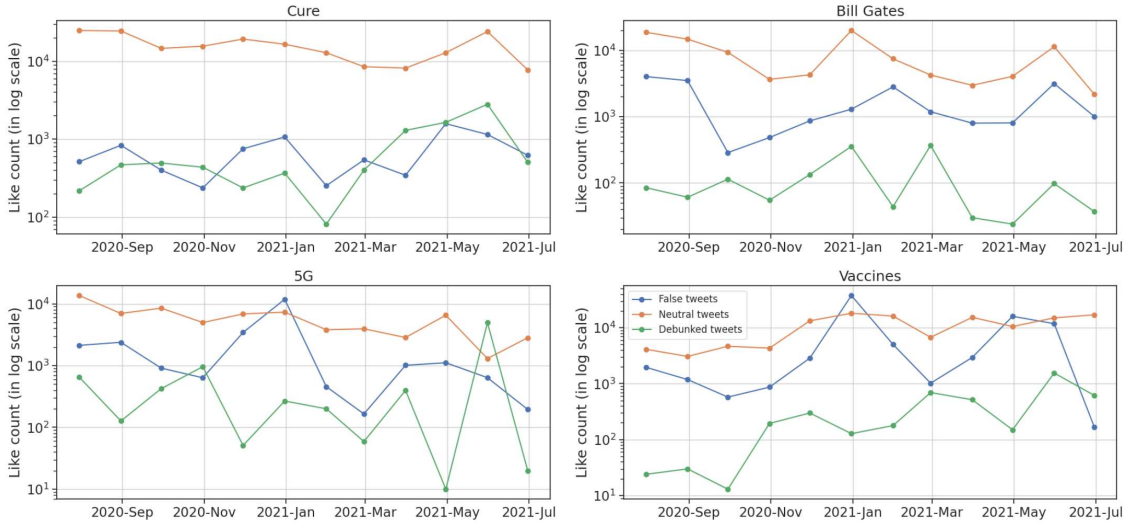


Figure 6.8: Like count per month (in log scale).

the topic of Bill Gates, the contrasting behavior between false and debunked tweets is easily seen. If the number of like count for false tweets tends to increase, the one for debunked tweets tends to decrease and vice versa. Similar trends were seen for the topics of 5G and Vaccines.

Reply count

In this part, the number of reply counts is calculated for each topic and visualized in Figure 6.9. Although the number of reply count in neutral tweets was always superior to the ones for false and debunked tweets, there was a highlight point in reply count for the topic of 5G where at the beginning of the period, the number of reply counts in debunked tweets reached its peak and this number was much higher compared to the general number of reply counts in four topics. Moreover, we also can notice the zigzagging behavior between the three groups. On the topic of misinformation related to Cure, starting at the peak throughout the period, the number of reply counts in false tweets declined until October 2020 and then inclined again in January 2021 and the same pattern appeared in the second half of the cycle. Whereas in debunked tweets, during the first half of the cycle, they did not change much, until February 2021, it went up and reached their highest point in June 2021 before dramatically dropping at the end of the period. In terms of the number of reply counts on the topic of Bill Gates, opposing behavior between the two groups

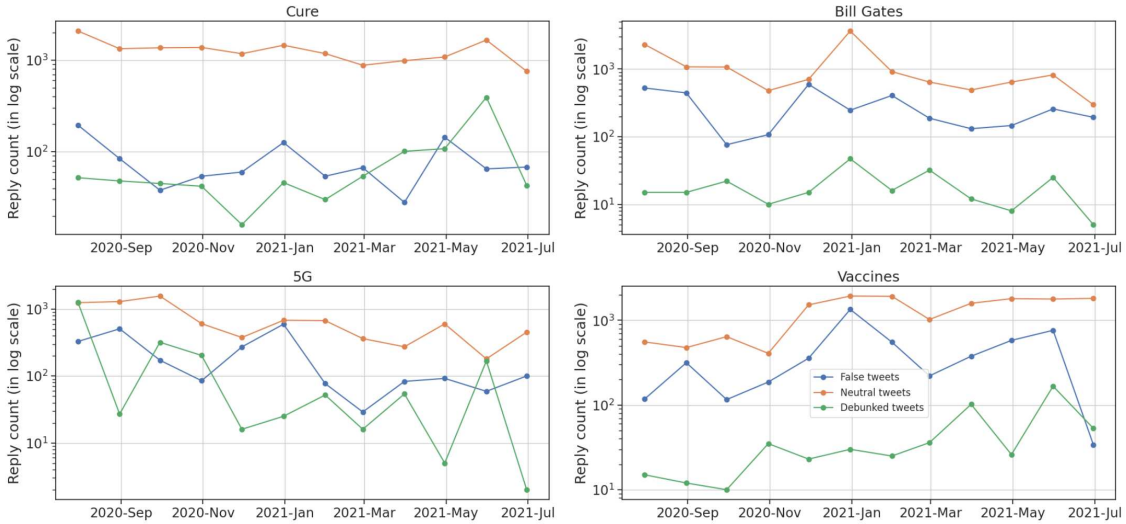


Figure 6.9: Reply count per month (in log scale).

false and debunked was observed, if the number in the false group tended to rise in the next period, the number in the debunked group tended to fall in the same period and vice versa. For the topic of misinformation related to 5G, in the first 6 months, the contrast behavior between the false and debunked groups can be easily seen but for the remaining period, the changes were not so significant. For the last dataset, in general, the number of reply counts in false tweets tended to increase and reached a peak in the middle of the period and then fell whereas the number of reply counts for debunked tweets increase after one year.

Combination

Since retweet count, like count, and reply count had quite similar behavior in general, we decided to combine these numbers into one graph which was visualized in Figure 6.10. At first glance, this graph looked exactly like Figure 6.8 for the number of like counts but indeed there were some differences though very small. In general, false tweets had more interactions with other people compared to debunked tweets, especially on the topics of misinformation related to Bill Gates and Vaccines. Furthermore, since Figure 6.10 mainly maintained the structure of Figure 6.8 of like count behavior, the number of likes was comparatively more than the retweet count and reply count.

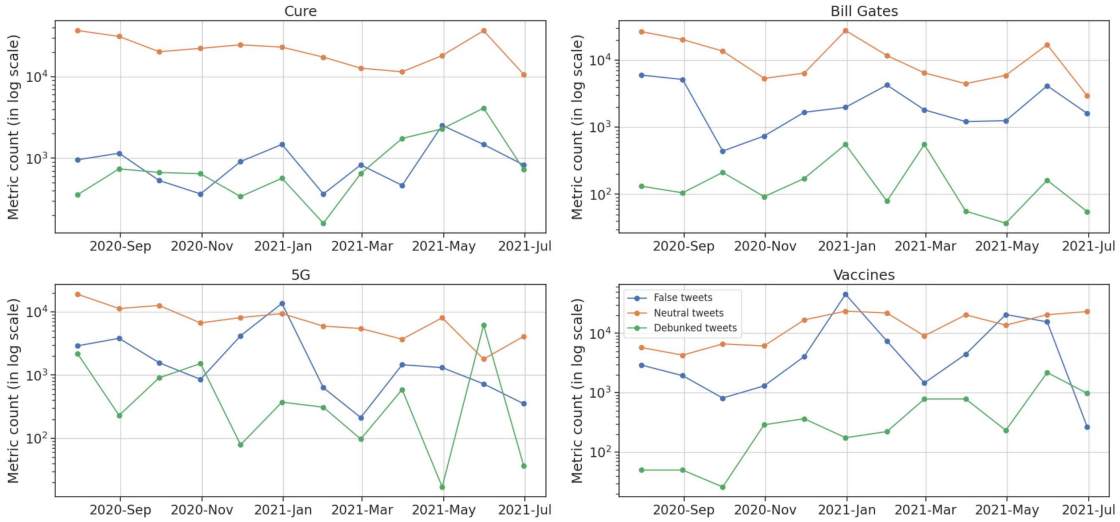


Figure 6.10: Combination of retweet count, like count, and reply count (in log scale).

6.4 Topic Visualization: The Documents

In this part, we used a function created by Selen Arslan[6] to visualize the documents inside the topics to see if they were assigned correctly or whether they made sense. Generally, this function recalculates the document embeddings and reduces them to 2-dimensional space for easier visualization purposes. Moreover, the color of each point represents the class to which the tweet belongs.

Cure

For the first topic of misinformation related to Cure, Figure 6.11 showed the documents and topics with the colors showing the labels of the tweets. Yellow indicates neutral tweets while red and blue indicate false and debunked ones respectively. Since there were 19 small topics belonging to this misinformation based on BERTopic results and some of them were not so meaningful for further analysis due to their irrelevant or generality, we decided to extract only five topics to analyze based on the subjective feelings about the specialness of these topics. Here, we chose five topics to indicate further which are topics 1, 6, 7, 9, and 16.

[6]selen.arslan@studenti.unipd.it

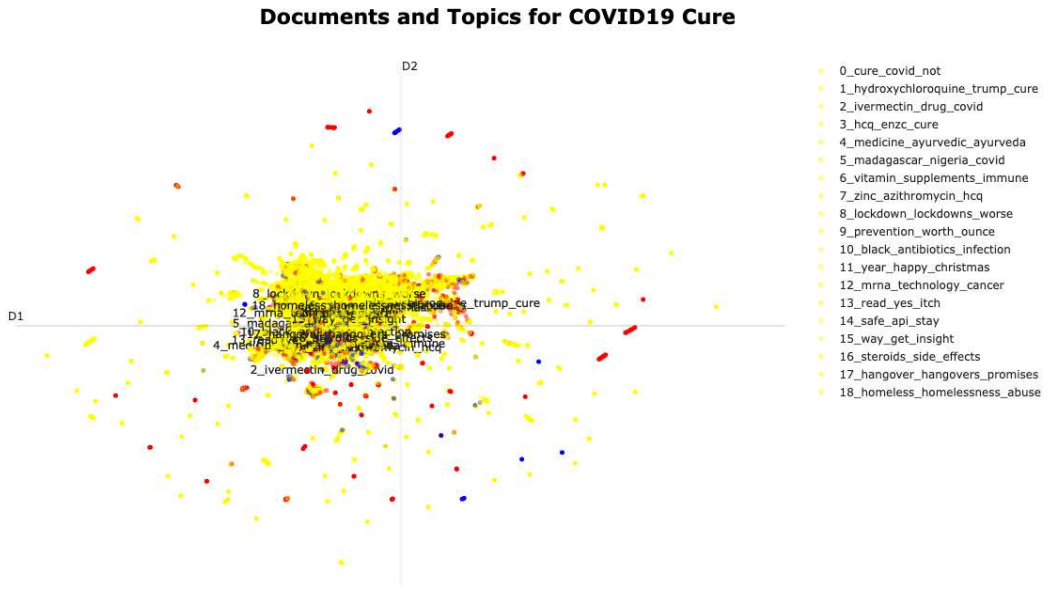
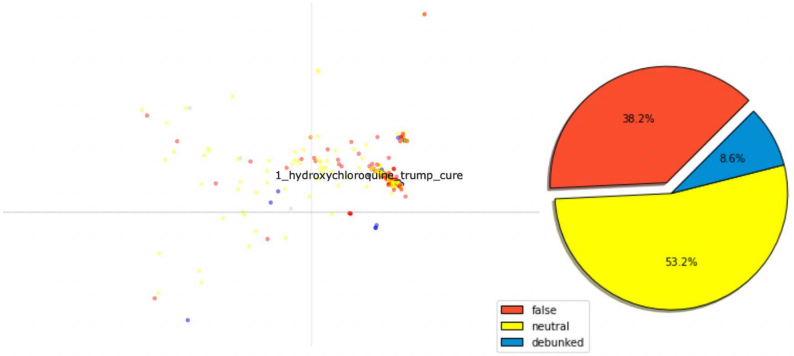


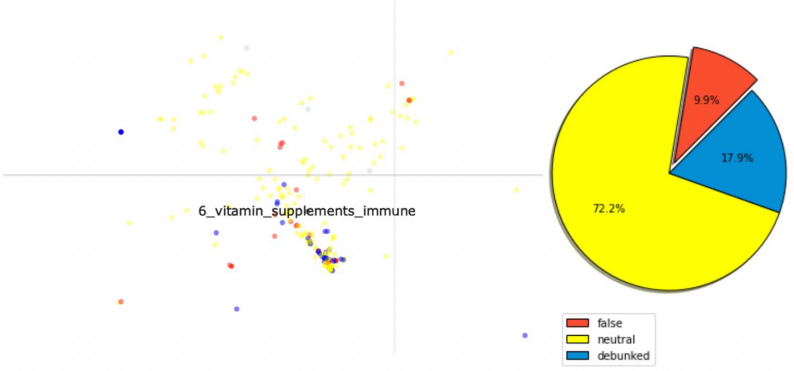
Figure 6.11: Topic Documents Visualization for Cure.

The top three words that appeared in topic 1 were "hydroxychloroquine," "trump," and "cure." As shown in Figure 6.12a, this topic consisted of mostly neutral tweets, at 53.2%, followed by false tweets and debunked tweets at 38.2% and 8.6% respectively. A study [1] showed that from March 1 to April 30, 2020, Donald Trump made 11 tweets about unproven therapies and mentioned these therapies 65 times in White House briefings, especially touting hydroxychloroquine and chloroquine. Moreover, their results also revealed that there was a substantial increase in purchases and searches for previously unpurchased and unsearched therapies by the general public following the backing of former US President Donald Trump. Therefore, a large portion of false tweets related to this topic was created, followed by debunked tweets since treatment with chloroquine or hydroxychloroquine, with or without a macrolide, appears to increase the risk of death in patients with COVID-19 [48].

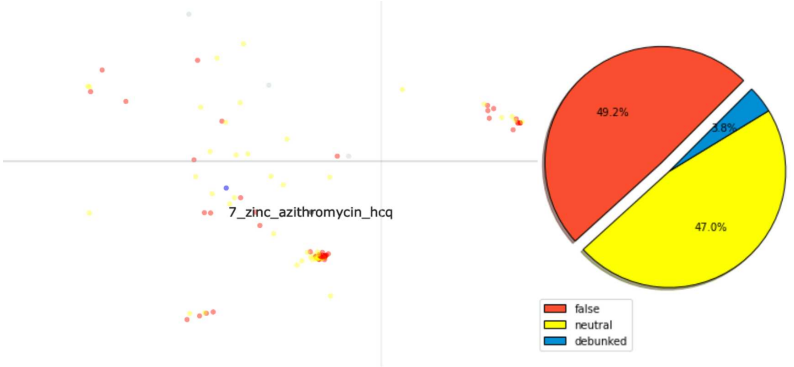
Topic 6, in which people discuss things related to "vitamin," "supplements," and "immune", had the highest percentage of debunked tweets - 17.9% among other topics. This is because people were encouraged to strengthen their immune systems



(a) Topic 1 - hydroxychloroquine_trump_cure.

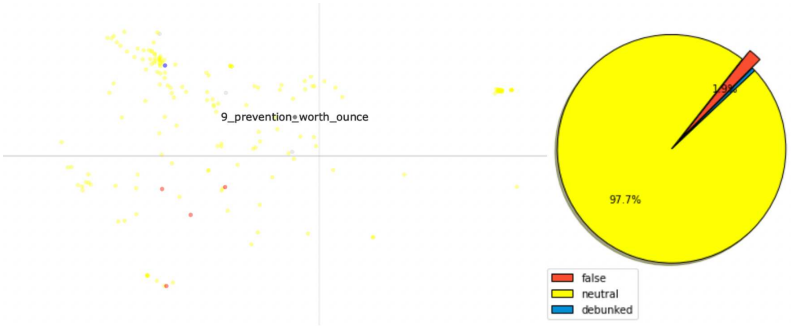


(b) Topic 6 - vitamin_supplements_immune.

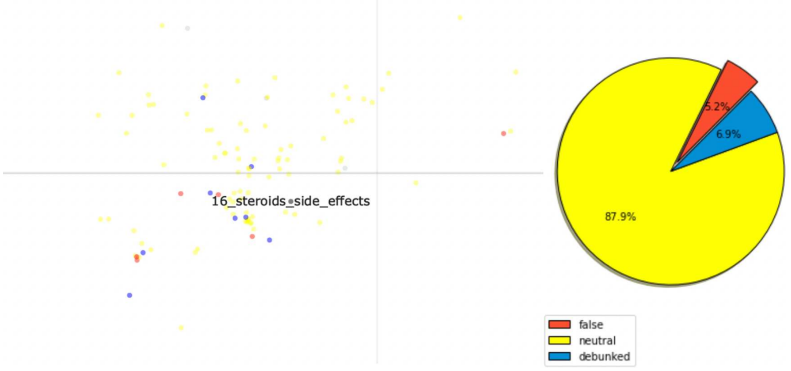


(c) Topic 7 - zinc_azithromycin_hcq.

Figure 6.12: Topic of tweets related to Cure.



(d) Topic 9 - prevention_worth_ounce.



(e) Topic 16 - steroids_side_effects.

Figure 6.12: Topic of tweets related to Cure (cont.).

by taking some vitamins and/or supplements rather than believing in misinformation on the Internet about the benefits of any kind of medicine in terms of curing COVID-19 since there is no cure for COVID-19 until now. Moreover, a very large portion of tweets on this topic was neutral, at 72.2%, and false tweets accounted for only 9.9%.

In topic 7, instead, people mostly talked about "zinc," "azithromycin," and "hcq" (shortly for hydroxychloroquine). Surprisingly, this topic had a high proportion of false tweets, at 49.2%, followed by a slightly lower number of neutral tweets, at 47%, and only 3.8% of debunked tweets. Given that there were many rumors about some nutritional supplements like zinc or vitamin D and some antibiotics like azithromycin that can be used against COVID-19, therefore, this topic has a high number of false tweets.

Topic 9 mostly discussed "prevention" which can be seen in the top three words used in Figure 6.12d, so that over 97% of tweets on this topic were neutral, and a very small portion of tweets belong to false and debunked, approximately 2.3% in total.

The final topic chosen here is topic 16, as seen in Figure 6.12e, the top three words used were "steroids," "side," and "effects." Despite the neutral group accounting for the highest percentage of tweets, at 87.6%, this topic also had many debunked tweets with a portion slightly higher than false tweets. Since the use of steroids for COVID-19 patients must be under the supervision and approval of a doctor rather than self-administered, many tweets on this topic were debunked tweets with the aim to provide some information about the side effects of using the medicine without any permission of the doctors.

Bill Gates

In this topic of misinformation, after performing BERTopic, we visualized all tweets with colors representing the classes of tweets, which was shown in Figure 6.13. It can be noticed that most of the tweets on this topic were visualized in the middle left of the graph based on the density of points in the Figure. Moreover, false tweets tended to have higher coordinates compared to neutral tweets since red points were more concentrated above the area with a dense density of points. Since the number of debunked tweets was small, we cannot see clearly the distribution of this group

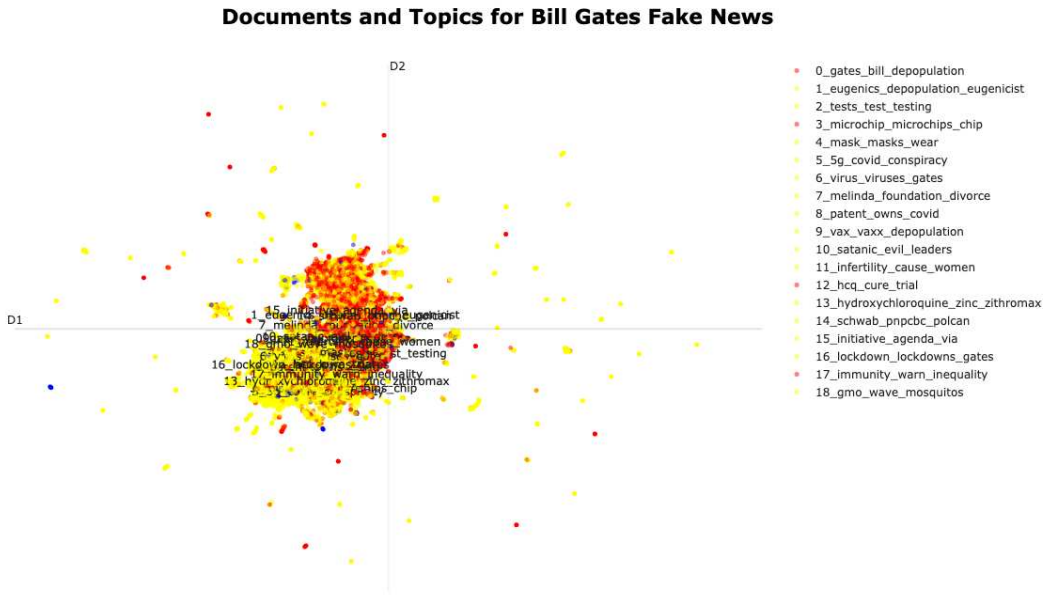
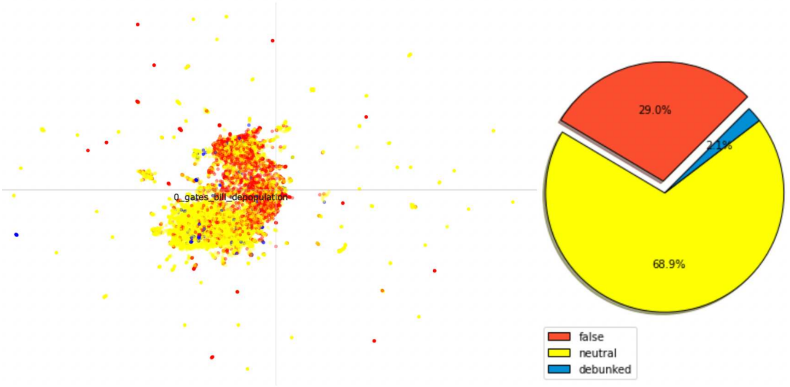


Figure 6.13: Topic Documents Visualization for Bill Gates.

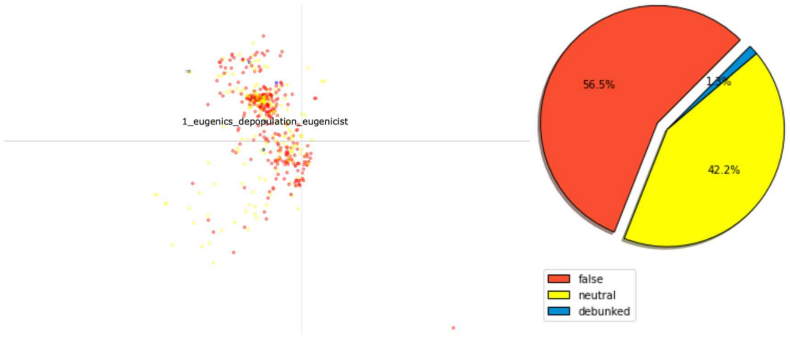
on the graph. In terms of the topics, since removing outliers, there were 19 small topics found when people discussed misinformation related to Bill Gates. Generally, we can see some topics such as topics 3 and 5 were directly related to 5G, topics 12 and 13 were related to Cure, and topic 9 was about Vaccines. Particularly, we chose five topics among these 19 topics which were different from the other five topics indicated above to see how tweets in the three groups behaved.

The first topic we chose to analyze is topic 0, as seen in Figure 6.14a, this topic contained most of the tweets that belong to the misinformation related to Bill Gates in which the top three words were “gates,” “bill,” and “depopulation”. From this Figure, we can notice that it maintained mostly the structure of the main graph 6.13 since this topic had the highest number of tweets. The pie chart showed that most tweets on this topic were neutral, with approximately 68.9%, whereas, false tweets accounted for 29% and a small portion for debunked tweets, only 2.1%.

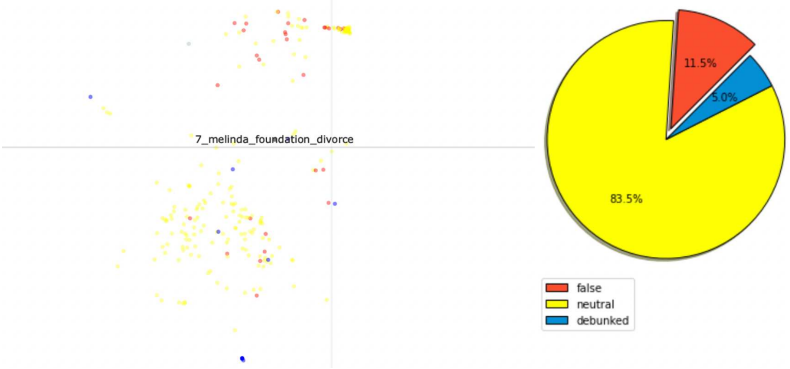
In topic 1, people discussed “eugenics,” “depopulation,” and “eugenicist” which is visualized in Figure 6.14b. This topic has mostly false tweets with approximately 56.5%, followed by neutral tweets at 42.2% and debunked tweets accounting for



(a) Topic 0 - gates_bill_depopulation.

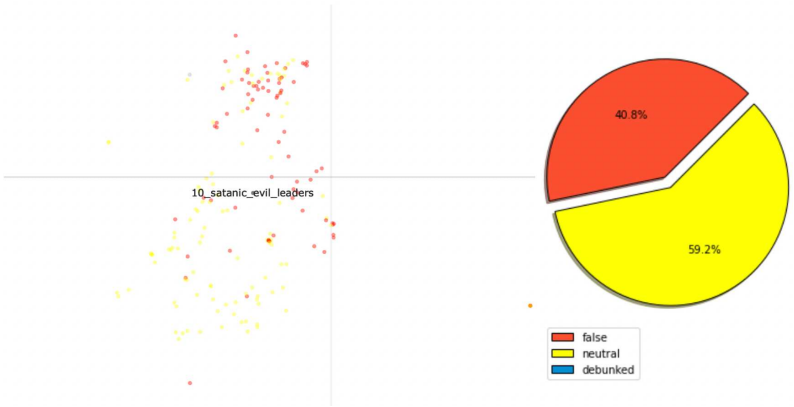


(b) Topic 1 - eugenics_depopulation_eugenicist.

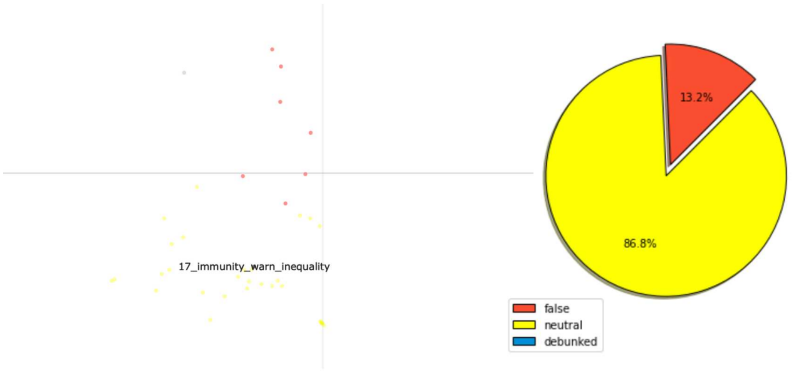


(c) Topic 7 - melinda_foundation_divorce.

Figure 6.14: Topic of tweets related to Bill Gates.



(d) Topic 10 - satanic_evil_leaders.



(e) Topic 17 - immunity_warn_inequality.

Figure 6.14: Topic of tweets related to Bill Gates (cont.).

1.3%. As seen by the top words in the graph, the word "Eugenics" basically refers to the scientifically erroneous and immoral theory of "racial improvement" and "planned breeding," which was directly related to misinformation about Bill Gates, therefore, in this topic, false tweets accounted for the most.

Topic 7 discussed the divorce of Bill Gates and his wife, Melinda Gates, which can be seen from the graph 6.14c. Not surprisingly, the highest portion of tweets on this topic belongs to the neutral group, with approximately 83.5%, followed by false tweets, at 11.5%. From the graph, we can see that almost neutral points with yellow color concentrated on the left below the horizontal line of the graph while red points represent the false group focused more on the left but above the horizontal line whereas debunked tweets distributed scattered.

In topic 10, from Figure 6.14d, we can notice that there were no debunked tweets on this topic. The top three words of this topic were "satanic," "evil," and "leaders," with approximately 59.2% of neutral tweets, and the rest are false tweets.

The last topic we chose here is topic 17, in which the top three words were "immunity," "warn," and "inequality." Once again, this topic did not have any debunked tweets with the percentage of neutral tweets being 86.8%. Since the top keywords did not directly relate to any misinformation.

5G

For the topic of misinformation related to 5G, from Figure 6.15, we can notice that most of the small topics of BERTopic result concentrated on the left of the graph along the horizontal line except the topic 6 which was isolated at the bottom left of the graph. Moreover, the colors representing the labels were scattered. While false tweets were distributed around the origin of the graph based on the dense density of the red points, the debunked tweets were concentrated more on the left above line D1 and on the right below line D1. Meanwhile, neutral tweets with yellow color were scattered.

Specifically, in topic 0, the majority of tweets are neutral which accounted for 79.2%, followed by false and debunked tweets, at 16.4% and 4.4% respectively. Since the top three words that appeared in tweets are "5G," "coronavirus," and "conspiracy," this topic mostly discusses some general things around the macro topic of 5G.

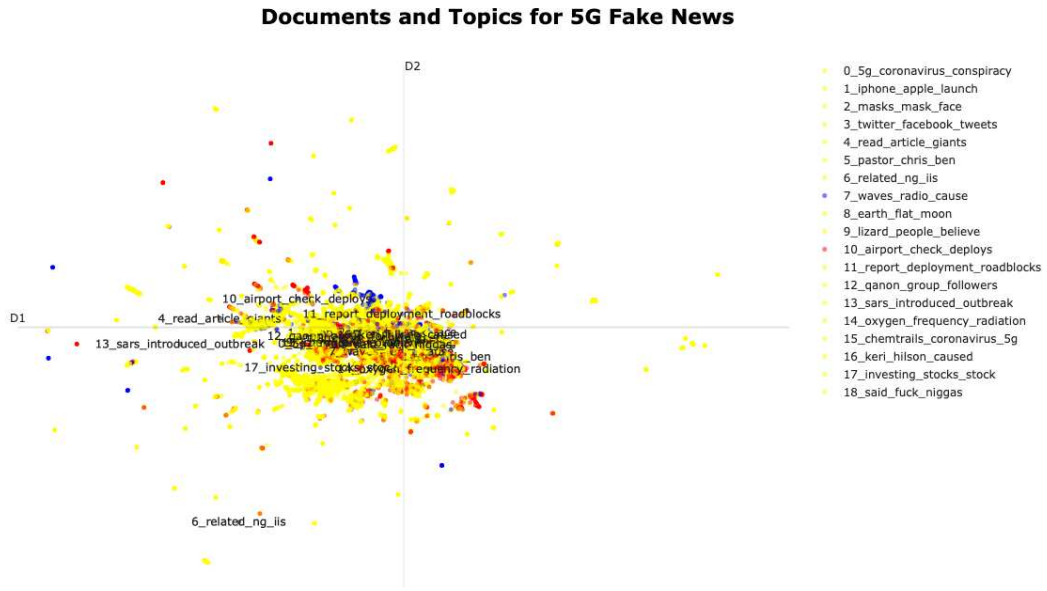
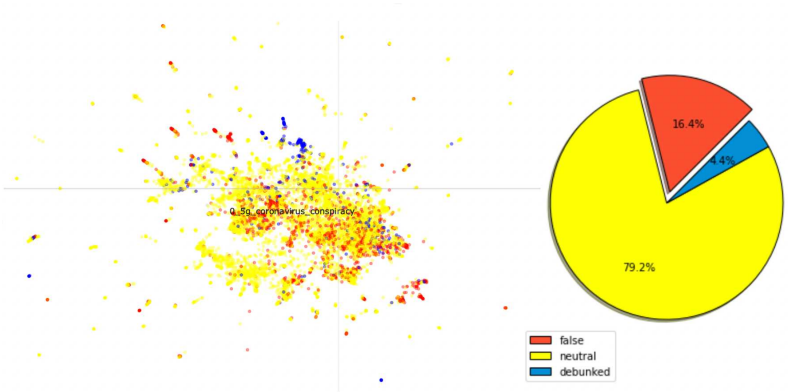


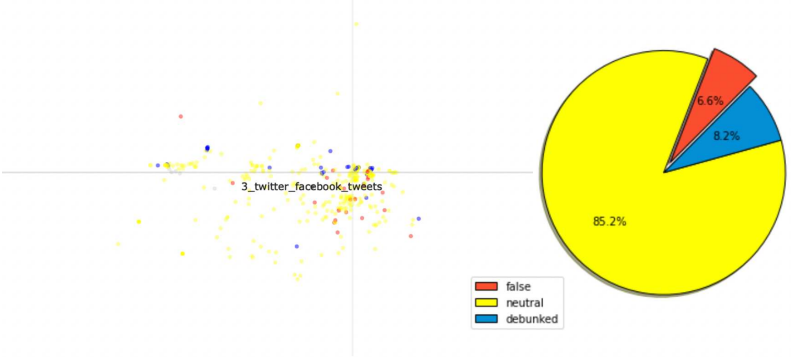
Figure 6.15: Topic Documents Visualization for 5G.

Next, BERTopic classified a portion of tweets in the topic of 5G into a smaller topic in which the top words that appeared were "twitter," "facebook," and "tweets". Clearly, in this subtopic, people discussed things related to social media and especially Twitter. Therefore, neutral tweets accounted for the most among the three groups, at 85.2%. Notably, from Figure 6.16b, the number of debunked tweets on this topic was slightly higher than the number of false tweets. Moreover, debunked tweets were frequently distributed in or above the horizontal line while red points that visualized false tweets tended to stay lower than the horizontal line. The reason is that generally, misinformation that spreads on social media and especially on Twitter started as a post, and for debunking, the author usually indicates the misinformation by its origin, for example, "A tweet claimed that …," and then debunks that misinformation by using some proper evidence. Consequently, in this specific topic, we had a higher number of debunked tweets rather than false tweets.

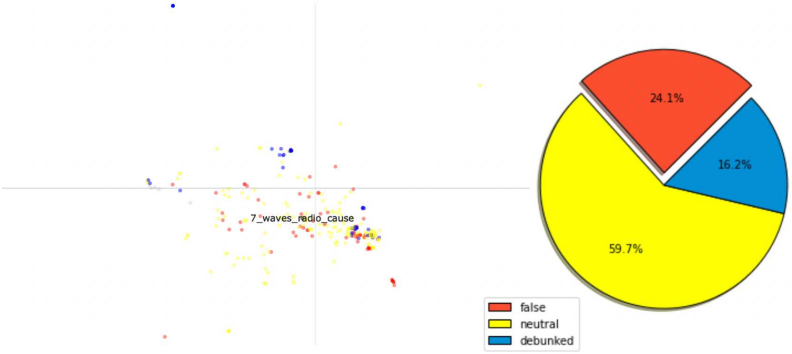
Figure 6.16c visualized topic 7 which belongs to the misinformation of 5G. This topic is directly related to 5G which is shown through the top three words "waves," "radio," and "cause." Unlike the previous two topics, this topic had a lower num-



(a) Topic 0 - 5g_coronavirus_conspiracy.

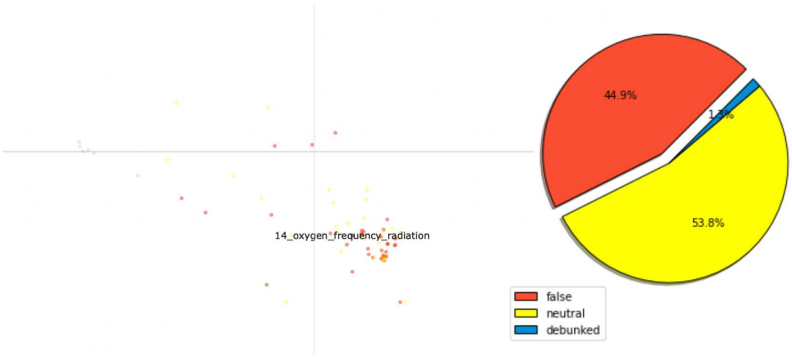


(b) Topic 3 - twitter_facebook_tweets.

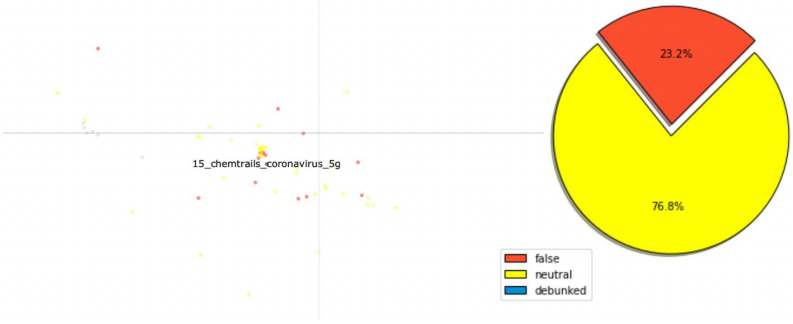


(c) Topic 7 - waves_radio_cause.

Figure 6.16: Topic of tweets related to 5G.



(d) Topic 14 - oxygen_frequency_radiation.



(e) Topic 15 - chemtrails_coronavirus_5g.

Figure 6.16: Topic of tweets related to 5G (cont.).

ber of neutral tweets and a higher number of false and debunked tweets. While debunked tweets concentrated both on the left above the horizontal line and on the right below the horizontal line, most false tweets were on the left and below the horizontal line. Particularly, many rumors talked about the connection between 5G and COVID-19, they said that the virus can travel on radio waves and/or mobile networks so that the 5G can spread COVID-19. For that reason, this topic has a very high portion of false and debunked tweets which discussed the connection between 5G and COVID-19.

Given that COVID-19 is caused by the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), which targets the human respiratory system and can induce multiple organ failure. Moreover, respiratory damage can give rise to a plethora of health issues in an infected patient, including silent hypoxia. Silent hypoxia is defined as a condition where an individual has an alarmingly lower oxygen saturation level than anticipated, however, the individual does not experience any breathing difficulty [49]. Moreover, there are many studies showing the effect of wireless radiation on the immune system. They showed that radiofrequency exposure affects the structure of hemoglobin, reducing its ability to bind to oxygen. After just two hours of exposure to cell phone radiation, human hemoglobin structure changed, decreasing its affinity to bind to oxygen in the lungs between 11-12% which reduces the amount of oxygen that would be carried from the lungs to the body's tissues, contributing to hypoxia [50]. Consequently, in the topic visualized in Figure 6.16d, where people discussed "oxygen," and "radiation," a very large number of tweets were false, accounting for the highest percentage of false tweets in a topic among 19 topics, at 44.9%. Since they tried to connect the information above and persuaded people to believe that 5G mobile technology is the cause of COVID-19, even though until now there is no scientific evidence showing that connection. However, in this topic, only 1.3% of tweets are debunked ones.

The final topic shown here is topic 15. From Figure 6.16e, we can notice that this topic did not have any debunked tweets. Looking at the top three words shown on the graph, this topic mostly talks about "chemtrails," "coronavirus," and "5g." Chemtrails is a visible trail left in the sky by an aircraft and believed by some to consist of chemical or biological agents released as part of a covert operation, rather than the condensed water of a vapor trail [7]. There were various rumors related to

[7]Oxford Languages Dictionary

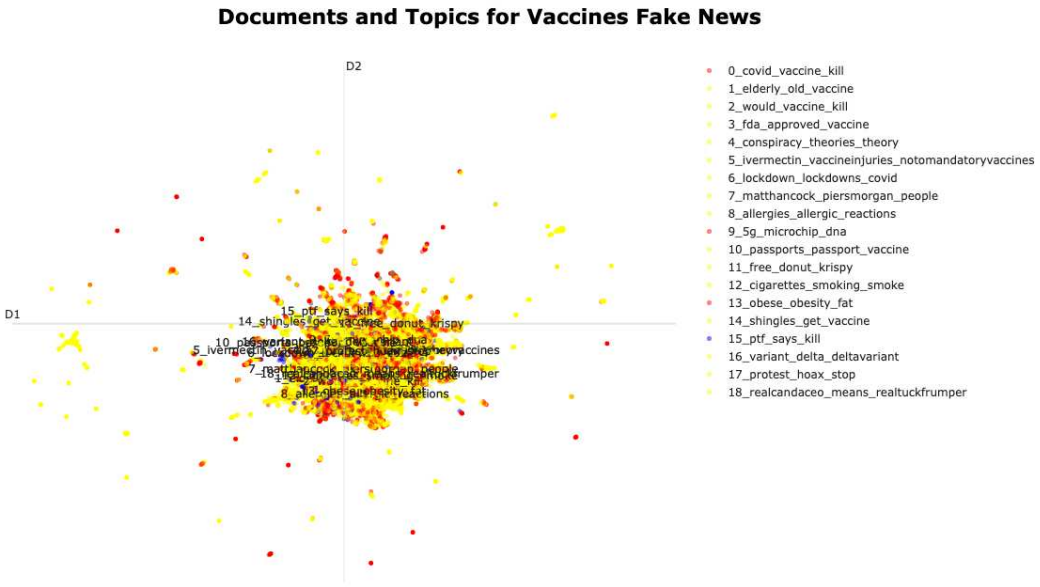


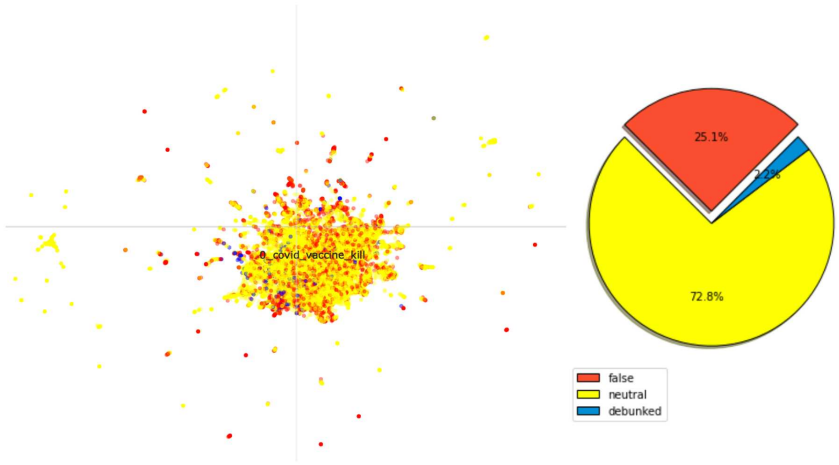
Figure 6.17: Topic Documents Visualization for Vaccines.

chemtrails, some of them are, for example, planes spray coronavirus into the air, and people got infected with the virus by inhaling the content of "chemtrails," moreover, a decade-old conspiracy theory alleging that the Australian government had approved the use of "chemtrails" to vaccinate the population forcibly has resurfaced with a fresh COVID-19 twist. Until now, there is no evidence for the existence of chemtrails [51], so tweets belonging to this topic often classify as false along with a large portion of neutral tweets, accounting for 76.8%.

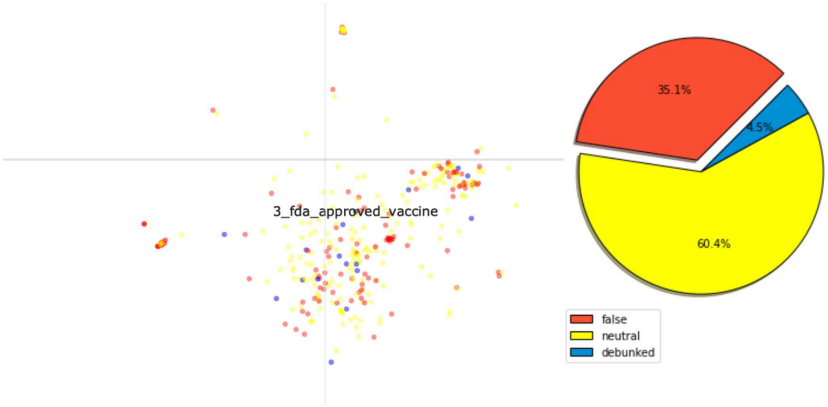
Vaccines

For the final topic, Figure 6.17 visualized the documents and topics for misinformation related to Vaccines. In general, neutral, false, and debunked tweets were scattered around the origin of the graph.

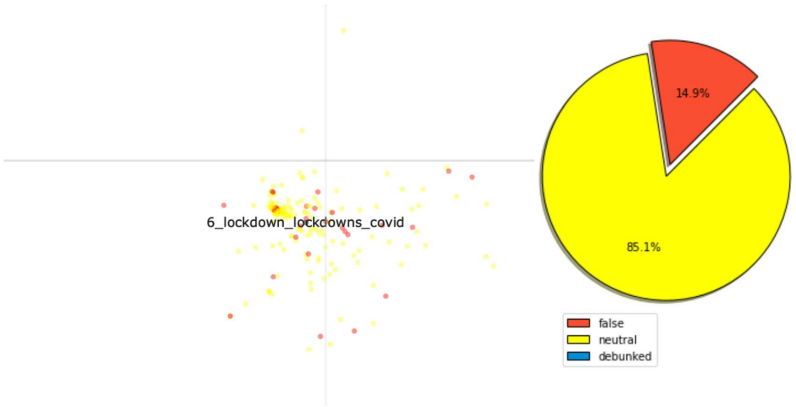
Particularly, considering the first small topic from the BERTopic algorithm, this topic mostly discussed vaccines related to COVID-19 and targeted on misinformation community by having the word "kill." The pie chart in Figure 6.18a shows that neutral tweets are the majority with approximately 72.8% in the total, while false



(a) Topic 0 - covid_vaccine_kill.

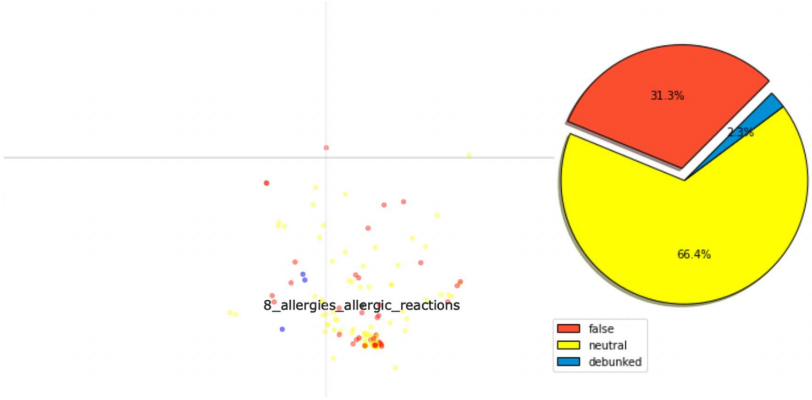


(b) Topic 3 - fda_approved_vaccine.

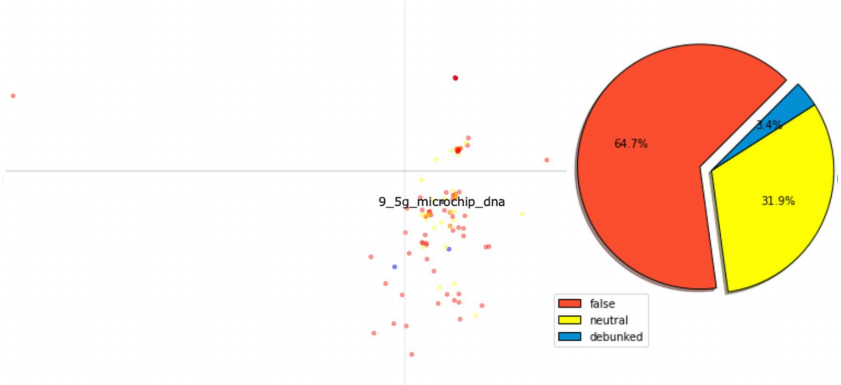


(c) Topic 6 - lockdown_lockdowns_covid.

Figure 6.18: Topic of tweets related to Vaccines.



(d) Topic 8 - allergies_allergic_reactions.



(e) Topic 9 - 5g_microchip_dna.

Figure 6.18: Topic of tweets related to Vaccines (cont.).

tweets accounted for 25.1% and a very small portion for debunked tweets, at 2.2%.

In the next topic, as we can see from Figure 6.18b, this topic is represented by the top three words "FDA," "approved," and "vaccine." Therefore, they likely talked about the approval of vaccines by the FDA - the Food and Drug Administration. Despite neutral tweets were still the majority with approximately 60.4%, false tweets played also an important role with 35.1% and only 4.5% of debunked tweets. Since that moment, the year of 2020 and early 2021, COVID-19 vaccines attracted public attention a lot. Consequently, taking this advantage, a part of people spread misinformation, confusing the public that the approved vaccines had problems.

Topic 6, in which people discussed the lockdown of COVID-19, did not have debunked tweets, and neutral tweets were the majority with 85.1%. This is because this topic was very generalized, mostly talking about the pandemic and lockdown, without considering any aspect of misinformation.

Next, we consider topic 8, which was directly related to "allergy," and "vaccines reactions." Similar to the behavior of topic 3, this topic also had a large portion of neutral tweets, at 66.4%, while false tweets accounted for 31.3% and only 2.3% of debunked tweets. Given that the authors of false tweets tended to say that they or the people they know have been suffering from the allergy after taking the vaccines which made the information more reliable. Therefore, not only false tweets but also debunked tweets talked about this topic. Furthermore, from the graph, it can be seen that the majority of false and neutral tweets were concentrated on the right below the horizontal line while debunked tweets were likely to be on the other side of the left of the graph.

The last topic we consider here is topic 9. This topic is related to the main topic of 5G misinformation since the top three words shown are "5G," "microchip," and "DNA." As discussed before, a portion of people believe that the COVID-19 vaccines can change our DNA and also the vaccines contain a 5G microchip that Bill Gates wants to implant in humans. Since the main keywords were directly linked to the misinformation so that this topic had a significantly high number of false tweets as seen in Figure 6.18e, at 64.7% and the percentage of debunked tweets was 3.4%.

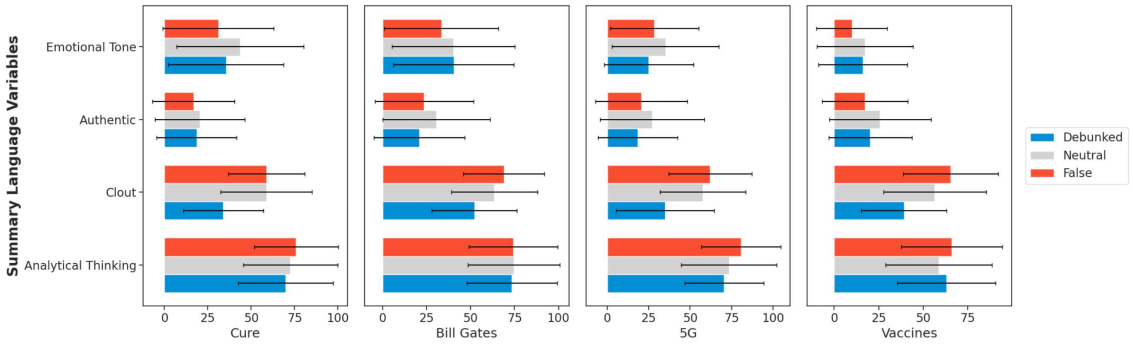


Figure 6.19: Summary Language Variables.

6.5 LIWC results comparison

Linguistic Inquiry and Word Count (LIWC) is a text analysis program that calculates the percentage of words in a given text that fall into one or more of over 80 linguistic, psychological, and topical categories indicating various social, cognitive, and affective processes. In this section, we compared LIWC results between four topics of misinformation by using LIWC2015. Assuming that words contained in texts that are read and analyzed by LIWC2015 are referred to as target words and words in the LIWC2015 dictionary file will be referred to as dictionary words. The default LIWC2015 Dictionary is composed of almost 6,400 words, word stems, and select emoticons. Each dictionary entry additionally defines one or more word categories or subdictionaries. For example, the word "cried" is a part of five word categories: sadness, negative emotion, overall affect, verbs, and past focus. Hence, if the word "cried" is found in the target text, each of these five subdictionary scale scores will be increased [52].

Summary Language Variables

First of all, we consider the Summary Language Variables category which is shown in Figure 6.19. In this category, we chose to visualize four subcategories which are Emotional Tone, Authentic, Clout, and Analytical Thinking. The Emotional Tone category measures the tone of written messages; it is a psycholinguistic variable that summarizes the presence of positive and negative emotions in texts as the difference between positive-emotion words and negative-emotion words. The higher the score the more positive the tone [52]. In general, the neutral group showed a

higher value compared to the other two groups since the authors of these tweets tend to express their thought and their feeling more than other groups, and then it is likely to be the debunked group. Given the fact that scores below 50 suggest a more negative emotional tone, therefore, the discussion between tweets in the three groups tends to have a negative emotional tone because they all discussed important issues being concerned by society at that time, and more specifically, it was the COVID-19 pandemic.

Next, Authentic scores are used to detect writing that is honest and personal in nature. Basically, when people reveal themselves in an "authentic" or honest way, they tend to speak more spontaneously and do not self-regulate or filter what they are saying. From Figure 6.19, we can notice that Authentic scores in tweets are small, which are always 30 or below. These numbers show that tweets may include prepared texts and people are being socially cautious. Once again, for Authentic scores, the neutral group is higher than the other two groups which shows that neutral tweets tend to be spontaneous conversations between friends or political leaders with little-to-no social inhibitions.

Clout refers to the relative social status, confidence, or leadership that people display through their writing or talking. Clout words suggest the author is speaking confidently and with expertise. Generally, false tweets and neutral tweets have higher Clout scores than debunked tweets in which false tweets scores are slightly higher than neutral tweets. While false tweets are created to persuade people to believe in misinformation, their words must show a high level of confidence and be convincing by a significantly high level of Clout in texts compared to the debunked group.

The last variable considered in this group is Analytical Thinking. Analytical Thinking indicates the degree to which people use words that suggest formal, logical, and hierarchical thinking patterns. People low in Analytical Thinking tend to write and think using language that is more intuitive and personal. Language scoring high in Analytical Thinking tends to be rewarded in academic settings and is correlated with things like grades and reasoning skills. Language scoring low in Analytical Thinking tends to be viewed as less cold and rigid, and more friendly and personable [53]. Generally, in all four topics, Analytic scores are very high, above 60 for three types of categories.

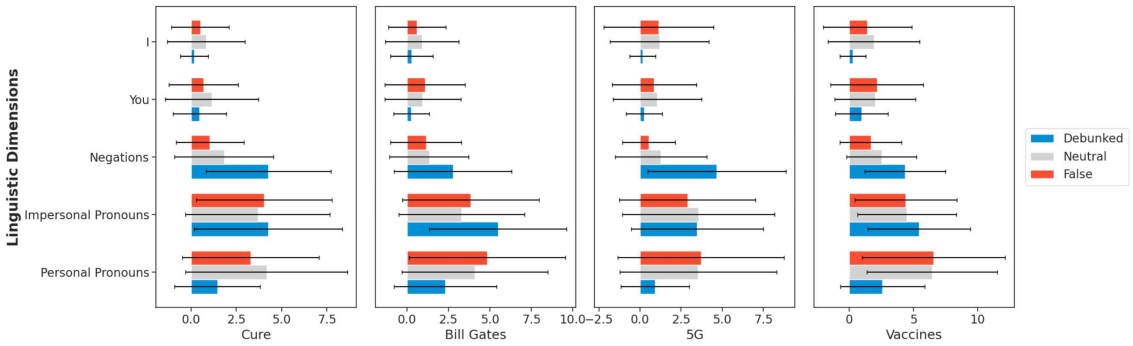


Figure 6.20: Linguistic Dimensions.

Linguistic Dimensions

As shown in Figure 6.20 when considering Linguistic Dimensions, neutral and false groups tend to use pronouns that focus on others (e.g., "you") rather than the debunked group, which often uses impersonal pronouns (e.g., "it"). This is because they usually express their thoughts and opinion as in neutral tweets, and also refer to the person they are addressing or to other people and things related to the misinformation as in false tweets. Moreover, the use of first personal pronouns (e.g., "I") and second personal pronouns (e.g., "you") in neutral tweets are similar and neutral tweets tend to use "I" more than the other two groups which can show that people in this group basically discuss between each other, likely to be a conversation between friends. Whereas debunked tweets mostly use more impersonal pronouns as these pronouns describe or stand for a thing or verb or any nonliving thing but not for a person since debunking often shows the falseness of an idea or belief rather than talking about a particular person. In terms of the use of negations such as "no," "not," and "never," debunked group is significantly higher than the other two groups. Since debunked tweets are often used to expose false information and provide readers with authentic information or cite sources that contain real information to counter misinformation, therefore, they usually use negation words to firmly reject the misinformation.

Psychological Processes

In Psychological processes, as shown in Figure 6.21, we considered five subprocesses which are Biological processes, Perceptual processes, Cognitive processes, Social

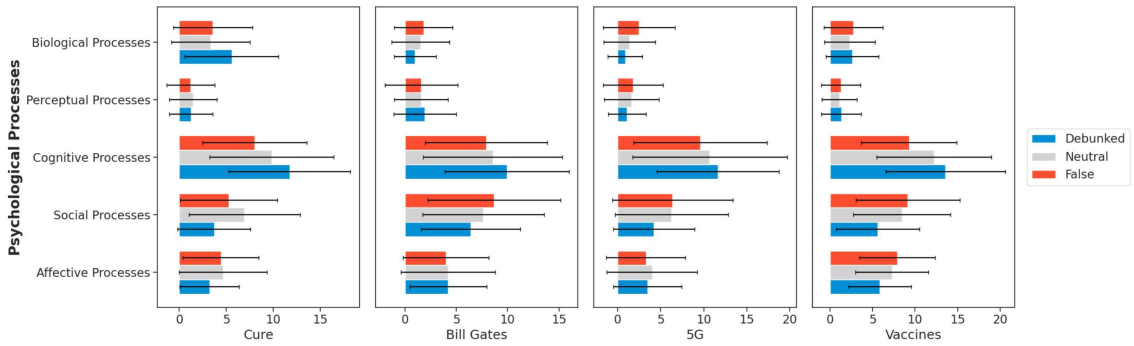


Figure 6.21: Psychological Processes.

processes, and Affective processes. In Biological processes (e.g., eat, blood, pain), the frequency of using words related to these processes is higher in the topics of Cure and Vaccines since these two topics are directly related to Health which has a strong connection with Biological processes, especially in debunked tweets on the Cure topic where they mostly talked about the direct bodily consequences of mistakenly believing misinformation about cures for COVID-19. Next, for Perceptual processes (e.g., look, heard, feeling), this category has the lowest scores compared to other processes in this part, and particularly, there is not too much difference between the four topics and also between the three types of tweets because this category seems very general. Then, Cognitive processes consist of words such as cause, know, ought, and think. Generally, the behaviors in Cognitive processes between the four topics are similar, and we will talk about it in more detail in the next part. Moving to the next category, Social processes contain some words related to family and friends such as mate, talk, they, girl, and boy. Depending on the topic but in general, neutral and false tweets tend to use words related to these processes than debunked tweets because debunked tweets, more or less, refer to scientific evidence and expose false news rather than social conversations where they can discuss and express their own feeling. Finally, for Affective processes (e.g., happy, cried, love, hurt), there is also not too much difference between the three types of tweets in the four topics, and to be more specific about these processes, we consider the next part.

Affective Processes

Affective processing is fundamental to human behavior, which consists of the natural feeling of humans which are Sadness, Anger, Anxiety, Negative emotion, and

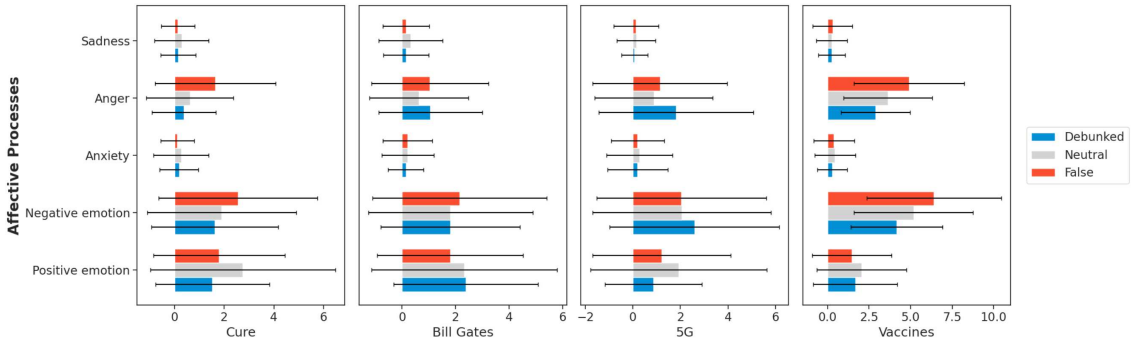


Figure 6.22: Affective Processes.

Positive emotion, as shown in Figure 6.22. It can be noticed that Sadness, which contains words such as "crying," "grief," and "sad", and Anxiety words such as "worried" and "fearful" did not appear much in the text of tweets in all four topics. In terms of Anger, topics with misinformation related to Cure and Vaccines used words in the Anger dictionary such as "hate," "kill," and "annoyed" mostly in false tweets, while on the topics of Bill Gates and 5G, the authors of neutral tweets showed less anger feeling rather than the other two groups. In general, negative and positive emotions were expressed a lot through tweets and they varied depending on the topic. Particularly, on the topic of Cure, Bill Gates, and Vaccines, false tweets had more negative emotion while on the topic of 5G, debunked tweets showed more negative emotion. In terms of positive emotion, the topics of Cure and Bill Gates had higher scores than the other two topics.

Cognitive Processes

Cognitive processes are made up of six sub-scores (Insight, Causation, Discrepancy, Tentativeness, Certainty, and Differentiation).

In Figure 6.23, we can easily notice the difference in the behavior of false and debunked tweets in the Differentiation index. Particularly, the Differentiation index shows the contrast in the sentence by using words such as "but," "not," "if," and "else". Therefore, from the Figure, debunked tweets have significantly higher scores than false tweets in all four topics since debunking exposes the falseness of an idea or belief, in order to do so, the authors tend to make the contrast between sentences by using these words and moreover, they want to emphasize the information they are debunking is false.

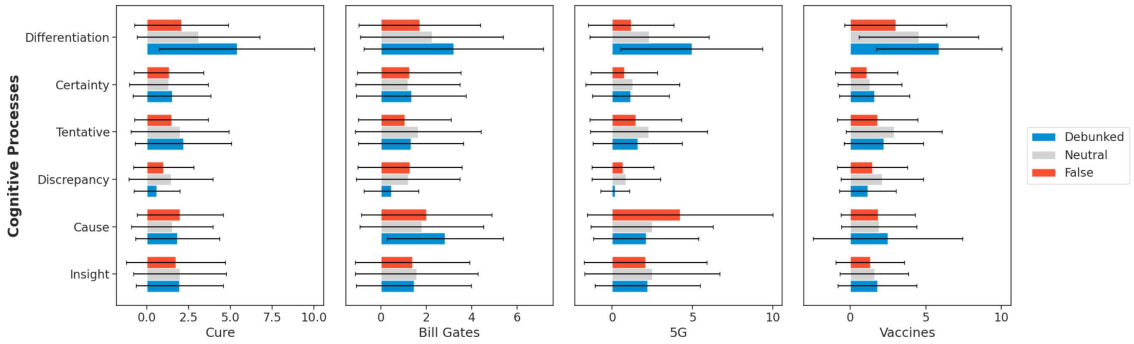


Figure 6.23: Cognitive Processes.

Furthermore, in terms of Certainty (e.g., always, never) in tweets, there is no significant difference between the four topics as well as the three types of tweets. However, the debunked group is slightly superior to the false group in Certainty since in the way of writing, as mentioned above debunked tweets usually use scientific evidence or refer to reliable sources so that they use more words that reflect certainty. By contrast with Certainty, Tentativeness is the quality or state of uncertainty or hesitancy. In the topics of misinformation related to Bill Gates, 5G, and Vaccines, Tentative (e.g., maybe, perhaps, guess) scores of neutral tweets are higher than the other two groups, whereas, in the topic of Cure, debunked tweets' scores are slightly higher. In general, the levels of certainty and tentative expressed are not significantly different among topics and also among the three types of tweets, however.

Discrepancy generally shows the difference between the present (i.e. what is now) and what could be (i.e. what would, should, or could be). In the Discrepancy index, neutral and false tweets have significantly higher scores than the debunked tweets which are already shown in the WordCloud section. The reason is that debunked tweets reveal false information rather than talking about ability, possibility, or necessity.

Basically, Causation means one thing is a reason why something else happens. In LIWC, Causation's dictionary contains words such as because, effect, why, and how. As seen in the Figure, the Causation index shows a significantly high score for false tweets on the topic of 5G. Because, in this specific topic, a portion of people believe that there is a link between the 5G mobile network (particularly the cell towers) to the coronavirus pandemic and this conspiracy theory led to an event

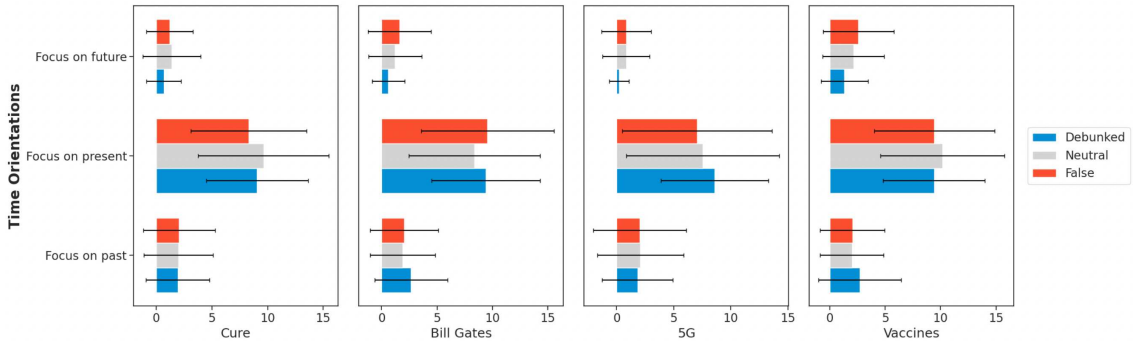


Figure 6.24: Time Orientations.

that many 5G towers were set on fire. On Twitter, they discussed that "5G causes COVID-19" as we saw in the WordCloud section that the word "cause" appeared a lot in this topic. Whereas, neutral tweets got a little lower score than the other two groups in topics related to Cure and Bill Gates. On the topic of Vaccines, the debunked group had a higher score than the other two, but the standard deviation is very large.

Finally, for Insight words such as "think," and "know," there is not too much difference among groups and between topics.

Time Orientations

In this part, we divide into 3 sub-categories which are focused on the future, focus on the present, and focus on the past which is visualized in Figure 6.24. In general, tweets were mostly written with content focused on the present since the scale of that category is significantly higher than the other two categories. Furthermore, false and neutral tweets tended to focus on the future more than debunked tweets. To explain this, Heraclitus said "There is nothing permanent except change," as one thing may be true at this exact moment but will be false in the future and vice versa. This is the reason why in order to guarantee the truth of information, debunked tweets rarely focus on the future as the truth may change over time. Whereas, there is no considerable difference in the behavior of tweets in terms of words related to the past.

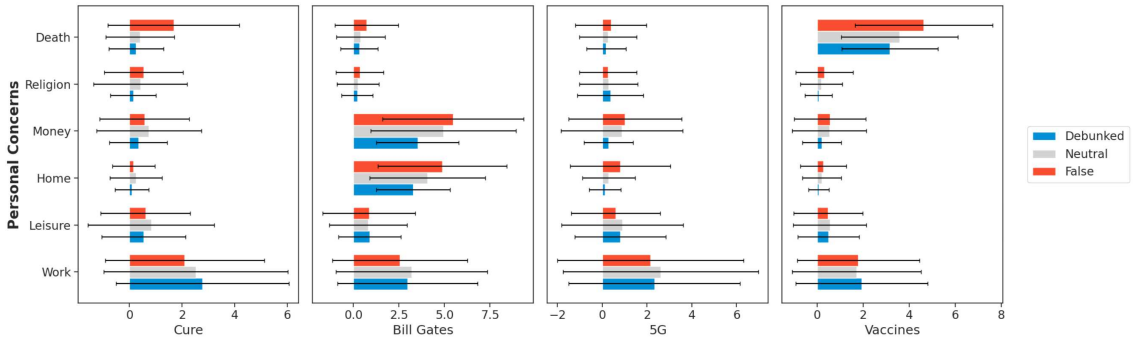


Figure 6.25: Personal Concerns.

Personal Concerns

Personal concerns describe the issues people care about as well as the purpose of the conversations.

As shown in Figure 6.25, each category behaves differently depending on the topic of misinformation. In terms of Death, topics directly related to health like Cure and Vaccines have higher scores than other topics, and especially false tweets tend to use words belonging to this category more than the other two types. There were significantly high scores of the Death index for tweets on the topic of Vaccines because we used keywords containing the word "kill" to target the misinformation community. Moreover, words that belong to Religion and Leisure did not appear much in tweets as shown by lower scores compared to other indices. For Money and Home, words related to these categories were used most in the misinformation topic of Bill Gates with a significantly high proportion of words seen in false tweets compared to debunked ones. Finally, for words related to the Work category, there is no considerable difference between the topics.

Informal Language

In the Informal Language category, we compare the language used in tweets which are Swear words, Netspeak, and Assent, as visualized in Figure 6.26. It is worth noting that there are mostly debunked tweets using swear words on the topics of Bill Gates and 5G and a significantly high percentage of false tweets using swear words on the topics of Cure and Vaccines compared to debunked tweets, interestingly. Furthermore, the proportion of using Netspeak (e.g., btw, lol, thx) in both false

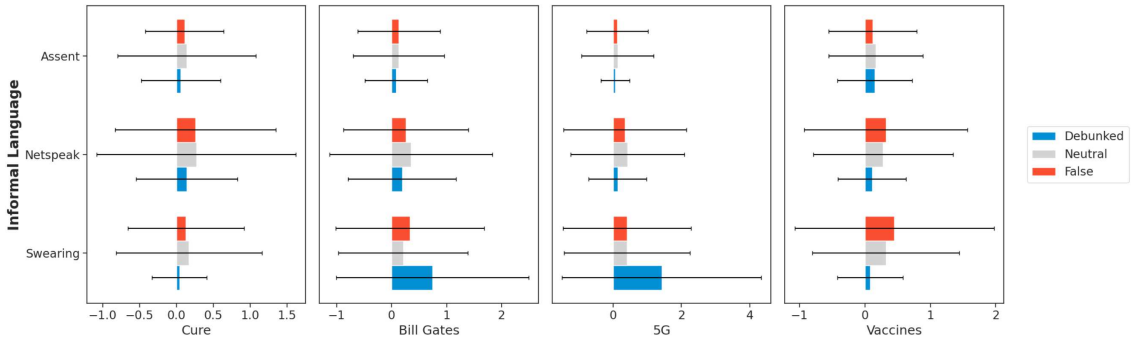


Figure 6.26: Informal Language.

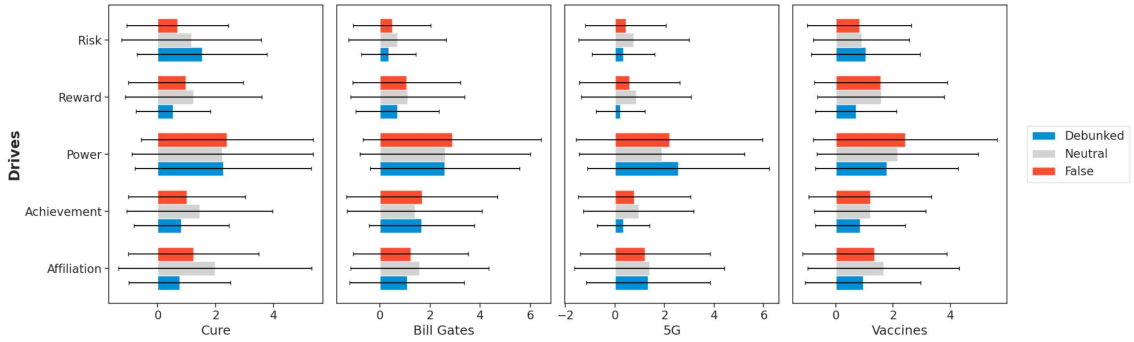


Figure 6.27: Drives.

and neutral tweets is higher than debunked ones. For the last category in this part, Assent, which contains words such as "OK," "agree," and "yes", there is no considerable difference between the topics.

Drives

In general, Drives refer to increased arousal and internal motivation to reach a particular goal. Psychologists differentiate between primary and secondary drives. Primary drives are directly related to survival and include the need for food, water, and oxygen. Secondary or acquired drives are those that are culturally determined or learned, such as the drive to obtain money, intimacy, or social approval. These drives motivate people to reduce desires by choosing responses that will most effectively do so. For instance, when a person feels hunger, he or she is motivated to reduce that drive by eating; when there is a task at hand, the person is motivated

to complete it[8]. In LIWC algorithm, five sub-categories are used to describe the Drives in text, which are Risk, Reward, Power, Achievement, and Affiliation.

Firstly, Risk is the probability or likelihood that a negative event will occur, it consists of some words such as "danger," and "doubt" in text. From Figure 6.27, Risk has somehow very low scores compared to other categories. Moreover, we can easily notice that on the macro topics of Cure and Vaccines, the total Risk scores in general and in debunked tweets specifically seem higher than the other topics. This is because these two topics are directly related to Health and the authors of the debunked tweets want to raise health awareness so that people do not believe misinformation that can be harmful to their health.

A reward is something given or done in return for good (or, more rarely, evil) received[9]. In LIWC, Reward's dictionary contains words such as "take," "prize," and "benefits." In this category, debunked tweets have lower scores than false and neutral tweets. Moreover, setting aside neutral tweets, false tweets basically refer to false information that the author usually tries to convince the public to believe that this information is accompanied by a conspiracy theory about a specific group of people or government that will benefit from defrauding the community and people should not believe it.

Power scores have the largest scale among other categories. Power is the capacity to influence others, even when they try to resist this influence[10], some words belonging to its dictionary are "superior," and "bully." Generally, there is not too much difference between the groups and also between the three types of tweets in this category.

Achievement is the desire to perform well and be successful[10], it is performed by using words such as "win," "success," and "better." Similarly to Power, the Achievement scores vary depending on the topics.

And for the last category in Drives, we consider Affiliation scores in which Affiliation's dictionary contains words such as "ally," "friend," and "social". Affiliation is a social relationship in which a person joins or seeks out one or more other individuals, usually on the basis of liking or a personal attachment rather than perceived material benefits[10]. Due to that reason, based on the Figure above, we can see

[8]https://psychology.iresearchnet.com/social-psychology/social-psychology-theories/drive-theory/

[9]Collins Dictionary

[10]American Psychological Association

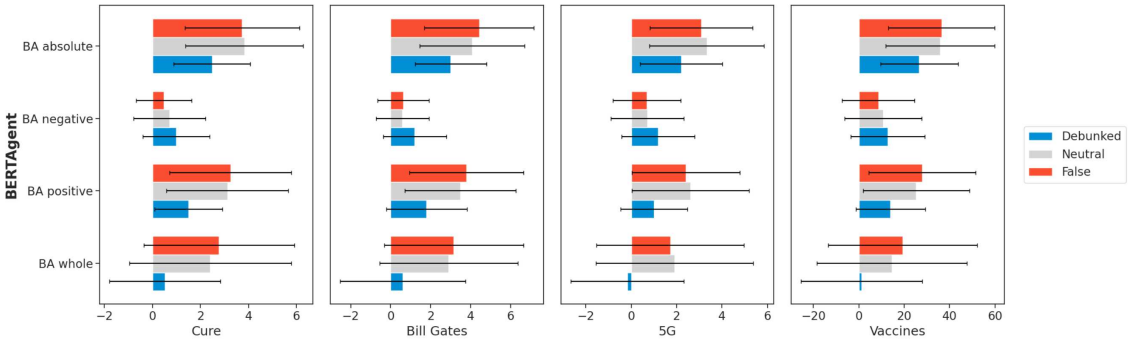


Figure 6.28: BERTAgent.

that in all four topics, neutral tweets tend to have higher Affiliation scores than the other two groups.

BERTAgent

Agency is defined as the ability to assign goals and plan and execute their achievement. Moreover, it refers to the ability of agents to "act on their own behalf," generate causal effects and participate in interactions with their environment while maintaining their own integrity and a considerable degree of autonomy [54]. In this part, we used the BERTAgent algorithm which was introduced in the paper [54], and the results were visualized in Figure 6.28.

In BERTAgent, we consider four indices which are BA absolute (a sentence-wise average of the absolute value of all scores), BA negative (a sentence-wise average of all negative scores), BA positive (a sentence-wise average of all positive scores), and BA whole (a sentence-wise average of all scores). And since BA whole is the difference between BA negative and BA positive, this index plays an important role in defining the agency score of the text. As mentioned above, Agency refers to the ability of agents to "act on their own behalf," generate causal effects and participate in interactions with their environment while maintaining their own integrity and a considerable degree of autonomy. In particular, Agency-positive (BA positive) indicates the feeling in control of one's body, mind, and environment. Therefore, false and neutral tweets have significantly higher BA positive than debunked tweets since in debunked tweets, the authors often do not express their own feeling or thought, they can only say about the truth rather than an individual's opinion. By contrast, in Agency-negative (BA negative) these feelings are not under one's

control. That is the reason why the debunked group has a higher score than the other two groups. From the Figure, we can easily notice that debunked tweets have significantly lower BA whole scores compared to the false and neutral groups. Since BA whole scores are calculated by subtracting negative agency from positive agency to provide overall agentic saturation in texts. Consequently, false tweets seem to tell and encourage people to do something such as "taking hydroxychloroquine as a cure for COVID-19," "5G causes the pandemic, let's burn the cell towers," "Bill Gates' vaccines aim to implant microchips for depopulation, do not take the vaccines," whereas debunked tweets basically tell and inform people about the misinformation and what is the truth instead.

7 Conclusion

Social media plays an important role to help people stay connected and informed about events happening across the globe or in other people's lives. People have become more conscious thanks to social media. It serves as a channel for information, thus paving the way to innovation and success via developing their knowledge and abilities. Social media well-covers global events, making people more aware of their surroundings. However, misinformation on social media rose to prominence in 2016 during the United States of America presidential election, leading people to question science, true news, and societal norms. Misinformation is increasingly affecting societal values, changing opinions on critical issues and topics as well as redefining facts, truths, and beliefs [12]. Moreover, the outbreak of the SARS-CoV-2 novel coronavirus (COVID-19) recently has been accompanied by a large amount of misleading and false information about the virus. In addition, this misinformation during the epidemic negatively affects human health since a portion of people mistakenly believed in the misinformation which led to painful consequences.

Understanding the importance of authentic news and the debunking of misinformation in the time of the pandemic, this thesis aims to build a BERT classification model to distinguish between false, neutral and debunked tweets on Twitter and we especially focus on four macro topics related to COVID-19 which are "Cure," "Bill Gates," "5G," and "Vaccines." Particularly, neutral tweets were the majority in each dataset of the four macro topics, followed by false tweets and debunked tweets only

accounting for no more than 4.2% in the total of tweets. Moreover, words used in false and neutral tweets are not so different but the neutral group likely uses the word "conspiracy" in their tweets when discussing things related to misinformation. While in debunked tweets, the similarity between the four macro topics is that they mostly had the word "claim" to indicate the misinformation that needs to debunk, together with some other words such as "fact," "evidence," and "false.". Next, false tweets had a tendency to use longer hashtags that directly targeted the misinformation than the other two types of tweets. Meanwhile, debunked tweets used mostly #fakenews and #factcheck. There is a significant contrast between the false and debunked groups when considering their temporal behavior. In particular, if the number of debunked tweets is prone to increase over a certain time, the number of false tweets decreases respectively, and vice versa. Similar trends are seen for the number of retweet counts, like counts, and reply counts during the concerned time among the four macro topics. Furthermore, false tweets are likely to have more interactions with other people than debunked tweets and people also prefer to like a post rather than retweet and reply considering the same post.

Furthermore, the linguistic aspects of tweets were also studied using BERTAgent and LIWC to better understand the differences between the three types of tweets. In general, similar behaviors are seen in the four macro topics. Particularly, false tweets have a higher level of confidence conveyed in the texts since they convince people to believe in misinformation. Additionally, this group of tweets tends to use personal pronouns that focus on others because misinformation usually targets a person or a group of people who are involved in issues of social concern to distort the truth. False tweets are also more "reader-friendly" with a high frequency of words related to Social processes compared to debunked tweets and they focus more on the possibility with a higher frequency of using words "would, should, or could." Moreover, because they are not scientific texts, Internet slang (NetSpeak) is more likely to be used. Meanwhile, debunked tweets mostly used impersonal pronouns and a high frequency of negations appeared in their texts rather than in the other two groups. Furthermore, they used certainty words and did not refer to the possibility of an event like false tweets. This is because debunked tweets are used to expose the falseness of an idea or belief rather than talking about an opinion, additionally, they have to use scientific evidence or refer to a reliable source. Since the truth may change over time, debunked tweets rarely use words that focus on

the future. From the Drives and BERTAgent scores, false tweets seemed to tell and encourage people to do something while debunked tweets just aimed to tell and inform people about the misinformation and what is the truth instead. Neutral tweets seemed to have a higher emotional tone rather than the other two groups. In addition, they likely used the first person singular pronoun "I" and also had a higher level of words related to Social processes since they tend to express their own thought and opinion. Sadness and anxiety scores were also seen in this group even though they are very low. Moreover, neutral tweets contained hesitancy words, and Internet slang was used more frequently.

Finally, the possible future work is increasing the learning capacity of the BERT model to distinguish between false, neutral, and debunked tweets by training it with a higher number of samples. Moreover, since misinformation related to COVID-19 is not limited to just four topics, a more diverse dataset is essential to improve the generalization of the model when applied in practice. Since the model is still misclassified between false and neutral tweets, improving the ability to learn the different patterns between these tweets is needed to increase the performance of the model.

8 Appendix

8.1 Evaluation metrics

Confusion Matrix

Confusion Matrix is a performance measurement for machine learning classification. This matrix helps us to compare the resulting classification outcomes with the actual values of the given observation to judge the performance of the classification model.

From the Table 8.1, there are four possible outcomes:

- True Positive (TP) indicates the model predicted an outcome of true, and the actual observation was true.
- False Positive (FP) indicates the model predicted a true outcome, but the actual observation was false.

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Table 8.1: Confusion Matrix

- False Negative (FN) indicates the model predicted a false outcome, while the actual observation was true.
- True Negative (TN) indicates the model predicted an outcome of false, and the actual outcome was also false.

Furthermore, confusion matrices can be used to calculate performance metrics for classification models. The most common are Accuracy, Precision, Recall, and F1-score.

Accuracy

From the Table 8.1, Accuracy can be calculated as followed

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN},$$

which simply divides all true positive and true negative cases by the total number of all cases. Even though accuracy is commonly used to judge model performance, there are a few drawbacks that must be considered before using accuracy liberally. When dealing with unbalanced datasets where one class, either true or false, is more common than the other causing the model to classify observations based on this imbalance. For example, if 90% of cases are false and only 10% are true, there is a very high possibility of our model having an accuracy score of around 90%. Naively, it may seem like we have a high rate of accuracy, but in reality, we are just 90% likely to predict the "false" class, so we do not actually have a good metric. Therefore, there are other metrics that can be considered in this case which are Precision, Recall, and F1-score.

Precision

Precision is the measure of the number of true positives over the total of positives predicted by the model. The formula for precision can be written as followed

$$\text{Precision} = \frac{TP}{TP+FP}.$$

This metric allows you to calculate the rate at which the positive predictions are actually positive.

Recall

Recall, which is all known as sensitivity, is the measure of the true positive over the total of actual positive outcomes. It can be calculated as followed

$$\text{Recall} = \frac{TP}{TP+FN}.$$

This formula allows us to know how well the model is in terms of identifying the actual true results.

F1-score

The F1-score is the harmonic mean between precision and recall. The formula for the F1-score can be expressed as

$$\text{F1-score} = \frac{2(pr)}{p+r},$$

where p is precision and r is recall. This score can be used as an overall metric that incorporates both precision and recall. The reason why F1-score uses the harmonic mean rather than the regular mean is that the harmonic mean punishes values that are further apart.

References

[1] K. Niburski and O. Niburski, "Impact of trump's promotion of unproven covid-19 treatments and subsequent internet trends: observational study," *Journal of medical Internet research*, vol. 22, no. 11, p. e20044, 2020.

[2] Rani Horev. Bert explained: State of the art language model for nlp. [Online]. Available: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] Nicolo Cosimo Albanese. Fine-tuning bert for text classification. [Online]. Available: https://towardsdatascience.com/fine-tuning-bert-for-text-classification-54e7df642894

[5] WHO. Coronavirus disease 2019 (covid-19) situation report - 86. [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200415-sitrep-86-covid-19.pdf

[6] L. Garrett, "Covid-19: the medium is the message," *The lancet*, vol. 395, no. 10228, pp. 942–943, 2020.

[7] G. K. Shahi, A. Dirkson, and T. A. Majchrzak, "An exploratory study of covid-19 misinformation on twitter," *Online social networks and media*, vol. 22, p. 100104, 2021.

[8] M. S. Islam, T. Sarkar, S. H. Khan, A.-H. M. Kamal, S. M. Hasan, A. Kabir, D. Yeasmin, M. A. Islam, K. I. A. Chowdhury, K. S. Anwar *et al.*, "Covid-19–related infodemic and its impact on public health: A global social media analysis," *The American journal of tropical medicine and hygiene*, vol. 103, no. 4, p. 1621, 2020.

[9] J. Vincent, "Something in the air: Conspiracy theorists say 5g causes novel coronavirus, so now they're harassing and attacking uk telecoms engineers," *New York, NY: The Verge*, 2020.

[10] S. Shahsavari, P. Holur, T. Wang, T. R. Tangherlini, and V. Roychowdhury, "Conspiracy in the time of corona: Automatic detection of emerging covid-19 conspiracy theories in social media and the news," *Journal of computational social science*, vol. 3, no. 2, pp. 279–317, 2020.

[11] P. Institude. The coronavirusfacts database. [Online]. Available: https://www.poynter.org/ifcn-covid-19-misinformation/

[12] F. Olan, U. Jayawickrama, E. O. Arakpogun, J. Suklan, and S. Liu, "Fake news on social media: the impact on society," *Information Systems Frontiers*, pp. 1–16, 2022.

[13] J. Li and X. Chang, "Combating misinformation by sharing the truth: a study on the spread of fact-checks on social media," *Information systems frontiers*, pp. 1–15, 2022.

[14] Anish Agarwal. Covid-19 misinformation: The flip side of 'knowledge is power'. [Online]. Available: https://www.pennmedicine.org/news/news-blog/2022/october/covid-misinformation-the-flip-side-of-knowledge-is-power

[15] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–236, 2017.

[16] E. Ortiz-Ospina, "The rise of social media," *Our World in Data*, 2019, https://ourworldindata.org/rise-of-social-media.

[17] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 2, pp. 1–36, 2018.

[18] P. Hernon, "Disinformation and misinformation through the internet: Findings of an exploratory study," *Government information quarterly*, vol. 12, no. 2, pp. 133–139, 1995.

[19] C. Wardle and H. Derakhshan, *Information disorder: Toward an interdisciplinary framework for research and policymaking*. Council of Europe Strasbourg, 2017, vol. 27.

[20] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[21] J. Jeyasudha, P. Seth, G. Usha, and P. Tanna, "Fake information analysis and detection on pandemic in twitter," *SN Computer Science*, vol. 3, no. 6, pp. 1–10, 2022.

[22] J. A. Nasir, O. S. Khan, and I. Varlamis, "Fake news detection: A hybrid cnn-rnn based deep learning approach," *International Journal of Information Management Data Insights*, vol. 1, no. 1, p. 100007, 2021.

[23] H. Saleh, A. Alharbi, and S. H. Alsamhi, "Opcnn-fake: Optimized convolutional neural network for fake news detection," *IEEE Access*, vol. 9, pp. 129 471–129 489, 2021.

[24] G. Sansonetti, F. Gasparetti, G. D'aniello, and A. Micarelli, "Unreliable users detection in social media: Deep learning techniques for automatic detection," *IEEE Access*, vol. 8, pp. 213 154–213 167, 2020.

[25] A. Agarwal, M. Mittal, A. Pathak, and L. M. Goyal, "Fake news detection using a blend of neural networks: an application of deep learning," *SN Computer Science*, vol. 1, pp. 1–9, 2020.

[26] Y. Liu and Y.-F. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[27] M. Choudhary, S. S. Chouhan, E. S. Pilli, and S. K. Vipparthi, "Berconvonet: A deep learning framework for fake news classification," *Applied Soft Computing*, vol. 110, p. 107614, 2021.

[28] S. Kumari, H. K. Reddy, C. S. Kulkarni, and V. Gowthami, "Debunking health fake news with domain specific pre-trained model," *Global Transitions Proceedings*, vol. 2, no. 2, pp. 267–272, 2021.

[29] J. Xing, S. Wang, X. Zhang, and Y. Ding, "Hmbi: a new hybrid deep model based on behavior information for fake news detection," *Wireless Communications and Mobile Computing*, vol. 2021, pp. 1–7, 2021.

[30] Z. Kvetanová, A. K. Predmerská, and M. Švecová, "Debunking as a method of uncovering disinformation and fake news," *Fake News Is Bad News-Hoaxes, Half-truths and the Nature of Today's Journalism*, 2020.

[31] S. Jiang and C. Wilson, "Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–23, 2018.

[32] S. Lewandowsky, U. K. Ecker, C. M. Seifert, N. Schwarz, and J. Cook, "Misinformation and its correction: Continued influence and successful debiasing," *Psychological science in the public interest*, vol. 13, no. 3, pp. 106–131, 2012.

[33] B. Swire-Thompson, N. Miklaucic, J. P. Wihbey, D. Lazer, and J. DeGutis, "The backfire effect after correcting misinformation is strongly associated with reliability." *Journal of Experimental Psychology: General*, 2022.

[34] G. Pennycook, T. D. Cannon, and D. G. Rand, "Prior exposure increases perceived accuracy of fake news." *Journal of experimental psychology: general*, vol. 147, no. 12, p. 1865, 2018.

[35] B. Nyhan and J. Reifler, "When corrections fail: The persistence of political misperceptions," *Political Behavior*, vol. 32, no. 2, pp. 303–330, 2010.

[36] A. Roets *et al.*, "'fake news': Incorrect, but hard to correct. the role of cognitive ability on the impact of false information on social impressions," *Intelligence*, vol. 65, pp. 107–110, 2017.

[37] E. Thorson, "Belief echoes: The persistent effects of corrected misinformation," *Political Communication*, vol. 33, no. 3, pp. 460–480, 2016.

[38] J. Paynter, S. Luskin-Saxby, D. Keen, K. Fordyce, G. Frost, C. Imms, S. Miller, D. Trembath, M. Tucker, and U. Ecker, "Evaluation of a template for countering misinformation—real-world autism treatment myth debunking," *PloS one*, vol. 14, no. 1, p. e0210746, 2019.

[39] P. Smith, M. Bansal-Travers, R. O'Connor, A. Brown, C. Banthin, S. Guardino-Colket, and K. M. Cummings, "Correcting over 50 years of tobacco industry misinformation," *American journal of preventive medicine*, vol. 40, no. 6, pp. 690–698, 2011.

[40] H. Yousuf, S. van der Linden, L. Bredius, G. T. van Essen, G. Sweep, Z. Preminger, E. van Gorp, E. Scherder, J. Narula, and L. Hofstra, "A media intervention applying debunking versus non-debunking content to combat vaccine misinformation in elderly in the netherlands: A digital randomised trial," *EClinicalMedicine*, vol. 35, p. 100881, 2021.

[41] N. Walter and S. T. Murphy, "How to unring the bell: A meta-analytic approach to correction of misinformation," *Communication monographs*, vol. 85, no. 3, pp. 423–441, 2018.

[42] M.-p. S. Chan, C. R. Jones, K. Hall Jamieson, and D. Albarracín, "Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation," *Psychological science*, vol. 28, no. 11, pp. 1531–1546, 2017.

[43] N. Walter, J. J. Brooks, C. J. Saucier, and S. Suresh, "Evaluating the impact of attempts to correct health misinformation on social media: A meta-analysis," *Health Communication*, vol. 36, no. 13, pp. 1776–1784, 2021.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[45] Britney Muller. Bert 101 state of the art nlp model explained. [Online]. Available: https://huggingface.co/blog/bert-101

[46] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[47] T. Joachims, "A probabilistic analysis of the rocchio algorithm with tfidf for text categorization." Carnegie-mellon univ pittsburgh pa dept of computer science, Tech. Rep., 1996.

[48] M. R. Mehra, S. S. Desai, F. Ruschitzka, and A. N. Patel, "Retracted: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of covid-19: a multinational registry analysis," 2020.

[49] A. Rahman, T. Tabassum, Y. Araf, A. Al Nahid, M. A. Ullah, and M. J. Hosen, "Silent hypoxia in covid-19: pathomechanism and possible management strategy," *Molecular biology reports*, vol. 48, no. 4, pp. 3863–3869, 2021.

[50] A. Tsiang and M. Havas, "Covid-19 attributed cases and deaths are statistically higher in states and counties with 5th generation millimeter wave wireless telecommunications in the united states." *Medical Research Archives*, vol. 9, no. 4, 2021.

[51] . Chemtrails conspiracy theory. [Online]. Available: https://keith.seas.harvard.edu/chemtrails-conspiracy-theory

[52] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," Tech. Rep., 2015.

[53] . Liwc analysis. [Online]. Available: https://www.liwc.app/help/liwc

[54] T. E. L. D. Jan Nikadon, Caterina Suitner and M. Formanowicz, "Bertagent: A novel tool to quantify agency in textual data," *The paper is submitted to the Behavior Research Methods.*, 2023.

Acknowledgments