MISINFORMATION IDENTIFICATION USING NATURAL LANGUAGE PROCESSING

JIA YING OU

A THESIS SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

GRADUATE PROGRAM IN ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
YORK UNIVERSITY
TORONTO, ONTARIO

October 2021

Abstract

The popularity of social media has accelerated the speed and scope of fake news propagation, and exacerbated the harm caused by false information. Identifying misinformation is crucial to maintain a country's political, social, financial stability and democracy. In this thesis, we study the problem of misinformation identification using natural language processing (NLP). Given a claim, our approach classifies a claim as true, partly true or false using a set of news articles whose contents are related to the claim. The set of related articles, collected from reputable sources, serves as the ground truth to assess the validity of the claim.

Using this approach of misinformation identification, the contributions of this thesis is to address the following research problems:

- We constructed a new large-scale, feature-rich dataset of COVID-19 news and facts for research on COVID-19 misinformation, which is named COVMIS. We provide a comprehensive analysis of the dataset to better understand the data, including claim contents, article contents, publication dates, news sources, and country distribution. We also discuss potential use cases to demonstrate the benefits of the dataset for research on misinformation-related COVID-19 and other areas.

- We conducted two sets of extensive experiments to evaluate several state-of-the-art transformer-based NLP models using the COVMIS dataset. The models that were evaluated are BERT (Bidirectional Encoder Representations from Transformers), DistilBERT, XLNet (Generalized Autoregressive Pretraining for Language Understanding), ALBERT (A Lite BERT), and RoBERTa (Robustly Optimized BERT Pre-training Approach). The first set of experiments shows that BERT performs the best in terms of F1 score. In the second set of experiments, we evaluated an optimization: instead of inputting all articles related to a claim to classify the claim, we extracted and input only a subset of K sentences (e.g., K = 5) that are the most relevant to the claim. Experimental results show that this optimization improves the performance of the models in terms of accuracy, F1 score, precision and recall, given different values of K.

- We conducted two sets of extensive experiments on a news classification model based on BERT and evaluated the performance of the model in terms of accuracy, F1 score, precision, and recall. We used two datasets: (i) the general news dataset provided by the Fake News Challenge competition and (ii) the COVMIS dataset mentioned above. The first set of experiments was designed to answer the question of whether narrowing down the domain of knowledge (i.e., COVID-related news vs. general news) will improve the classification performance. Our experimental results show that the classification performance of the model improves significantly when the domain of knowledge of the dataset is narrowed down to a specific area of interest, COVID-19 in this case. The second set of experiments quantified how obsolete training data affect the classification performance. Our experimental results show that the more up-to-date the training data (relative to the test data), the better the classification performance.

Acknowledgments

First and foremost, I would like to thank my supervisor, U.T Nguyen, for all her support, patience and guidance throughout my Master's degree. I offer my gratitude to all committee members for their time and valuable comments that contributed to the improvement of my thesis. Also, a special thanks to Dr. Amir Chinaei for all help that he has provided. I am grateful to all of those I have had the pleasure to work with. Last but not least, thank you to my family members for their endless support and encouragement. I would like to express my thanks again to everyone who helped me, I wouldn't be here without you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter, we provide background information and the motivations for our research on misinformation identification, define the problems, and discuss the contributions of the thesis.

1.1 Background Information

Fake news is false information disseminated via various types of media such as newspapers, magazines, television, radio, podcast, blogs, and social networks. There are two forms of false information [2]: misinformation and disinformation. While both denote false information, disinformation results from bad intentions to mislead or cause harm, whereas misinformation is not intended to cause harm. For the sake of brevity, we will use the term ”misinformation“ in the remainder of the thesis to refer to all fake news, including disinformation.

Pew Research Center [3] found that 86% of Americans get news from social media platforms. While misinformation is a problem as old as human history, the popularity of social media has accelerated the speed and scope of misinformation propagation and exacerbated the harm caused by false information. Misinformation has had profound impacts on a country's political and social stability [4][5], democracy [6], financial markets [7][8], and public health [9].

Misinformation can be identified based on news contents (fact-checking), the credibility of the source of information, and patterns of information propagation on the Internet or in social networks [10]. In a global pandemic, misinformation can be as big a threat as the virus itself. According to the US National Institute of Health [11], *"(m)isleading information about treatment for COVID-19 has resulted in an increasing number of vitamin D abuse and even mass poisoning from methanol intake."* In addition, vaccine hesitancy and resistance caused by misinformation has hampered the recovery process in many parts of the world [12]. Therefore, it is crucial to intensify research in misinformation identification, prevention, and mitigation as part of the global effort to fight the pandemic. This thesis focuses on misinformation identification based on natural language processing (NLP) and datasets for misinformation identification models.

In literature, there currently exist four main approaches for NLP-based misinformation identification:

- Stance Detection [13][14][15] [16][17][18][19][20][21][22] [23][24][25]: given an article and its headline (or a claim), we determine whether the headline agrees or disagrees with the article. Depending on the dataset, there can be more categories than "agree" and "disagree". For example, the dataset provided by the 2017 Fake News Challenge [13] contains four categories: "agree","disagree", "discuss", and "unrelated".

- Source Credibility [26][27][28][29]: given a statement, we determine whether the statement is true or not by assessing the credibility of the source(s) from which the statement is extracted.

- Rumour Classification [30][31][32][33]: given a post from a social media platform, we classify whether the post is a rumour or not using user engagements, user activities, and/or contents of user comments. A rumour is defined as a claim or statement with an unverified truth value; it can be either true or false [34].

- Fact Checking [35][36][37][38][39][40][41][42][43]: given a claim (statement), we classify the validity of a claim (e.g., true, partly true or false) using evidence (e.g., a collection of articles whose contents serve as the ground truth).

Actual truth-labelling is a very challenging problem, heavily influenced by technical, political, social and cultural issues. The first three approaches are encouraging first steps towards solutions for misinformation identification, but they still fall short of effective truth labeling. In stance detection, even when a headline agrees with its article, they can be both misinformation. A rumour may be true or false. While rumours traditionally carry the negative notion of being false information, they can be true facts, such as leaks of new smartphone releases ahead of their actual release dates. Similarly, a credible source may make false statements, unintentionally or intentionally. For example, while being widely respected, President Obama told 18 different untruths during his presidency, either unintentionally or as exaggerations [44].

1.2 Motivations and Problem Definitions

In this thesis, we use a different approach that moves us closer to actual truth labelling, which is the automatic fact-checking approach. Given a claim (statement), for example, "NASA has just confirmed earth has a new moon" [45], our model classifies the claim as true, partly true or false using a set of news articles whose contents are related to the claim. The set of related articles, collected from reputable sources, serves as the ground truth to assess the validity of a claim. Table 1.1 shows an example claim and its metadata, and Figure 1.1 shows the set of articles related to this claim.

Two datasets are used for this thesis. The first dataset is from a competition named "2019 Leaders Prize: Fact or Fake News?" [45]. This dataset contains 15,555 claims, each associated with a set of related articles, for a combined total of 64,974 articles. The claims and articles were published before November 2019 and cover many topics such as politics,

Table 1.1: Metadata sample

ID	Claim	Claimant	Date	Label	Related Articles
544	We will have reached the number 25, 000 and thereby kept our promises.	John McCallum	2015-12-03	false	[97661, 97545, 97240]

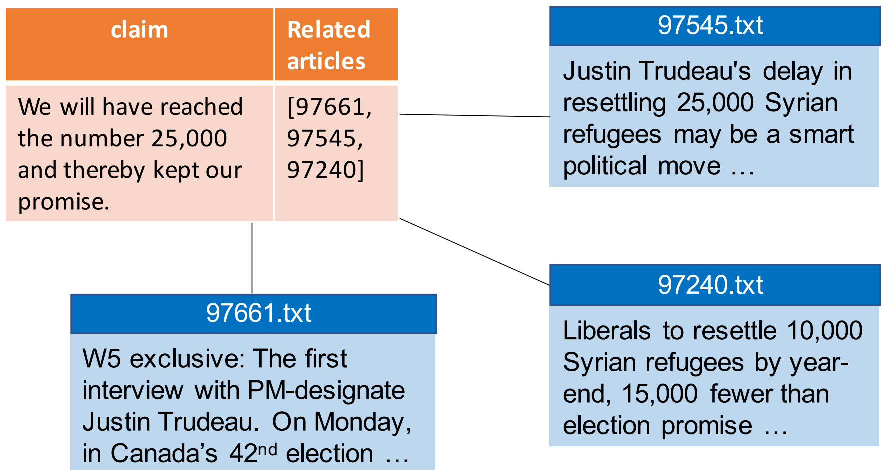


Figure 1.1: A sample of related articles

finance, entertainment, sports and health care. Therefore, we name this set "general news dataset". We collected data and constructed the second dataset, named COVMIS. This dataset contains news and facts related to only COVID-19. It contains 14,384 claims and 134,320 articles.

We are interested in building a NLP-based system to identify misinformation in real time using the fact checking approach stated above. Using this automatic fact-checking approach, we solve the following three research problems:

RQ.1 How do state-of-the-art transformer-based NLP models (e.g., BERT [1], DistilBERT [46], XLNet [47], RoBERTa [48], and ALBERT [49]) perform using the fact checking approach? We used the COVMIS dataset for this research problem. Given a claim, we first combine the articles related to the claim into one document, and then input this document into the above models to evaluate their performance in terms of accuracy, F1 score, precision and recall.

RQ.2 Would the performance of the above models improve if we input only the most relevant sentences extracted from the articles into the model instead of inputting related articles in their entirety? All the above models restrict the token input size to 512 tokens. Additional input beyond the first 512 tokens will be truncated, but the truncated data could be important for the classification task. Therefore, we use information retrieval algorithms (such as TF-IDF and cosine similarity) to overcome the input size limitation by retrieving K sentences that are the most relevant (most similar) to a claim from the set of articles related to the claim. In our experiments, K ranges from 5 to 50. In these experiments, we use the same dataset and models as for the first research question. However, instead of inputting related articles in their entirety, we input only the K most relevant sentences extracted from the set of articles related to a claim. We then analyzed the performance of the models in terms of accuracy, F1 score, precision and recall under this optimization.

RQ.3 How can data affect the performance of NLP-based misinformation identification models? In particular, if the domain of knowledge of a dataset is narrowed down to a specific topic, would the performance of a fact checking model in terms of accuracy, F1 score, precision, and recall be improved? Furthermore, we aim to quantify the performance degradation caused by obsolete training data. To answer question RQ.3, we use both the general news dataset and our newly constructed dataset, COVMIS, and solve the following two sub-problems:

a. **Can the performance (e.g., accuracy, precision, recall, F1 score) of a transformer-based NLP model be improved if the domain of knowledge is narrowed down to a specific area of interest?** We can define specific areas of interest such as the 2020 US election, the stock market, climate change, COVID-19, and construct datasets on each specific topic. A claim on the topic of climate change would then be evaluated using a model trained with a dataset containing only articles on climate change. We hypothesized that this method could result in higher performance than using a general news dataset to evaluate a claim. As an analogy, consider a panel of medical experts taking questions from an audience. An oncologist would answer questions about cancer treatments better than a general practitioner. Our objective is to quantify the performance difference between using a topic-specific dataset vs. a general news dataset for fact checking.

b. **How do obsolete training or testing data affect the performance of a transformer-based NLP model?** A general news dataset collected before November 2019 would not contain any information about the coronavirus pandemic. Therefore, if a COVID-19 related claim is evaluated using a model trained with the general news dataset mentioned above, the result may not be as good because the general news dataset is obsolete compared with the COVID-19 claim.

To solve RQ.3a and RQ.3b, we conducted experiments on a news classification model based on Bidirectional Encoder Representation from Transformers (BERT) and evaluated the performance of the model in terms of accuracy, F1 score, precision, and recall under various data configurations as follows. To answer RQ.3a, we compared the experimental results obtained from the general news dataset with those from the COVMIS dataset. To answer RQ.3b, we conducted several experiments using either obsolete training data or obsolete testing data. Obsolete data refer to those extracted from the general news dataset. The objective is to cause a temporal mismatch between the training and testing data in an experiment. For example, the training data was from the general news dataset, and the claims to be tested were from the COVMIS dataset or vice versa. We are interested in quantifying how such a temporal mismatch between training and testing data affects the classification performance.

1.3 Contributions

The main contributions of the thesis are as follows:

- We constructed a new large-scale, feature-rich dataset of COVID-19 news and facts for research on COVID-19 misinformation, which is named COVMIS. We provide a comprehensive analysis of the dataset to better understand the data, including claim contents, article contents, publication dates, news sources, and country distribution. We also discuss potential use cases to demonstrate the benefits of the dataset for research on misinformation related to COVID-19 and other areas.

- We conducted two sets of extensive experiments to evaluate several state-of-the-art transformer-based NLP models using the COVMIS dataset. The models that were evaluated are BERT (Bidirectional Encoder Representations from Transformers), DistilBERT, XLNet (Generalized Autoregressive Pretraining for Language Understanding), ALBERT (A Lite BERT), and RoBERTa (Robustly Optimized BERT Pre-training Approach). The first set of experiments shows that BERT performs the best in terms of F1 score. In the second set of experiments, we evaluated an optimization: instead of inputting all articles related to a claim to classify the claim, we extracted and input only K sentences (e.g., K = 5) that are the most relevant to the claim. Experimental results show that this optimization improves the performance of the models in terms of accuracy, F1 score, precision and recall, given different values of K.

- We conducted two sets of extensive experiments on a news classification model based on BERT and evaluated the performance of the model in terms of accuracy, F1 score, precision, and recall. We used two datasets: (i) the general news dataset provided by the Fake News Challenge competition and (ii) the COVMIS dataset mentioned above. The first set of experiments was designed to answer the question of whether narrowing down the domain of knowledge (i.e., COVID-related news vs. general news) will improve the classification performance. Our experimental results show that the classification performance of the model improves significantly when the domain of knowledge of the dataset is narrowed down to a specific area of interest, COVID-19 in this case. The second set of experiments quantifies how obsolete training data affect the classification performance. Our experimental results show that the more up-to-date the training data (relative to the test data), the better the classification performance.

1.4 Thesis Organization

The remainder of the thesis is organized as follows. In chapter 2, we review existing misinformation identification approaches, and datasets related to misinformation identification and COVID-19 information. In chapter 3, we introduce a new dataset named "COVMIS" for research in COVID-19 misinformation identification. In chapter 4, we address problems RQ.1 and RQ.2 by evaluating different state-of-the-art transformer-based NLP models using the COVMIS dataset. We quantify how topic-specific data and obsolete data affect the performance of transformer-based NLP models in chapter 5 to address problem RQ.3. We conclude the thesis and outline future research directions in chapter 6.

Chapter 2

Literature Review

This chapter provides a background overview of the fake news study in section 2.1, and then we review NLP-based misinformation identification approaches on misinformation identification in section 2.2. After that, we review datasets related to misinformation identification and datasets related to COVID-19 in section 2.3.

2.1 Misinformation Identification Approaches

Fake news study can be categorized into four groups: knowledge-based study, style-based study, propagation-based study and credibility-based study [10]. In this section, we will provide an overview of these groups.

2.1.1 Knowledge-based Study

A knowledge-based approach in fake news is fact-checking processing in detecting fake news. Currently, there are two kinds of fact-checking [10]: manual fact-checking by humans and automatic fact-checking by machine. There are two types of manual fact-checking [10], expert-based and crowd-source-based. For most fact-checking websites (e.g., PolitiFact, Snopes, Reuters, and Poynter), the validity of given news content is fact-checked manually by human experts. Expert-based manual fact-checking provides accurate results since fact-checkers follow rules when verifying news contents. For example, all fact-checkers who are a member of the International Fact-Checking Network (IFCN) follows the code of principles for nonpartisan and transparent fact-checking [50]: (1) a commitment to non-partisanship and fairness; (2) a commitment to transparency of sources; (3) a commitment to transparency of sources; (4) a commitment to transparency of methodology, and (5) a commitment to open and honest corrections. However, expert-based manual fact-checking is costly and labour intensive, mainly due to the increasing volume of to-be-verified data in our everyday life; this approach lacks scalability. A crowd-sourced manual fact-checking approach increases the scalability. The reason is that crowd-sourced manual fact-checking relies on a large group of individuals. However, this approach is less accurate than the expert-based manual fact-checking approach due to conflict annotations. Therefore, automatic fact-checking is introduced to address the scalability problem and increase the efficiency of the fact-checking process.

Most existing studies for automatic fact-checking rely on natural language processing (NLP) and information retrieval (IR) algorithms. An overview of these studies will be provided in Section 2.2. Before fact-checking a piece of news, we need to first construct a knowledge base by extracting facts from websites. Facts can be extracted from either a single source or multiple sources. It is more efficient to extract information from a single source, but it is less comprehensive than extracting information from multiple sources. Extracted facts from websites are also called raw facts, which need to be processed for five reasons [10]: (i) to reduce data redundancy, (ii) to eliminate outdated data, (iii) to solve label conflict, (iv) to remove unreliable data from untrusted sources, and (v) to address data incompleteness. Then we can use this constructed knowledge base to evaluate the validity of given news content.

2.1.2 Style-based Study

The style-based approach is to investigate the news content and figure out the intention, whether it has a bad intention to mislead or cause harm to the public or not. This approach can be divided into three categories [10]:

- Deception style theories [51]. How is the content style of fake news differ from that of trustworthy news?
- Style-based features and patterns [52–54]. What are the features and patterns that can represent and capture fake news?
- Deception detection strategies [55]. What are the strategies based on content style that can be used to detect fake news?

News articles cover a wide range of topics, such as politics, health care, economics, the financial market, and climate change. That being said, deception information in various fields is different [56]. Therefore, analysis of fake news should be cross-domain, cross-language, and cross-topic. Moreover, since fake news usually increase during significant events, such as elections; a real-time system to learn the content style of fake news should be considered.

2.1.3 Propagation-based Study

The propagation-based approach investigates how fake news propagates across different perspectives, such as domains, topics, websites, languages, and how it differs from the propagation of trustworthy news. There are some major patterns that distinguish fake news from trustworthy news by its propagation. For example, unconfirmed news often gets more noticed than true news; false news spreads farther, faster and more widely than trustworthy news; false political news spreads farther, faster and more widely than false news in other domains [10]. Fake news propagation models can be network-based [57, 58] or cascade-based [59, 60].

For network-based fake news propagation, it can be homogenous (consist of a single type of nodes and edges), heterogeneous (consist of multiple types of nodes and edges), or hierarchical (consist of different kind of nodes and edges that forms a hierarchy relationship) [61] . A homogenous network can be a stance network used to compare the similarity between a stance and other news, or a propagation tree that shows the relationship between the post and repost for each piece of news [10]. A heterogeneous network can be used to establish relationships between features such as the claim, the claimant, the source, the user, the reposts [10]. A hierarchical network can be used to turn fake news detection into a graph optimization problem [61]. Cascade-based fake news propagation is a tree or tree-like structure showing how users propagate fake news on social networks [61]. This approach can be represented by a hops-based fake news cascade that shows the number of steps the fake news has travelled, or by a time-based cascade that shows the number of times that a piece of fake news was posted [61]. Then fake news detection can be done after the discovery of propagation patterns and characteristics.

2.1.4 Credibility-based Study

The credibility-based approach is to assess fake news based on its source. There are four types of information being used for this approach [10]: (i) news headline, (ii) news source, (iii) news comments, and (iv) news spreaders.

News headlines are often used to detect clickbait, which is usually the case when the headline does not match the news content. Clickbait intends to attract people's eyes to increase the number of clicks to a particular website or to read a particular article. The contents on the website or in the article are often deceptive, leading to a high clicking rate for generating revenue and gaining public trust. Although not all articles associated with clickbait are false, clickbait seems to be an indirect way to detect fake news. Existing techniques [62, 63] to detect clickbait are based on linguistic features (i.e., term frequencies, readability and forward reference), non-linguistic features (i.e., webpage links, user interests and headline stances), and deep learning models.

News sources that publish fake news are non-reputable websites or websites intended to generate revenue by promoting certain kinds of information such as clickbait. A common technique to analyze source credibility to use web ranking algorithms [64, 65] to improve search engine results. However, this technique needs to be used with the accompany of web spam detection. Other techniques include assessing web credibility based on content [66], and link features using machine learning [67].

News comments often provide users' opinions about news content. However, comments can be made by "fake supporters" who were recruited to make specific comments under a piece of given news, such as for e-commerce websites. Therefore, this approach needs to be used with review spam detection [68] to ensure comment credibility. Review spam detection can be categorized into three approaches [10]: (i) content-based approach [69] that uses a collection of language features that are extracted from the comments; (ii) behaviour-based approach [70, 71] often uses features extracted from unreliable comments that are associated with user behaviour; (iii) relationship-based approach that take into account the relationships among reviewers, comments and products [72].

News spreaders are often involved in social activities, such as sharing, like, forward and commenting. News spreaders can be classified into two types: malicious users and naïve users. Malicious users intend to spread fake news. These users can be a bot running automatic scripts; a troll trying to provoke others into displaying emotional responses and normalizing tangential discussion; or a cyborg that combines both human and automatic activities, for instance, a cyborg can use a human-created account to execute scripts for online activities automatically. Naïve users do not intend to spread fake news but believe fake news as truth. These kinds of users are mostly influenced by two factors [10]: social influence of environmental and exogenous factors (e.g., network structure, peer pressure), or self-influence of internal and inherent attributes of users that can impact how they engage with fake news, such as political bias or their pre-existing knowledge.There exist classifiers used to classify news spreaders, such as [73].

2.2 NLP-based Misinformation Identification Approaches

Existing literature that studies misinformation identification based on NLP can be categorized into four groups: i) stance detection, (ii) source credibility, (iii) rumour classification and (iv) fact checking. We will review each category in this section.

2.2.1 Stance Detection

The Fake News Challenge (FNC-1)[13] held in 2017 is a milestone in fake news stance detection. FNC-1 defines stance detection as follows: given a headline and an article, we determine whether the headline relates to the article or not. If the article is related to the headline, we determine whether the article agrees with, disagrees with, or discusses the same topic as the headline. FNC-1 also defines its evaluation score, namely the FNC score. FNC score sums up to 1. A team receives 0.25 when correctly classifying whether a headline is related or unrelated to an article. A team receives an additional 0.75 when correctly classifying the related headline and article pair into one of the three categories specified above. This section reviews literature for stance detection from two directions: (i) feature engineering with a neural network and (ii) a BERT-based model.

Baird et al.[14] ranked the first place in FNC-1 with a FNC score of 82.02%. The authors used an ensemble of gradient boosted decision trees with a deep convolutional neural network combined with similarity features. However, this approach is limited to detect the agree and disagree labels. Shang et al. [15] built an agreement-aware search engine named Maester to overcome the limitation. This search engine is built according to the relatedness on word occurrences and level of agreement between the headline and body text. Hanselowski et al. [20] and Riedel et al. [16] both incorporated various concatenations of input features (e.g., n-grams, bag-of-words, lexical and similarity features) which are fed into a multi-layer perception neural network. Their solutions achieved an FNC score of 81.97% and 81.72%, respectively. Thota et al. [17] also found that similarity features such as TF-IDF vectors (based on unigrams and bigrams) and cosines similarity (between news article and headline) are valuable input features. Later on, Bhatt et al.[18] proposed to combine three sets of features using a deep neural network, including neural embedding features from the deep recurrent model; statistical features from the weighted n-gram and bag-of-words; and handcrafted features. This solution outperformed all other state-of-the-art models and achieved a FNC score of 83.08%. In addition to the features mentioned above, Ghanem et al.[19] used a feature set of cue words and Google News word2vec embedding, to achieve a macro F1 score of 59.6%. Hanselowski et al.[21] proposed to add a Long Short-Term Memory (LSTM) network to their previous model [20], which outperformed all other state-of-the-art models with a macro F1 score of 60.9%. Chaudhry et al.[22] also found that using conditional encoding in an independent encoding model can boost performance. Mohtarami et al. [23] proposed using similarity matrices to feed into memory networks for text ambiguity prevention.

Previous works on stance detection are mostly based on features like n-grams, bag-of-words, TF-IDF vectors, similarities, POS (part of speech) and NER (name entity recognition). However, these features have a few limitations. For example, n-grams is simple to use, but only using this feature is not enough to catch fine-grained linguistic information that shows in the fake news writing style [74]. The sole use of POS is weak at catching additional information such as emotiveness implicated writings; an example could be words like angry, horrible [74]. Bhatt et al. [18] and Hanselowski et al. [21] suggested that semantic understanding is more robust to resolve the detection problem than relying on lexical overlap features. Thus, some researchers have applied BERT to stance detection for semantic understanding. Jwa et al. [24] proposed exBAKE, a model that classifies data using weighted cross-entropy (WCE) and added more pre-training data to the BERT model. Kaliyar et al. [25] proposed a model named FakeBERT that combines BERT with three parallel blocks of a one-dimensional convolutional neural network.

2.2.2 Source Credibility

Source credibility is to detect misinformation by assessing the credibility of the publisher or author of a given statement or an article. Agarwal et al. [75] analyzed several classifiers for fake news detection by assessing the source credibility and concluded that the support vector machine (SVM) and logistic regression classifier performed the best. Yuan et al. [76] introduced a novel network called structure-aware multi-head attention network (SMAN), which contains news body, and publishing, reposting relations of publishers and users. Baruah et al. [77] used the BERT model to classify whether the author of a Twitter feed is a fake news spreader or not. Yang et al. [78] proposed an unsupervised framework, built a probabilistic graph model and introduced the Gibbs sampling approach to estimate user credibility using user engagement behaviour. Ruchansky et al. [28] proposed a model called CSI, which is made up of three modules, Capture (responses to a given article), Score (a score assigned to its source), and Integrate (combination of responses, text and source information). The Score module extracts a representation and assigns a score to each user using a neural network and an implicit user graph [28]. The assigned score is computed based on the user's tendency to join a source promotion group. Sitaula et al. [29] proposed to find signals of news credibility using the information on the number of authors and the author's publication history. The study found that articles with no authors are most likely to be fake news, and authors who are engaged in trustworthy news are less likely to work with authors associated with fake news. Gupta et al. [26] proposed a semi-supervised ranking model that assigns a credibility score to tweets based on their credibility. The authors deployed a real-time system named TweetCred that is available as a browser plug-in to compute a credibility score of a source.

2.2.3 Rumour Classification

Rumour classification is to determine whether a post (statement) is a rumour or not using social user's information (e.g., user engagements, user activities, and/or the contents of user comments) from social media platforms. Yuan et al. [30] proposed a heterogeneous network with local and global attention. The authors incorporated local attention into constructing new representations of source tweets and corresponding retweets, and created a global heterogeneous network by combining these new representations of source tweets with structural and semantic properties. Wu et al. [32] proposed a graph-kernel-based hybrid SVM classifier to detect false rumours; this classifier captures high-order propagation patterns in addition to semantic features such as topics and sentiments. This approach, on average, can detect false rumours just 24 hours after the initial broadcast. Kwon et al. [33] analyzed temporal, structural and linguistic characteristics of rumours for rumour classification. The authors proposed a novel periodic external shocks (PES) model to identify the temporal characteristic that can describe the periodic bursts unique to rumours due to the daily cycle and the external shock cycle. To identify structural characteristics, the authors extracted properties related to the propagation process, such as the promotion of isolated rumour spreaders. To identify the linguistic characteristics, the authors examined the world-level categories and sentiments. Nguyen et al. [41] proposed a factual news graph (FANG) that models large social interactions. The social context for the proposed graph construction includes news articles, news sources, social users, social interactions, and stance detection. This graph learning framework captures social structure and engagement patterns effectively.

2.2.4 Fact Checking

There are many studies on fact checking. Existing fact checking approaches mostly use knowledge graphs or relevant documents as evidence to classify the validity of a claim. In this section we first review works that fact-check claims using knowledge graphs and relevant documents. Then, we review studies that use other techniques, such as few-shot learning via perplexity score [37] and table-based classification [36]. We also review studies that do not produce a definite label for a claim, such as determining the check-worthiness of a claim, detecting whether a claim has been previously fact-checked [79], or finding external evidence to help humans fact-check claims more efficiently [80, 81].

Knowledge graph. Ahmadi et al. [35] proposed a model to assess whether a claim is true or false with a human interpretable explanation based on knowledge graphs or web information. The knowledge graphs are from DBpedia; when the knowledge in the knowledge graphs is not enough to assess a claim's validity, web articles relevant to the input claims will be extracted from a commercial search engine to assess a claim's validity. Gad-Elrab et al. [39] proposed a system that provides explanations when inputting a query (e.g., Sadiq Khan is citizen of United Kingtom). The explanations are ranked according to confidence scores based on knowledge graphs or text. The knowledge graphs can be from Yago or Wikidata, and the textual resource can be from Wikipedia or web search engine Bing. Nguyen et al. [41] proposed a novel framework called FANG that uses graph representation to classify whether the news title is fake or real based on the news body.

Relevant documents. Nie et al. [40] proposed a BiLSTM classification model to label a claim as "supported", "refuted" or "not enough information" based on evidence. The evidence is extracted from wiki documents. The system first retrieves documents relevant to a claim from a collection of wiki documents. It then selects the sentences that are the most relevant to the claim as evidence, and produces a label for the claim based on this evidence. Augenstein et al. [38] constructed a dataset and proposed using BiLSTM to validate a claim based on the evidence extracted from the top 10 results provided by Google searches. The dataset is from 26 different fact-checking websites, and the labels vary depending on the fact-checking websites.

Other studies. Other studies use different methods, such as few-shot learning via perplexity score [37] and table-based classification [36]. Lee et al. [37] classified whether a claim is supported or unsupported by evidence based on an evidence-conditioned perplexity score using language model (i.e., BERT, RoBERTa, and XLNet). Chen et al. [36] proposed the TABLE-BERT and Latent Program Algorithm models, with input being a triple (Table, Statement, Label) to classify whether a statement is confirmed or refuted by the table.

Some models are used to assist human fact-checkers instead of producing a definite label, such as determining the check-worthiness of a claim, detecting whether a claim has been previously fact-checked or not [79], and finding external evidence to help human fact-checkers to fact-check more efficiently [80, 81]. Shaar et al. [79] proposed a BERT-based model to solve a ranking problem for binary classification. The authors first determined whether the claim is check-worthy or not, then verified whether the check-worthy claim has been verified in the past. If the claim has been previously verified, the model will rank the previously fact-checked claims based on the usefulness against the input claim. This model helps human fact-checkers to work more efficiently by reducing time to read non-relevant information. Given a document with statistical claims and related datasets, Karagiannis et al. [80] proposed a system that generates SQL queries to assist a user in validating a claim. In case of an incorrect statement, the system also suggests to update the statement.

Technically, our misinformation identification approach in this thesis is very similar to [38,40] since our objective is to classify the validity of a claim (e.g., true, partly true, or false) using relevant documents as evidence. The differences between our work and [38,40] are as follows.

- The authors in [40] developed their own neural semantic matching network for document and sentence retrieval, using a bidirectional long-term memory model for classification. In our work, we use TF-IDF and cosine similarity for information retrieval, and pre-trained transformer-based models (e.g., BERT, DistilBERT, RoBERTa, XLNet, ALBERT) for classification. These pre-trained models will allow other researchers to repeat or extend our experiments more quickly and easily.

- The dataset used in [38] contains 34,918 claims collected from 26 fact-checking websites. Each of these websites has its own labeling system. The claims on some website have seven labels, while some may have only two labels. Therefore, the authors treated the data from each website as a separate dataset. They performed classification on the 26 datasets and took the averages of the results from these datasets. Unlike the approach used in [38], the data we collected from different fact-checking websites were filtered and processed so that there are exactly three labels (true, false and partly true) for the whole dataset.

- Using the fact-checking approach, each claim is associated with a set of related articles that are used as the ground truth to classify the claim. The authors of [38,40] collected the related articles from wiki documents or Google searches. We collected data from fact-checking websites where, for each claim, the human fact checker(s) provided resources (articles) that justify the label they assigned to the claim. These related articles, hand picked by human fact checkers, are supposed to be much more relevant to the claim than those provided by Google or wiki document searches.

2.3 Datasets

In this section, we review datasets related to misinformation identification and COVID-19 information.

2.3.1 Datasets for Misinformation Identification

Existing datasets can be broadly divided into three categories based on the type of data a dataset contains: (i) claims only, (ii) news articles only, and (iii) social user information in addition to claims or articles.

Claims only. These datasets contain political statements, and fact-checked or constructed claims generated by modifying sentences from Wikipedia, and political statements. Most of the datasets contain general news, which covers a wide range of topics, such as politics, economics, sports, finance, entertainment and health. Datasets such as Twitter [82], Twitter15 [83], Twitter16 [83] and Cred-1 [84] are annotated by Snopes and cover many topics. LIAR [85] is annotated by PolitiFact and covers political news only. FEVER [86] contains constructed claims fact-checked by trained or experienced annotators. Datasets in [82][83][86] are for rumour classification tasks (classifying a given Tweet as rumour or non-rumour). Datasets in [84][86] are for fact extraction tasks (classifying a claim as true or false by assessing the credibility of the source).

Articles only. Most of these datasets contain either political news (e.g., FakevsSatire [87], BuzzFeedPolitical [88], and Political-1 [89]) or general news (e.g., NELA-GT-2018 [90]). However, most of the datasets are very small: FakevsSatire [87] contains 486 news articles; BuzzFeedPolitical [88], 120; and Political-1 [89], 225. NELA-GT-2018 [90] and NELA-GT-2019 [91] are both large-scale datasets, containing 713,534 and 1.12 million articles collected in 2018 and 2019 from 194 and 216 news outlets, respectively.

Social user information in addition to claims or articles. These datasets contain social user information such as hashtags, user mentions, sentiments, and URLs in addition to claims or articles. NELA-GT-2020 [92] contains 1.8 million news articles collected from 519 sources in 2020 with tweets embedded in the news articles. FakeNewsNet [93] consists of 1,270 and 22,153 political statements from PolitiFact and Gossipcop, respectively. Both contain news content, social context (i.e., user names, posts, responses and networks) and spatiotemporal information (e.g., user locations and timestamps of news articles). FakeHealth [94] was curated from healthNewsReview.org, which contains news reviews (ground truth labels and explanations to the label), news contents, social engagements (tweets, replies, retweets), and user network information (user profiles, user timelines, user followings and user followers). Both FakeNewsNet [93] and FakeHealth [94] datasets can be used for multi-modal misinformation identification since they contain information from different modalities (e.g., news content and social engagements).

Claims and articles. Some datasets contain both claims and articles for fact checking purposes. Chen et al. [36] constructed a large-scale dataset that consists of Wikipedia tables and statements that are labelled by humans manually. This dataset can be used for table-based fact checking [36]. On the other hand, Augenstein et al. [38] built a real-world multi-domain dataset for evidence-based fact checking of claims. This dataset consist of 34,918 claims collected from 26 fact checking websites. However, this dataset comprises 26 small datasets; each dataset has its own set of labels based on the fact checking website from which it was extracted.

2.3.2 Datasets Related to COVID-19 Information

Many COVID-19 related datasets have been curated recently for research against COVID-19 misinformation. Several of these datasets were collected from Twitter. The dataset CoAID [95] has a wide range of features, including claims, news content, and social media information (i.e., posts, tweets and replies). TweetsCOV19 [96] contains rich semantic annotations including extracted entities (e.g., coronavirus_disease_2019, Wuhan, social_distancing), hashtags, user mentions, sentiments and URLs posted between October 2019 and April 2020. There are also other Twitter datasets that contain COVID-19 related keywords and hashtags [97][98][99][100][101]. Some of the Twitter datasets are in languages other than English (e.g., Arabic, Chinese) [102][103][104][105], and some contain geolocations [104][100][101]. ReCOVery [106] is a dataset based on the approach of assessing the credibility of sources for the purpose of identifying misinformation. CORD-19 [107] is an open research dataset that is collected dynamically, and consists of at least 52,000 papers, including papers published in more than 3,200 journals. This is a valuable dataset to help researchers develop new solutions to fight COVID-19 misinformation. There are also datasets containing only claims, for example, the COVID-19 Fake News Dataset [108], which contains only social media posts from social media platforms (e.g., Facebook, Twitter, Instagram).

Most COVID-19 datasets collected from Twitter contain user activities, user networks, or user comments, which do not cover all features of interest. For example, some have justifications along with the claims but no news articles [85], some have only one news article associated with each claim [93], while some have only news articles but no claims [87][88][90]. There are no COVID-19 related datasets that can be used with the fact checking approach.

We introduce a new dataset named COVMIS to conduct research using the fact checking approach. Compared to the existing datasets, COVMIS is the only dataset that contains all the following features: (i) it is a large-scaled dataset in the domain of COVID-19 (containing more than 10,000 samples), (ii) there is a set of articles related to each claim, and (iii) there is an explanation of the labelling result to help researchers identify hidden problems to improve model performance using the analysis included in the explanation.

Chapter 3

COVMIS - A Dataset for Research on COVID-19 Misinformation

3.1 Introduction

In a global pandemic, misinformation can be as big a threat as the virus itself. According to the US National Institute of Health [11], *"(m)isleading information about treatment for COVID-19 has resulted in an increasing number of vitamin D abuse and even mass poisoning from methanol intake."*. In addition, vaccine hesitancy and resistance caused by misinformation has hampered the recovery process in many parts of the world [12]. Therefore, it is crucial to intensify research in misinformation identification, prevention, and mitigation as part of the global effort to fight the pandemic.

The dataset reported in this thesis has been built for misinformation identification using the automatic fact checking approach. Given a claim (statement), e.g., "You are 300-900 times more likely to die after getting the COVID-19 vaccine compared to the flu vaccine", we classify whether the claim is true, partly true or false using a collection of articles whose contents are related to the claim. The set of related articles, collected from reputable sources, serves as the ground truth to assess the validity of the claim. An example of a claim and its related articles is shown in Figure 5.1. This approach mimics the act of fact checking by humans. Given a statement, we would perform an Internet search (e.g., Google search), obtain several articles and web pages, read them and try to ascertain whether the statement is true or false based on the obtained information. This approach thus moves us one step closer to effective solutions for misinformation identification.

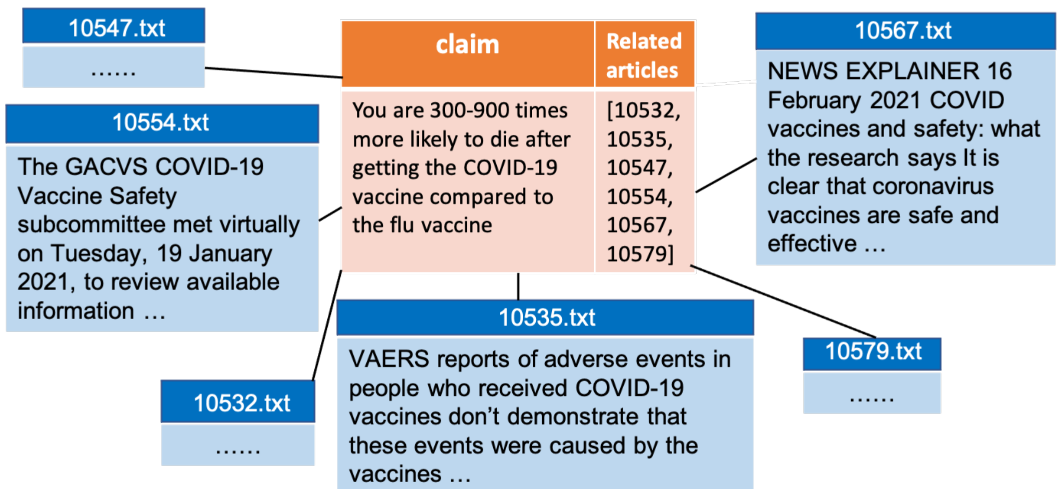


Figure 3.1: A claim and its related articles

We are interested in building a NLP-based system to identify misinformation in real time using the fact checking approach. Only one dataset is suitable for this purpose, published by the 2019 ”Leaders Prize: Fact or Fake News“ competition [45]. To conduct research on the fact checking approach with the ultimate goal of building an NLP-based system to identify misinformation in real time, we need more datasets of this type. This need motivated us to build a new dataset. The result is a large-scale dataset containing claims and articles on the topic of COVID-19 with 14,384 claims and 134,320 related articles. In this chapter, we present the methodology and process of data collection and statistics of the dataset.

We first searched for fact-checking websites and selected reputable sites according to three criteria: (i) whether they have high factual ratings from Media Bias/Fact Check (MBFC), (ii) whether they are a member of the International Fact-Checking Network (IFCN), and (iii) whether these websites are recommended by the Canadian Center for Occupational Health and Safety (CCOHS) as fact-checking tools. Based on these criteria, we identified nine reputable fact-checking websites (PolitiFact, Snopes, Agence France-Press, Africa Check, Polygraph.info, Reuters, The Washington Post, Poynter and Google Fact Check Tool) and four legitimate health-related websites (ScienceDaily, NIH, WebMD, and Medical News Today).

We collected claims and articles related to the claims from these websites. The claims were then manually labelled as true, partly true or false. The claims and their associated features (e.g., claimant, date, source) were formatted. Duplicate claims were removed from the dataset, and so were duplicate articles associated with a claim. The articles associated with a claim serve as the ground truth to assess the claim, so they must be relevant to the claim. Therefore, we applied the cosine similarity algorithm to remove articles that were not relevant to the claim with which they were associated. The result is a dataset that contains 14,384 claims and 134,320 articles, which we name COVMIS (from a combination of COVid and MISinformation).

The main contributions of this chapter are as follows:

- We constructed and released a large-scale dataset of COVID-19 news and facts for research on COVID-19 misinformation.
- We provide a comprehensive analysis of the dataset, including claim contents, article contents, publication dates, news sources, and country distribution, to better understand the data.
- We discuss potential use cases to demonstrate the benefits of the dataset for research on misinformation related to COVID-19 and other areas.

The remainder of the chapter is organized as follows. We first discuss the methodology and process of data collection in terms of news site selection, news content collection, news annotation and data preprocessing. Then we present statistics and an analysis of the dataset. Lastly, we discuss potential use cases for the dataset.

3.2 Data Collection

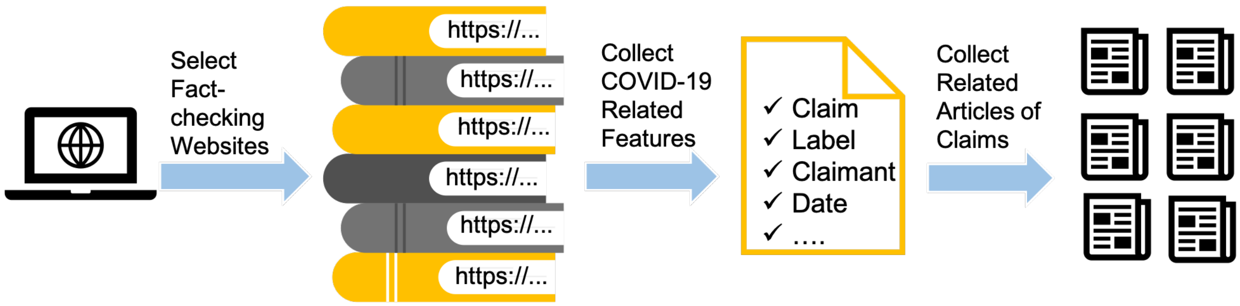


Figure 3.2: Data collection pipeline

In this section, we describe in detail how data was collected and processed in three stages: i) selection of news sites; ii) collection of news contents (e.g., claims, labels, news articles); iii) news annotation. We follow the pipeline as shown in Figure 3.2 for data collection and processing.

3.2.1 Selection of News Sites

We collect data from the following nine fact-checking websites and four health-related news sites:

Fact-checking Websites

- **PolitiFact** [109]: fact-checking the statements made by US politicians.
- **Snopes** [110]: known as a "well-regarded reference for sorting out myths and rumours", and fact-checking urban legends and similar stories in American pop culture.

- **Agence France-Presse** [111]: fact-checking claims posted on Facebook and downgrading information classified as"false" in the news feeds to reduce the number of people reading.
- **Africa Check** [112]: fact-checking claims made by public figures, institutions and media.
- **Polygraph.info** [113]: fact-checking quotes, stories, reports distributed by government officials, government-sponsored media and other high-profile individuals.
- **Reuters** [114]: fact-checking visual material and claims posted on social media.
- **The Washington Post** [115]: fact-checking political statements and stories.
- **Poynter** [50]: database which fact-checks claims in English and many other languages.
- **Google Fact Check Tool** [116]: same as *Poynter*.

The news sites that we selected must meet at least one of the following three criteria:

1. High factual rating from Media Bias/Fact Check (MBFC) [117]: MBFC assigns a factual rating to a news source ranging from "Very Low", "Low", "Mixed", "Mostly Factual", "High" to "Very High".
2. Member of International Fact-Checking Network (IFCN) [50]: IFCN promotes excellence in fact-checking worldwide. Members that belong to IFCN follow the code of principles to maintain consistency: "a commitment to non-partisanship and fairness, transparency of sources, transparency of funding & organization, transparency of methodology, open and honest corrections" [50].
3. Recommendation from the Canadian Centre for Occupational Health and Safety (CCOHS) [118]: CCOHS provides relevant tools and resources to support Canada's health, safety and wellness programs.

The details of the selected news sites are listed in Table 3.1.

Table 3.1: Fact-checking website descriptions

News Sites	MBFC Factual Rating	Member of IFCN	Recommended by CCOHS
PolitiFact	High	Yes	Yes
Snopes	High	Yes	Yes
Poynter	High	Yes	Yes
AFP Fact Check	High	Yes	-
Reuters	Very High	Yes	-
The Washington Post	Mostly Factual	Yes	-
Africa Check	Very High	Yes	-
Polygraph.info	High	-	Yes
Google Fact Check Tool	-	-	Yes

Legitimate Health-Related News

Most of the claims collected from the above fact-checking websites are labelled as false. In order to obtain a balanced dataset, we needed to add claims that can be labelled as true. To do this, we collected claims (statements) and articles from the following four news sites: ScienceDaily [119], National Institutes of Health (NIH) [120], WebMD [121], and Medical News Today [122]. **ScienceDaily** publishes the latest discoveries in health and science from leading scientific journals and universities. **NIH** is the primary medical research agency sponsored by the U.S. government. **WebMD** provides content regarding health and health care topics and is best known as a health information news site. **Medical News Today** features medical news and information to the public and physicians, and it is one of the fastest-growing health information sites in the US.

3.2.2 News Content Collection

Our news collection process was comprised of two major stages. The first was to collect claims and their features from the above-mentioned websites. The second was to collect the articles related to each claim. Algorithm 1 illustrates the process of extracting claims and their corresponding features and related articles.

Algorithm 1 News contents collection

Input A set of selected news sites $\mathcal{N} = \{n_1, n_2, \cdots, n_K\}$

Output return the extracted claim set $\mathcal{C}$, feature set $\mathcal{F}$ and related article set $\mathcal{A}$

procedure EXTRACT($\mathcal{C}, \mathcal{F}, \mathcal{A}$)
    for each news site $n_i \in \mathcal{N}$ **do**
        extract a set of claims $\vec{c_i} \subset C$ and a set of claim features $\vec{f_i} \subset \mathcal{F}$
    for each claim $c_j \in \mathcal{C}$ **do**
        if reference or source set $\vec{r_j} \neq \emptyset$ **then**
            $\vec{a_j} \leftarrow \vec{r_j}$ ▷ Extract references or sources as related articles of c_j
        else if embedded URLs set $\vec{e_j} \neq \emptyset$ **then**
            $\vec{a_j} \leftarrow \vec{e_j}$ ▷ Extract content from embedded URLs as related articles of c_j

Collecting Claim Features

We collected claims along with their features, including:

- ID: a unique ID of each claim.
- claim: the statement.
- claimant: the person or entity that made the statement.
- date: the date when the claim was made.
- label: truth label of the claim (true, partly true or false).

- source: the fact-checking website from which the claim was collected.

For claims collected from the Poynter website, we collected two more features in addition to the above:

- country: the country where the claim originated.

- explanation: explanation of the labelling result.

Collecting Related Articles

On the fact-checking websites, each claim (or statement) is accompanied with explanations made by journalists justifying the ground truth assigned to the claim. The explanations can be followed by a list of references or sources used for the justification; one example is shown in Figure 3.3. We followed the URLs listed in the reference section and downloaded the contents of the pages indicated by the URLs, and use the downloaded contents as articles related to the claim. In many cases, the URLs of related articles are not listed in a separate section but embedded in the explanation text, as shown in Figure 3.4. We followed the embedded links to download articles related to the claim.

3.2.3 News Annotation

All the claims are annotated manually by human experts on the fact-checking websites that we used. Each fact-checking website has its own rating system and criteria for labelling. For example, the website *The Washington Post* uses "geppetto checkmark" and "four pinocchios" to indicate that a claim is true or false, respectively. Therefore, we converted website-specific labels into three equivalent categories: true, partly true or false. Not all labels can be mapped into one of the above three categories, e.g., "missing context", "satire", or "uncertain". The claims associated with these labels were thus excluded from the dataset. A list of label mappings is shown in Table 3.2.

Data cleaning was applied to the raw data collected and is discussed next in section 3.2.4.

Table 3.2: Label mapping

Labels in the Dataset	Labels from the News Sites
true	“true”, “vrai”, “correct”, “geppetto checkmark”, “checked”, “accurate”
partly true	“partly true”, “half-true”, “mostly true”, “mixed”, “unclear and partially true”, “spinning the facts”, “misrepresentation”, “three Pinocchio’s”, “mixture”, “not exactly”, “not the whole story”, “not quite right”, “twists the facts”, “questionable”, “partially false and misleading”, “understated/exaggerated”, “disputed data”, “mixture of true and false information”, “mostly correct”, “likely false”, “misleading”, “mostly false”, “misleading headline”, “unlikely”, “half-truth”, “partly false claim”, “unclear”, “misleading due to missing context”
false	“pants on fire”, “false”, “faux”, “fake”, “four Pinocchio’s”, “not true”, “wrong”, “fake news”, “fiction”, “legend”, “scam”, “probably false”, “incorrect”, “hoax”, “false headline”, “distorts the facts”, “inaccurate”, “false headline claim”
Excluded from the dataset	“missing context”, “lacks context”, “needs context”, “altered”, “unsubstantiated”, “misattributed”, “miscaptioned”, “cherry-picked”, “lacks evidence”, “mislabeled”, “satire”, “uncertain”, “flip-flop”, “manipulated photo”, “false comparison”

Our Sources

Wisconsin Assembly GOP, Wisconsin Legislature Takes Gov. Evers to Court, April 21, 2020

Institute for Health Metrics and Evaluation, COVID-19 projections for Wisconsin, updated April 17, 2020

Milwaukee Journal Sentinel, GOP lawmakers ask Supreme Court to block Tony Evers' order to stay home, April 21, 2020

STAT, Influential Covid-19 model uses flawed methods and shouldn't guide U.S. policies, critics say, April 17, 2020

Milwaukee Journal Sentinel, Tracking coronavirus in Wisconsin, accessed April 21, 2020

Wisconsin Hospital Association, COVID-19 Situational Awareness Update, April 21, 2020

Email exchange with Kit Beyer, spokeswoman for Robin Vos, April 21, 2020

Figure 3.3: Example of a list of references

On February 25, the day Brazil recorded the highest number of deaths from COVID-19 to date – close to 1,600 – President Jair Bolsonaro said that "studies have started to appear" about the harmful effects mask wearing has on children.

"A German university says that [masks] are harmful to children, and show several aspects like irritability, headache, difficulty concentrating, decreased perception of happiness, refusal to go to school or daycare, discouragement, impaired learning ability, dizziness, fatigue," he said.

Bolsonaro, who has repeatedly minimized the risks of COVID-19 and derided protective masks, added that he would "not go into details because everything flows into criticism over me, and I have my opinion about masks, and each one has his own. But we are waiting for a more in-depth study on this by competent people."

Bolsonaro's statements are misleading.

Although he did not name the German study, Bolsonaro was clearly referring to a report titled "Corona children studies 'Co-Ki': First results of a Germany-wide registry on mouth and nose covering (mask) in children" first published in December and updated this month, examining 26,000 children's reactions to wearing masks in the country. The study is based on results posted to an online registry; it was not peer reviewed.

Figure 3.4: Example of an explanation with embedded URLs

3.2.4 Data Preprocessing

We applied a three-step data cleaning process to the collected raw data. We first formatted the data into a JSON file. The data format is shown in Table 3.3. In the second step, we removed all duplicate claims with the same label and kept all claims with different labels even though the claims' contents were the same. In the last step, for each claim, we computed the cosine similarity between the claim and each of its related articles. If the cosine similarity is less than 1×10^{-6}, that article is removed from the dataset.

For the non-English articles collected from the Poynter website, we first translated them into English using Google Translate API and then applied the three-step data cleaning described above.

Table 3.3: Metadata sample

ID	Claim	Claimant	Date	Label	Source	Related Articles	Counrty	Explanation
76	You are 300-900 times more likely to die after getting the COVID-19 vaccine compared to the flu vaccine.	Alex Berenson	2021-02-15	false	Science Feedback	[10532,10535, 10547,10554, 10567,10579]	United States	Pfizer-BioNTech and Moderna COVID-19 vaccines demonstrated a high level of safety and efficacy during clinical trials in order to receive emergency use authorization from the U.S. Food and Drug Administration...

3.3 Data Descriptions and Statistics

This section provides a description and statistics of the dataset.

3.3.1 Data Descriptions

COVMIS contains 14,384 claims and 134,320 articles related to the claims. The class distribution is 70.6% false, 15.4% partly true and 14.0% true, as illustrated in Figure 3.5.

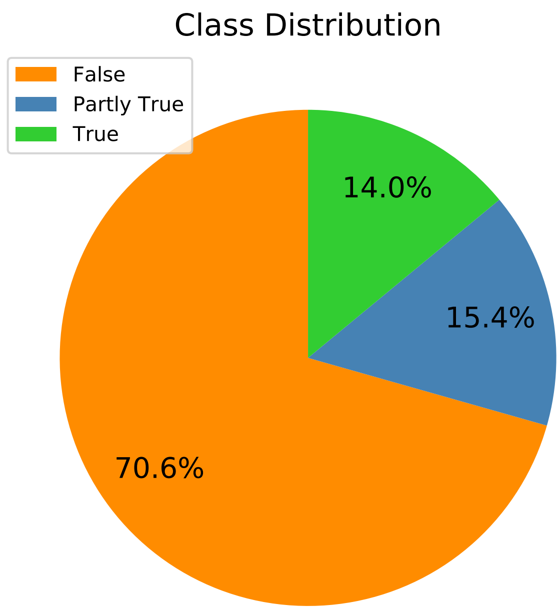


Figure 3.5: Class distribution

3.3.2 Data Statistics

This section provides visualization for different components of the dataset.

Claim Contents

Figure 3.6 shows the word cloud of the claims. As we collected only claims that are related to COVID-19, the most commonly used words are "COVID", "coronavirus", "covid vaccine", "time", "people", "pandemic", "hospital", "death". The font size for each word is scaled to its appearance frequency in the dataset. Figure 3.7 shows the distribution of character counts of claims. We observed that the number of characters per claim concentrates between 20 and 300, with the highest number of claims having 75 characters per claim and an average of 101 characters per claim.

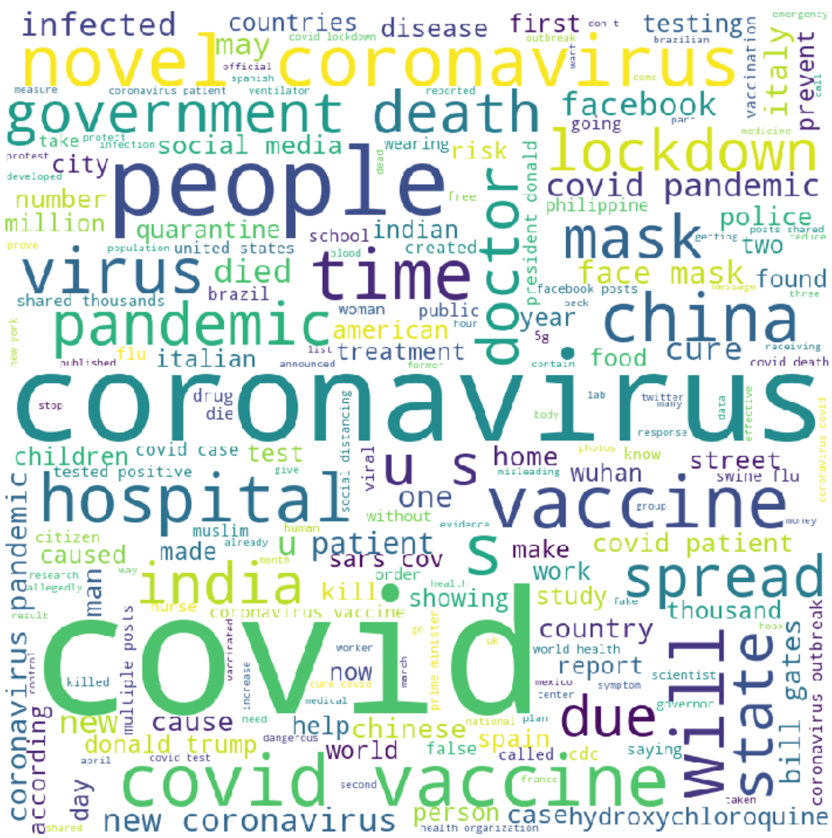


Figure 3.6: Most commonly used words in the dataset

Article Contents

The range of word counts of the articles is broad with the longest article having 1,042,420 words. The average number of words per article is 6,764. Therefore, we only show the distribution of the word count for articles with fewer than 15,000 words. As we can see from Figure 3.8, the histogram is right-skewed, indicating the majority of the articles contain 6,000 words or fewer.

Publication Dates

Figure 3.9 illustrates the distribution of claims according to their publication dates. As we can see, the graph has two waves, and this maybe related to the first two waves of COVID-

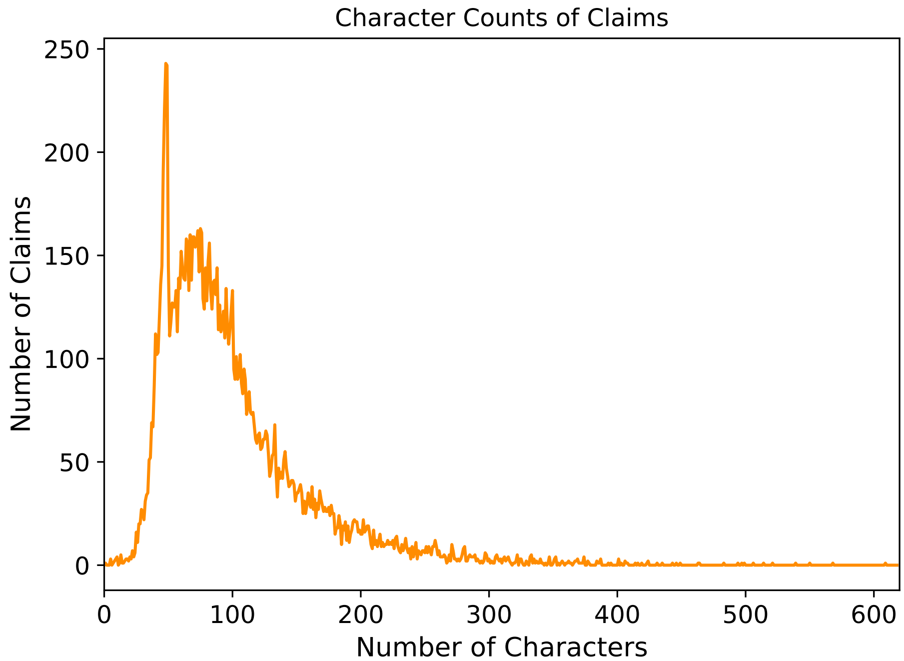


Figure 3.7: Distribution of character counts of claims

19 infection between February 2020 and February 2021. The first wave happened between January 2020 and June 2020 when not much was known about coronavirus; hence a large volume of misinformation circulating on the Internet and social media platforms.

Distribution of News Sources

There are a total of 100 news sources. Figure 3.10 shows the top 25 news sources, which have each contributed more than 100 claims to the dataset.

Country Distribution

As mentioned above, the claims collected from the Poynter website were fact-checked by journalists across different countries, and thus, the articles are in English and many other languages. Figure 3.11 shows the country distribution with the circle sizes scaled to the number of claims. There is a total of 143 different countries. As we observed from the

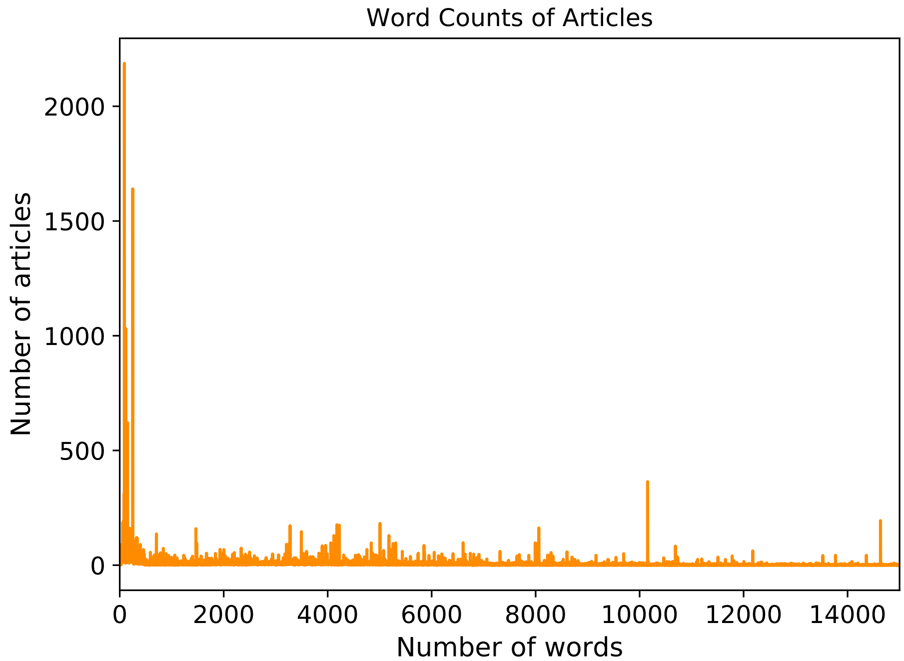


Figure 3.8: Distribution of word counts of articles

diagram, the majority of claims came from India, United States, Spain, Brazil, Colombia, France, the Philippines and Italy. The commonality across these countries is the seriousness of the pandemic and high numbers of COVID-19 cases in those countries. People tend to spread misinformation when they feel anxiety toward uncertainties [123].

3.4 Potential Use Cases

We aim to provide a useful dataset that helps the research community to fight against misinformation. This large-scale dataset can be used for various applications.

3.4.1 Ground Truth Data for NLP-based Fact Checking Models

The automatic fact checking approach is the closest to the act of fact checking by humans and the most effective for truth labelling. However, to the best of our knowledge, using

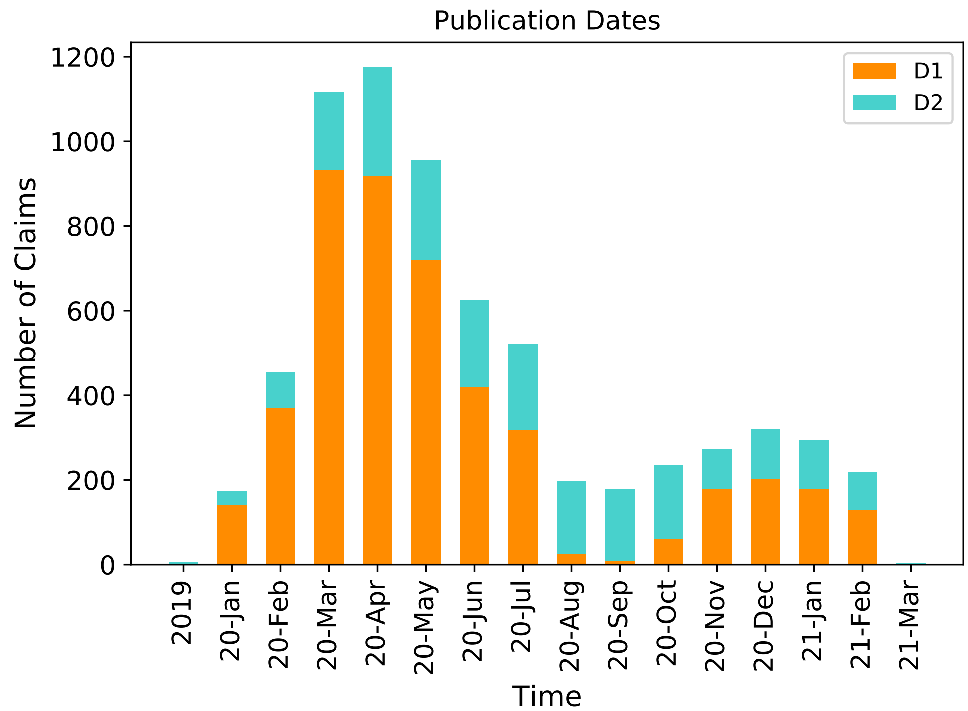


Figure 3.9: Distribution of publication dates

related articles from the source section or embedded links of fact-checking websites for the fact checking approach has not been studied in literature. Its first use was in the 2019 "Leaders Prize: Fact or Fake News?" competition [45]. COVMIS is the second dataset constructed using the same methodology as the competition mentioned above. We are using COVMIS to study and fine-tune a fact checking model based on BERT [1]. We anticipate that the dataset will be useful for developing future fact checking models.

3.4.2 Data Collection Methodology and Process

Although the fact checking approach is the most effective for truth labelling, there was only one dataset [45] built using related articles from the source section or embedded links of fact-checking websites, and COVMIS is the second of such datasets. This paper is the first that presents the methodology and process for data collection, formatting and pre-processing

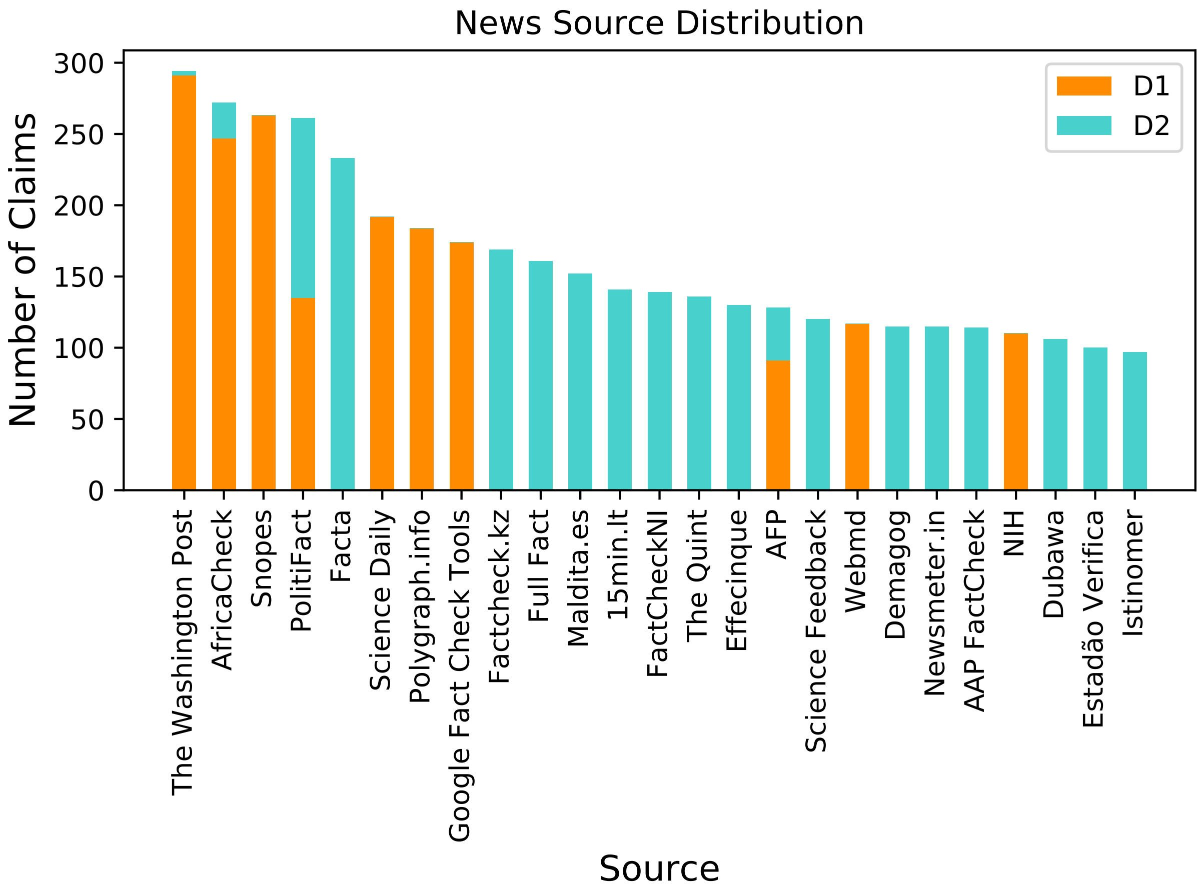


Figure 3.10: Distribution of news sources

to textbfconstruct data for the fact checking approach. This data collection approach can be used to build more datasets of this type, in many other domains such as US politics, stock markets, sports, entertainment, and environment. This will enable more comprehensive and in-depth research in misinformation identification using the fact checking approach and other approaches, including stance detection.

3.4.3 Explainable Misinformation Identification

Most research on misinformation identification has focused on improving the performance of machine learning models. Few works address the ability to explain the classification results

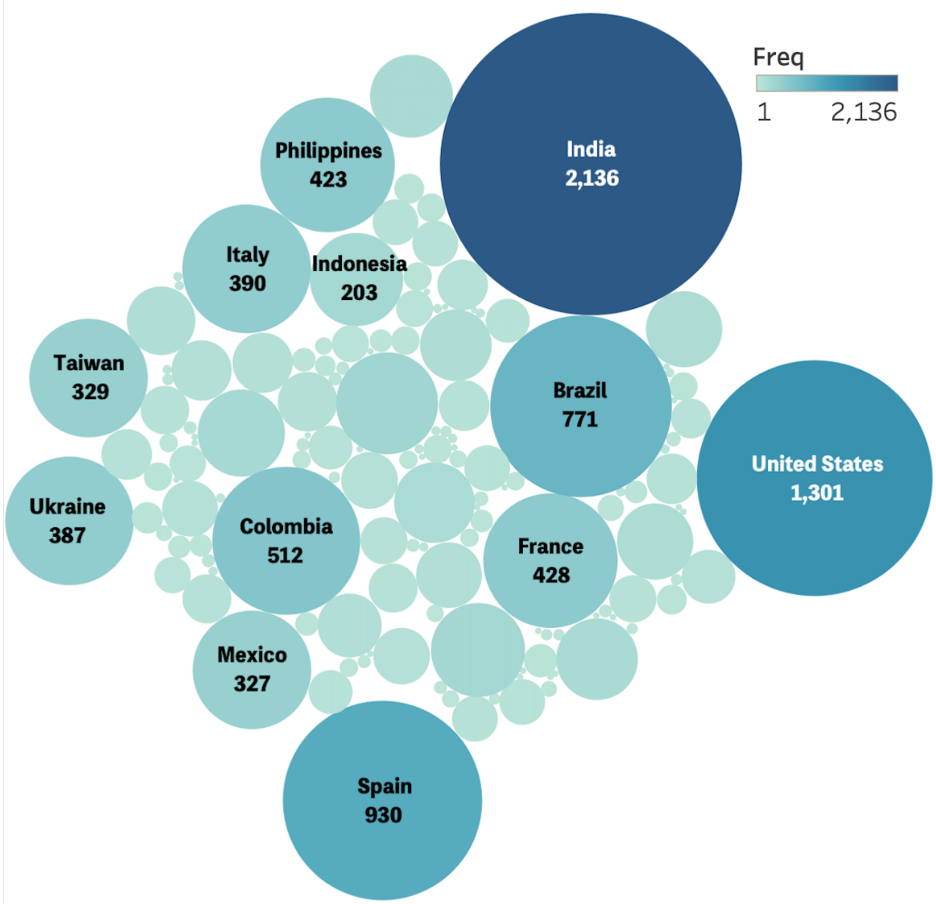


Figure 3.11: Country distribution

of a machine learning model. Explainable machine learning allows users to understand, interpret and appropriately trust the machine's predictions. It also enables researchers to debug and improve the performance of a model. Explanations of machine learning models and predictions improve transparency and users' trust in the system. Part of the dataset, which we collected from the Poynter website, can be used to train and test an explainable misinformation identification model. For each claim in the Poynter subset of data (which contains 10,943 claims and 47,390 articles), there is an explanation associated with the claim, justifying the truth label assigned to it.

3.4.4 Sentiment Analysis

Sentiment analysis is a process that automates the mining of attitudes, opinions, views and emotions from text, speech, tweets and database sources through natural language processing (NLP) [124]. One example is to classify a statement as "positive", "negative", or "neutral" [124]. Sentiment analysis has a broad range of applications, such as recommender systems, business intelligence [125][126][127][128], and analysis of financial news [129]. The claims from COVMIS can be used as an additional corpus for training unsupervised classification models for sentiment analysis such as [130][131][132][133].

3.4.5 Multi-language Misinformation Identification

Most of existing NLP-based models for misinformation identification deal with datasets in English, either originally written in or translated to English. However, misinformation is a global problem, and identifying misinformation in languages other than English is just as important. The non-English articles we collected from the Poynter website can be used for this purpose. For example, it can be used to study textual features of different languages for the task of fake news identification [134][135], or to identify fake news in other languages such as Russian [136] or Portugese[137].

3.4.6 Text Summarization

Text summarization refers to the task of generating a summary of an article or multiple articles on the same topic [138][139][140]. Automatic summarization allows a large volume of text to be condensed to a short summary. Automatic text summarization has a wide range of applications for enterprises and organizations to mitigate the problem of information overload, providing summaries of social media monitoring, financial research, social media marketing, video scripting, and patent research. Given a large number of articles in the COVMIS dataset, it can be used as a corpus for text summarization using unsupervised learning approaches [141][142].

3.4.7 Machine Translation

Machine translation is automated translation from one language to another by a computer. This task can be handled using various approaches, for example, supervised learning (where a parallel corpus is required), unsupervised learning (where only monolingual data is required), or semi-supervised (where both a parallel corpus and monolingual data is required). COVMIS can be helpful for all the approaches mentioned above. For example, both English and corresponding articles in foreign languages in the data set can be used to train supervised learning models [143] and semi-supervised learning models [144]. Articles in foreign languages can serve as an additional training corpus for news translation tasks using unsupervised learning approaches [145][146][147][148].

3.5 Chapter Summary

In this chapter, we presented a large-scale, publicly available dataset for research on COVID-19 misinformation. COVMIS contains 14,384 claims, 134,320 related articles and many features associated with the claims such as claimant, source, date and explanation. We provided statistics and a detailed analysis of the dataset, and discussed a variety of its potential use cases. COVMIS supports NLP-based misinformation identification that relies on fact checking for truth labelling.

Chapter 4

Evaluate Different State-of-the-Art NLP Models

4.1 Problem Definition

In this chapter, we evaluate the performance of several state-of-the-art NLP models, namely, BERT, DistilBERT, XLNet ALBERT and RoBERTa. We conduct our research using the automatic fact-checking approach discussed in Section 2.2.4. For example, given a claim, "NASA has just confirmed earth has a new moon" [45] and a set of articles related to the claim, we classify whether the claim is true, partly true or false. The related articles were collected from reputable sources and used as the ground truth to assess the claim's validity. We use the COVMIS dataset that we constructed and described in Chapter 3 for the experiments. COVMIS consists of 14,384 claims and 134,320 articles; each claim is associated with a set of articles. This dataset contains only COVID-19 related information.

We use transformer-based NLP models to predict the validity of a claim using its set of related articles as evidence. In this study, we evaluate the performance of BERT, DistilBERT, XLNet, RoBERTa and ALBERT models for the task of fake news classification.

For each claim to be classified, we combine the set of articles related to the claim into one large document, then input this document into the models. Yet any of the five models can accept only 512 tokens as input. As a result, the set of articles related to a claim will be truncated once the model has input 512 tokens. However, the truncated data can be important for classification. Therefore, we use an information retrieval algorithm to extract from the set of articles related to a claim sentences that are the most relevant to the claim. The K most relevant sentences are then input into the model instead of whole articles related to the claim. We name this improvement the MRS (most relevant sentences) method, and the former way of inputting whole articles into a model the WA (whole articles) method.

The remainder of the chapter is structured as follows. We provide a detailed explanation of the methods used in the experiments and the architecture of models in section 4.2. We introduce the experiment settings in section 4.3. Then, we present the results and discussion in section 4.4. And lastly, we summarize the chapter in section 4.5.

4.2 Methods

In this study, we first evaluate the performance of BERT, DistilBERT, XLNet, RoBERTa and ALBERT models using the WA method. The WA method combines the set of related articles of each claim into one large document and inputs this document into the models. Then, we re-evaluate these five models using the MRS method. In this method, we use an information retrieval algorithm to retrieve K sentences that are the most relevant to a claim from the set of related articles. Firstly, we combine all the related articles of each claim into a large document. Secondly, we convert the document to TF-IDF vector representations; then, we sort the sentences based on the cosine similarity of each sentence against the claim. Lastly, we input K sentences that have the highest similarity scores into a model.

In this section, we explain how we use TF-IDF and cosine similarity for information retrieval. Also, we provide an overview of the state-of-the-art models used in the experiments: BERT, DistilBERT, XLNet, RoBERTa, and ALBERT. Figure 4.1 illustrates the steps of the experiments.

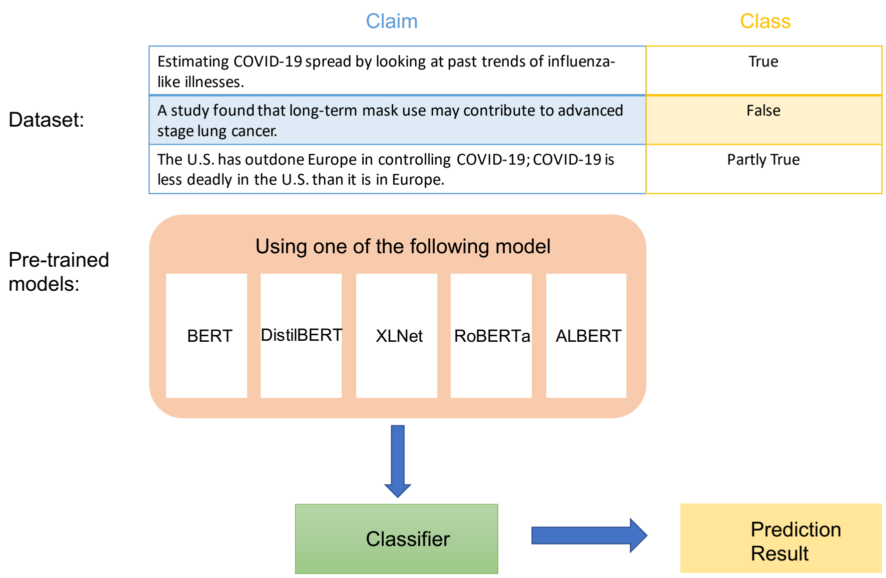


Figure 4.1: Algorithm workflow

4.2.1 Information Retrieval Algorithm

In this section, we describe the proposed information retrieval approach based on TF-IDF.

TF-IDF [149] is an algorithm used to transform a piece of text into a number representation. It is also an effective statistical method that quantifies the importance of a word to a document. TF-IDF consists of two parts: (1) term frequency (TF) that indicates how frequently a word appears in a document; (2) inverse document frequency (IDF) that aims to reduce the weight of commonly used words in a dataset and increase the weight of rarely used words.

In this work, we apply the TF-IDF algorithm to measure the importance of a word in a sentence. Let c denote a claim and S denote the set of all sentences in all the articles related to claim c. We define D as $D = S \cup \{c\}$. We use a simple example with three sentences to describe the TF-IDF algorithm.

- Sentence s_1 (claim c): The ocean is blue.
- Sentence s_2: The sky is blue.

- Sentence s_3: What colour is the ocean?

Sentence s_1 is claim c. Sentences s_2 and s_3 form set S, the set of all sentences in all articles related to claim c. D denotes the set of all three sentences.

The TF-IDF computation for this example is shown in Table 4.1. We first extract all distinct words from D, which are listed in the first column of Table 4.1. Then, we calculate the term frequency and inverse document frequency of each word.

word	TF			IDF	TFIDF		
	S1 (c)	S2	S3		S1 (c)	S2	S3
the	1/4	1/4	1/5	log(3/3)=0	0	0	0
ocean	1/4	0/4	1/5	log(3/2)=0.18	0.045	0	0.036
is	1/4	1/4	1/5	log(3/3)=0	0	0	0
blue	1/4	1/4	0/5	log(3/2)=0.18	0.045	0.045	0
sky	0/4	1/4	0/5	log(3/1)=0.48	0	0.12	0
what	0/4	0/4	1/5	log(3/1)=0.48	0	0	0.096
color	0/4	0/4	1/5	log(3/1)=0.48	0	0	0.096

Table 4.1: TF-IDF computation for the above example

The term frequency of a word w in a sentence s_i, denoted by $f(w, s_i)$ is the number of times w appears in the sentence divided by the length of the sentence. For example, the word "ocean" appears once in sentence s_3 whose length is five words. Therefore the term frequency of "ocean" in s_3 is $f(\text{"}ocean\text{"}, s_3) = 1/5$.

The inverse document frequency of a word w in set D, denote by $f(w, D)$, is computed as follows:

$$f(w, D) = log\frac{|D|}{|s_i \in D, w \in s_i|}$$

where $|D|$ is the cardinality of set D and $|s_i \in D, w \in s_i|$ indicates the number of sentences in set D in which word w appears. For example, the word "ocean" appears in two sentences s_1 and s_3 in set D. Thus the IDF of w in set D, whose cardinality is 3, is $f(\text{"}ocean\text{"}, D) = log(3/(2)) = 0.18$.

The IDF metric measures how common a word is in a set of sentences (documents).

Commonly used words such as “the”,“be”, “to”, “of”, “and”, etc. should be given lower weights in terms of similarity scores than less commonly used words.

The TF-IDF score of a word w in a sentence s_i is computed as $f(w, s_i) \times f(w, D)$. For example, the TF-IDF score of the word “ocean” in sentence s_3 is $f(\text{“}ocean\text{”}, s_3) \times f(\text{“}ocean\text{”}, D) = 1/5 \times 0.18 = 0.036$.

In Table 4.1, the last three columns show the TF-IDF score of every word in each sentence s_1, s_2 and s_3. The TF-IDF scores in each column forms the TF-IDF vector representation of the corresponding sentence. Table 4.2 shows the TF-IDF vector representations of s_1, s_2 and s_3, each vector corresponding to a column in Table 4.1. We use these obtained TF-IDF vector representations of s_1, s_2 and s_3 to compute the similarity score between claim c (sentence s_1) and each other sentence.

Cosine similarity is an effective and widely used metric in information retrieval. It is used to compare documents with respect to a given vector of words. In the thesis, cosine similarity is used to compute similarity scores between a claim c and each sentence in the set of articles related to the claim (i.e., each sentence in set S). The cosine similarity computation of our example is shown in Table 4.2.

Sentence ID	Sentence	Vector Representation	Cosine Similarity with S1
S1	The ocean is blue.	[0, 0.045, 0, 0.045, 0, 0, 0]	1
S2	The sky is blue.	[0, 0, 0, 0.045, 0.12, 0, 0]	0.248
S3	What color is the ocean?	[0, 0.036, 0, 0, 0, 0.096, 0.096]	0.181

Table 4.2: Cosine similarity computation for the above example

The cosine similarity between a claim c and a sentence $s_i \in S$, denoted by CS(c, s_i), is computed as follows:

$$\mathbf{CS}(c, s_i) = \frac{\vec{c} \cdot \vec{s_i}}{||\vec{c}|| \times ||\vec{s_i}||} \tag{4.1}$$

where $\vec{c}$ and $\vec{s_i}$ represent the TF-IDF vector representation of claim c and sentence s_i, respectively; $\vec{c} \cdot \vec{s_i}$ is the dot product of vector $\vec{c}$ and $\vec{s_i}$; and $||\vec{v}||$ is the Euclidean norm of vector $\vec{v} = (v_1, v_2, ..., v_n)$, a.k.a. the length of vector $\vec{v}$, which is computed as $||\vec{v}|| = \sqrt{v_1^2 + v_2^2 + ... + v_n^2}$, where n is the number of dimensions of the vector.

We use the above example to compute the cosine similarity between sentence s_1 (claim c) and sentence s_3. Using the TF-IDF vector representations of claim c and sentence s_3 given in Table 4.2, we have

$$\mathbf{CS}(s_1, s_3) = \frac{[0, 0.045, 0, 0.045, 0, 0, 0] \cdot [0, 0.036, 0, 0, 0, 0.096, 0.096]}{\sqrt{0.045^2 + 0.045^2} \times \sqrt{0.036^2 + 0.096^2 + 0.096^2}} = 0.181. \tag{4.2}$$

The last column of Table 4.2 shows the cosine similarity between sentence s_1 (claim c) and each sentence in set D. Note that the cosine similarity between sentence s_1 and itself is 1.

After computing the cosine similarity between a claim c and each sentence in the set of articles related to the claim, we sort the similarity scores and select the sentences with the highest scores to input into the models. The number of selected sentences was varied from 5 to 50 in the experiments described in Section 4.3.4 and 4.4.2. These sentences have the strongest similarity to the claim and are considered the most relevant for the task of classifying the claim.

4.2.2 Pre-trained Natural Language Models

In this section, we describe the transformer-based language models to be evaluated.

BERT

BERT [1] stands for Bidirectional Encoder Representation from Transformers, known as Masked Language Model (MLM). The pre-trained BERT model takes the entire sentence(s) as input, then tokenizes the sentence, masks out 15% of the input words, and then predicts the masked words. The [MASK] token does not appear during fine-turning to prevent a mismatch between pre-training and fine-tuning. The input words are replaced with [MASK] 80% of the time replaced with a random word for 10% of the time, and kept the same for the remaining 10%. A binarized next sentence prediction task can be generated for downstream tasks based on the understanding between sentences.

The pre-trained model contains data from Wikipedia (2,500M words) and BookCorpus (800M words), totalling 16 gigabytes. The batch size is 131,072 words; the training time is 1 Million steps; the optimizer is Adam with 10^4 learning rate and linear decay. There are two types of BERT models; one is the base version, the other is the large version. There are 12 layers, 768 hidden units per layer, 12 self-attention heads, and 110 million parameters for the base version model. And there are 24 layers, 1024 hidden units per layer, 16 self-attention heads, and 340 million parameters for the large version.

BERT model can be fine-tuned inexpensively for many NLP tasks, one of which is sequence classification, which is the function used for this study. It adds a linear layer on top of the pooled output to perform text classification to classify the validity of a claim. Figure 4.2 illustrates the architecture of the BERT model.

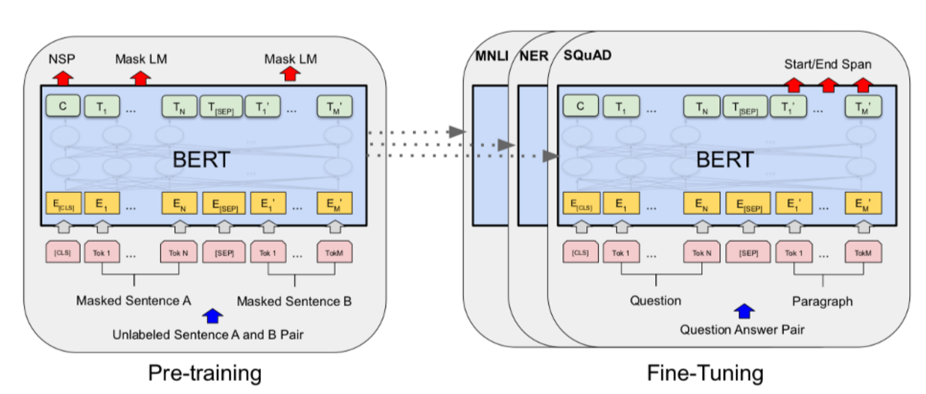


Figure 4.2: BERT Model (Source: courtesy of [1])

DistilBERT

DistilBERT [46] is a distilled version of BERT. DistilBERT is the same as the BERT model in terms of general architecture but with three differences: (1) the removal of token-type embeddings, (2) the removal of pooler, and (3) the reduction of the number of layers by two. DistilBERT is distilled on very large batches leveraging gradient accumulation using dynamic masking and without the next sentence prediction objective. DistilBERT is trained on the same corpus as the BERT model. Sanh et al. [46] show that the DistilBERT model reduces the size of a BERT model by 40% while retaining 97% of its language understanding capabilities and being 60% faster.

XLNet

XLNet [47] is a generalized auto-regressive model for natural language understanding. XLNet is a large bidirectional transformer that uses nearly ten times more data than the BERT model. XLNet is trained with a batch size eight times larger for half as many optimization steps, which ends up with four times as many sequences in pre-training compared to the BERT model. XLNet uses permutation language modelling in comparison to the masked language model that BERT uses. Permutation language modelling means that all tokens are predicted in random order, which is different from traditional language models in which all tokens were predicted in sequential order instead. The random ordering for token prediction helps the model learn bidirectional relationships, therefore, better handle dependencies and relations between words.

RoBERTa

RoBERTa [48] stands for Robustly Optimized BERT Approach. RoBERTa is a modified model based on the BERT model. Liu et al. [48] found that BERT was significantly undertrained. Therefore, they proposed the RoBERTa model to match or the performance of other BERT-based models. The modification includes:

- Training model with larger mini-batches and learning rates with a longer time.
- Removal of the next sentence pre-training objective.
- Training on longer sequences.
- Dynamically changing the masking pattern applied to the training data.

Training with larger batches improves perplexity for masked language modelling objectives and the accuracy of the end-ask. The reason is that large batches can run in parallel using the disturbed data-parallel training approach. In [48], RoBERTa outperforms BERT because RoBERTa was trained without the next sentence prediction loss and with blocks of text from a single document, while BERT incurred the next sentence prediction loss and used text from single or multiple documents. For data, RoBERTa uses five English-language corpora: BookCorpus, Wikipedia, CC-News, OpenWebText, and Stories. The combined size of all five corpora is 160 gigabytes under uncompressed conditions. The data for the BERT model is only 16 gigabytes. For masking, BERT uses a static mask, and the training data was duplicated ten times and masked in ten different ways over 40 epochs of training to avoid using the same mask for each training instance in every epoch. RoBERTa, on the other hand, uses dynamic masking, it generates a masking pattern whenever there is a sequence fed to the model, which reduces time when pre-training for more steps and massive datasets. The pre-trained steps are also increased from 100,000 to 300,000, and 500,000 compared to BERT, which shows a significant improvement in downstream task performance [48]. Furthermore, 300,000 and 500,000 pre-train steps models outperform the XLNet large model across most tasks [48].

ALBERT

ALBERT [49] stands for A Lite BERT. This model is aimed to solve memory limitations on available hardware and the problem of long training time in BERT. The ALBERT model scales the pre-trained BERT model by lowering memory consumption and increasing BERT's training speed based on two-parameter reduction techniques. The first technique is to factorize embedding parameterization. This technique decomposes a large vocabulary embedding matrix into two mall matrices, then separates the hidden layer's size from the size of vocabulary embedding. The second technique is cross-layer parameter sharing. This technique prevents the size of parameters from growing with the network's depth. The configuration of the ALBERT model is identical to the large version of the BERT model but with 18 times fewer parameters and 1.7 times lesser training time. [49].

4.3 Experiment Setting

In this section, we provide an overview of the dataset, parameters and evaluation metrics used for the experiments.

4.3.1 Dataset Overview

We use the COVMIS dataset described in Chapter 3 for the experiments. This dataset consists of 14,384 claims and 134,320 articles related to the claims. Each claim is associated with a set of related articles that were collected from reputable sources. The number of characters per claim concentrates between 20 and 300, with the highest number of claims having 75 characters per claim and an average of 101 characters per claim. Each claim is associated with at least one article. The average number of words per article is 6,764, and the majority of the articles contain 6,000 words or fewer. Each claim can be categorized into one of the three categories: true, partly true or false. The class distribution is 14.0% true, 15.4% partly true and 70.6% false.

4.3.2 Parameters

We use the base version of BERT, DistilBERT, XLNet and RoBERTa, and version one of ALBERT [49] pre-trained models to run our experiments. The batch size, learning rates and activation function for each experiment are the same throughout the experiments, and details are shown in Table 4.3. The activation function GELU [150], which stands for Gaussian Error Linear Unit, is a neuron activation function based on a Gaussian function. These models were trained and tested on an Nvidia GTX 2080Ti GPU on a local machine. For all the experiments, we ran ten trials each using different testing sets and took the average of the ten runs.

Parameters	BERT	DistilBERT	XLNet	RoBERTa	ALBERT
Batch Size	8	8	6	8	8
Learning Rate	4×10^{-5}	4×10^{-5}	4×10^{-5}	4×10^{-5}	4×10^{-5}
Activation Function	GELU	GELU	GELU	GELU	GELU

Table 4.3: Parameter setting

4.3.3 Evaluation Metrics

We use accuracy, precision, recall and F1 score as the metrics to evaluate the performance of the misinformation identification models. We define the following variables:

- TP: true positive, the number of claims being correctly predicted as positive out of the total actual positives.
- TN: true negative, the number of claims being correctly predicted as negative out of the total actual negatives.
- FP: false positive, the number of claims being wrongly predicted as positive out of the total actual negatives.
- FN: false negative, the number of claims being wrongly predicted as negative out of the total actual positives.

The definitions of the performance metrics are as follows:

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

$$Precision = \frac{|TP|}{|TP| + |FP|}$$

$$Recall = \frac{|TP|}{|TP| + |FN|}$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Accuracy indicates that the number of correctly predicted samples. However, "predictive models with a given level of accuracy may have greater predictive power than models with high accuracy" [151]. This kind of situation is called the accuracy paradox, especially when the classifier has an imbalanced class distribution. For example, a class contains 95% of true labels and 5% of false labels, and the classifier achieves 95% accuracy by classifying all labels as true. However, this classifier has no predictive power to predict false labels, which says that a classifier is not necessarily a robust model even if it has high accuracy. Therefore, only considering accuracy as the only evaluation metric for a classifier is not enough. We added precision, recall and F1 score as evaluation metrics in the experiments. Precision indicates that the number of samples being predicted as positive is labelled positive, recall indicates the number of positive samples is predicted correctly, and F1 score is the harmonic mean of precision and recall.

One may wonder which metric is more important, and the answer depends on the application. For example, accidentally giving a person access to top-tier confidential information that the person does not have permission to do can be crucial since confidential information in tier I is sensitive. This example is a false positive case in a classifier. A high precision rate is required to lower the false positive rate to minimize the above risk. Another example is when someone has permission to access some general information, but he cannot do so. This example increases labour cost to provide permissions to authenticated users manually. As this is a false negative case in a classifier, a high recall is required to minimize the false case negative rate. To summarize, a high precision rate is needed when a high false positive rate is problematic, and a high recall rate is needed when a high false negative rate is problematic.

As described above, the importance of the metrics depends on the topic. For example, during the early pandemic stage, a statement claimed a shortage in toilet paper. This statement caused many people to stock up toilet paper even though they do not need it just yet. In this case, no matter if this statement has been classified as a false positive or a false negative, a shortage of toilet paper does not cause a considerable impact. However, what if one claimed that there is a shortage in face masks? If this statement is classified as false positive, and people believe that the statement is valid, they will start to rush into the clinic or store and start purchasing all the face masks. This act will harm health care workers because they will not have enough protective equipment when serving society. In the thesis, we are studying misinformation in COVID-19; therefore, recall and precision are both essential.

Since our dataset is highly unbalanced, especially there is only 14.3% of the data is labelled as true, we consider all four metrics (accuracy, precision, recall and F1 score) to evaluate the performance of our misinformation identification models.

4.3.4 List of Experiments

We conducted two sets of experiments using transformer-based NLP models (i.e., BERT, DistilBERT, XLNet, ALBERT, and RoBERTa). Set I uses the WA method, which inputs whole articles into the models, and may truncate articles to accept only 512 tokens per claim. Set II uses the MRS method, which inputs only K sentences that are the most relevant to a claim into a model. In each experiment, 90% of the data is used for training, and the remaining 10%, for testing.

- Set I was designed to evaluate the performance of different state-of-the-art models by feeding the whole articles into the model. We conducted five experiments in this set, each of which corresponds to one of the five models mentioned above. The purpose is to find out the best-performing model for misinformation identification.

- Set II was designed to evaluate the performance of different state-of-the-art models by feeding only K sentences that are the most relevant to a claim into a model. We conducted five sets of experiments, each of which corresponds to one of the five models. Each set contains eight sub-experiments. We extracted K sentences that are most relevant to the claim, where K = 5, 10, 17, 25, 30, 35, 40, 50. There are eight different values of K, hence, eight sub-experiments for each set. After we have the results from Set I and Set II, we then compare the results of Set I against those of Set II. Set I feeds the whole articles into the model, whereas Set II feeds only top K relevant sentences to the claim into the model. The purpose is to determine whether feeding short but relevant data into the model improves the performance.

4.4 Results and Discussion

4.4.1 Set I - Performance of the Models Using the WA Method

In this set of experiments, we evaluated the performance of state-of-the-art models using the WA method, which inputs the whole articles into the models. Table 4.4 shows the results which are illustrated by the chart in Figure 4.3. As shown in Figure 4.3, BERT has the best performance among the five models in terms of accuracy, F1 score, precision, and recall with a score of 0.8336, 0.6746, 0.7390 and 0.6699, respectively, followed by DistilBERT, ALBERT, RoBERTa, and XLNet. Among the five models, there is a difference of 0.37 to 1.64 percentage points in terms of accuracy (0.8336 vs. 0.8172 vs. 0.8202 vs. 0.8299 vs. 0.8290, respectively). The performance difference ranges from 0.3 to 4.6 percentage points in terms of F1 score (0.6746 vs. 0.6287 vs. 0.6319 vs. 0.6593 vs. 0.6487, respectively); 0.9 to 10.3 percentage points in terms of precision (0.7390 vs. 0.6766 vs. 0.7057 vs. 0.7709 vs.

0.7796, respectively); and 0.5 to 3.0 percentage points in terms of recall (0.6699 vs. 0.6452 vs. 0.6402 vs. 0.6586 vs. 0.6489, respectively). Other than the accuracy, the performance difference among the models in terms of F1 score, precision, and recall can be significant.

	BERT	XLNet	RoBERTa	DistilBERT	ALBERT
Accuracy	0.833633	0.817249	0.8201518	0.8299411	0.8289866
F1 score	0.6745598	0.6286821	0.631931	0.6593045	0.6487258
Precision	0.7390233	0.6765532	0.7057496	0.7709005	0.7796106
Recall	0.6698787	0.6451857	0.6402173	0.6586035	0.6488558

Table 4.4: Performance of the NLP models using the WA method

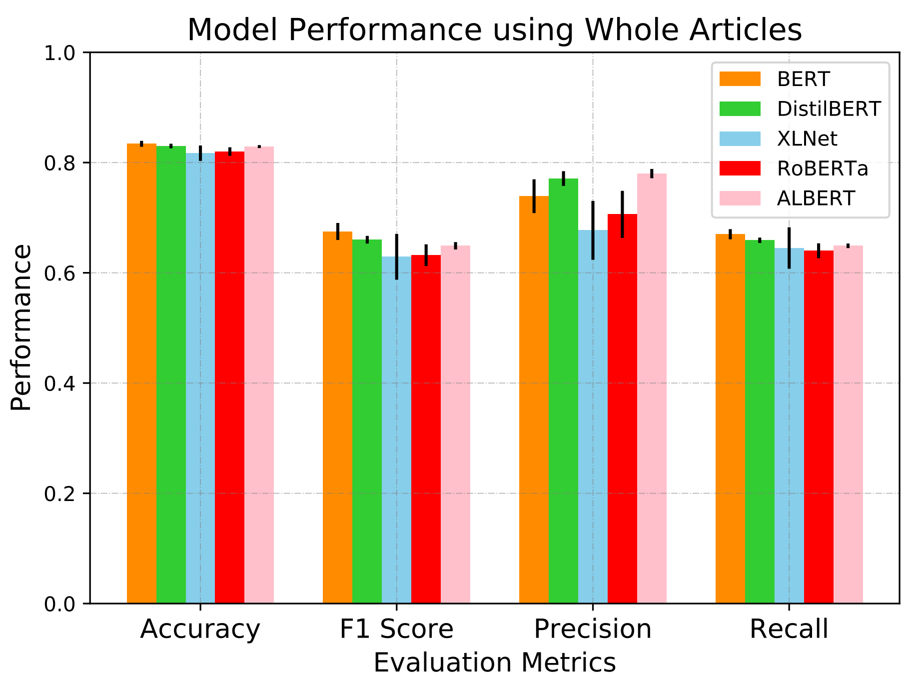


Figure 4.3: Performance of the NLP models using the WA method

4.4.2 Set II – Performance of the Models Using the MRS Method

In this set of experiments, we improved the performance of the models using the MRS method instead of the WA method. Using the MRS method, we input K sentences that are the most relevant to a claim into a model, instead of inputting whole articles. We retrieve K sentences that are the most relevant to a claim using the information retrieval algorithm that we discussed in Section 4.2.1. First, we applied the TF-IDF algorithm to obtain a vector representation of each claim and relevant sentences from the corresponding articles. Next we applied the cosine similarity algorithm to obtain the similarity score between the claim and each sentence of the articles related to the claim. We then selected K sentences with the highest similarity scores. Finally, we input these top K sentences into a model in the order from the highest score to the lowest score.

We varied the number of the most relevant sentences, K, from 5 to 50, and investigated how the value of K affects the performance of the models in terms of accuracy, F1 score, precision and recall. The graphs in Figure 4.4 illustrate the results. We observed that there is no pattern indicating which K value performs the best, for all five models. Therefore, we will use the result given by K = 5 for the analysis below. Figure 4.5 shows the performance of the five models in terms of accuracy, F1 sore, precision and recall when K = 5.

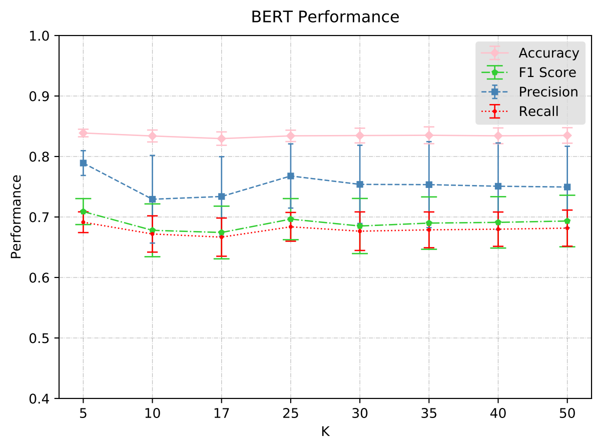
The performance differences between using the WA method and the MRS method (K = 5) are shown in Figure 4.6. The graphs show an improvement in accuracy, F1 score, precision, and recall using the MRS method. The accuracy of the MRS method for BERT, XLNet, RoBERTa, DistilBERT, and ALBERT increases by 0.5 (0.834 vs. 0.839), 1.9 (0.817 vs. 0.836), 0.7 (0.82 vs. 0.827), 0.6 (0.83 vs. 0.836), and 0.2 (0.829 vs. 0.831) percentage points, respectively. The F1 score for BERT, XLNet, RoBERTa, DistilBERT, and ALBERT increases by 3.4 (0.675 vs. 0.709), 6.5 (0.629 vs. 0.694), 3 (0.632 vs. 0.662), 4.2 (0.659 vs. 0.701) and 2.5 (0.649 vs. 0.674) percentage points, respectively.

The precision increases by (0.739 vs. 0789), 8.1 (0.677 vs. 0.758), 2.5 (0.706 vs. 0.731), 2 (0.771 vs. 0.791), and 0.5 (0.78 vs. 0.785) percentage points, respectively. The recall increases by 2.1 (0.67 vs. 0.691), 3.9 (0.645 vs. 0.684), 1.8 (0.64 vs. 0.658), 2.6 (0.659 vs. 0.685), and 3.2 (0.649 vs. 0.681) percentage points, respectively.

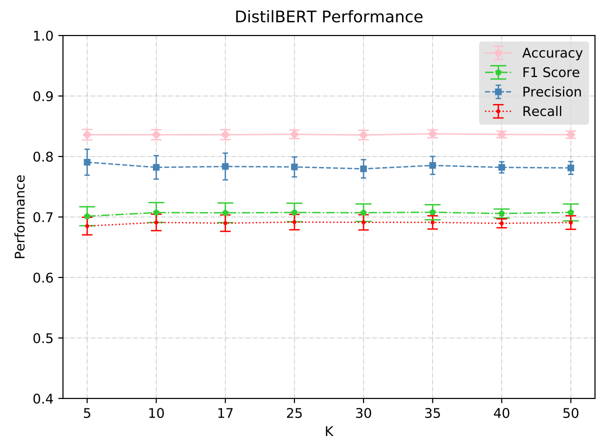
In summary, the MRS method offers noticeably improved performance over the WA method. In particular, the F1 score can be improved by up to 6.5 percentage points, and the precision, by up to 8.1 percentage points in our experiments. Using the WA method, the models truncate the input articles related to a claim and keep only the first 512 tokens, many of which may not be relevant to the claim. The MRS method, on the other hand, extracts from the articles the sentences that are the most relevant to the claim to input into the models. As a result, it offers better classification performance than the WA method.

4.5 Chapter Summary

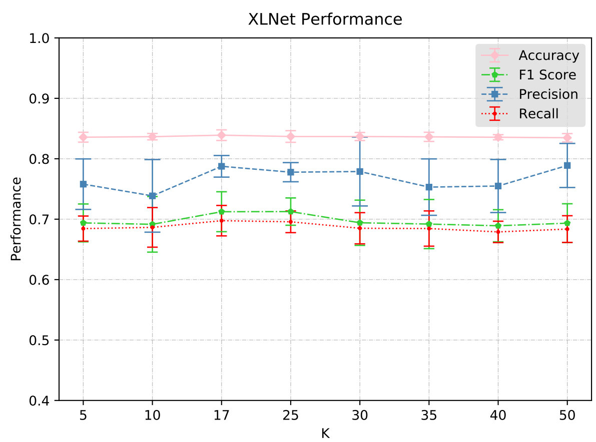
In this chapter, we evaluated the performance of the states-of-the-arts transformer-based models using two methods: (1) the WA method by inputting the whole articles into the models, and (2) the MRS method by inputting K sentences that are the most relevant to a claim into a model. Our experimental results show that the MRS method noticeably improves the classification performance of the models compared to the WA method.



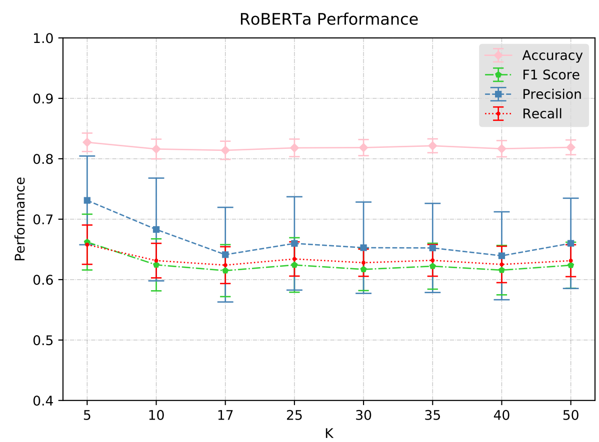
(a) BERT Model Performance



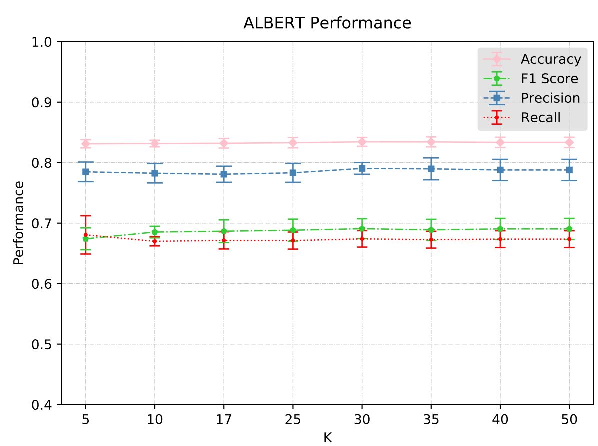
(b) DistilBERT Model Performance



(c) XLNet Model Performance



(d) RoBERTa Model Performance



(e) ALBERT Model Performance

Figure 4.4: Performance of the NLP models using the MRS method

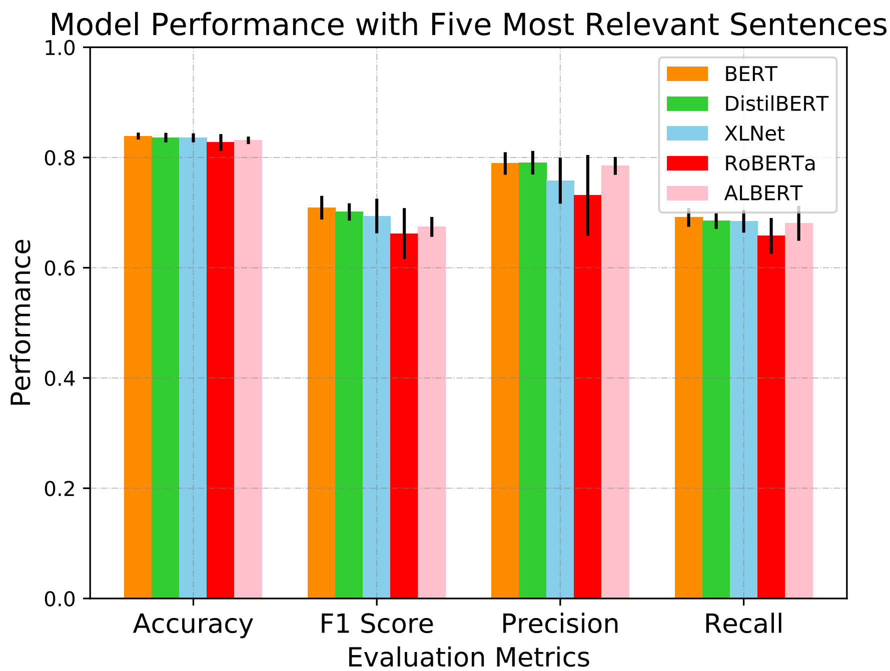


Figure 4.5: Model performance with five most relevant sentences

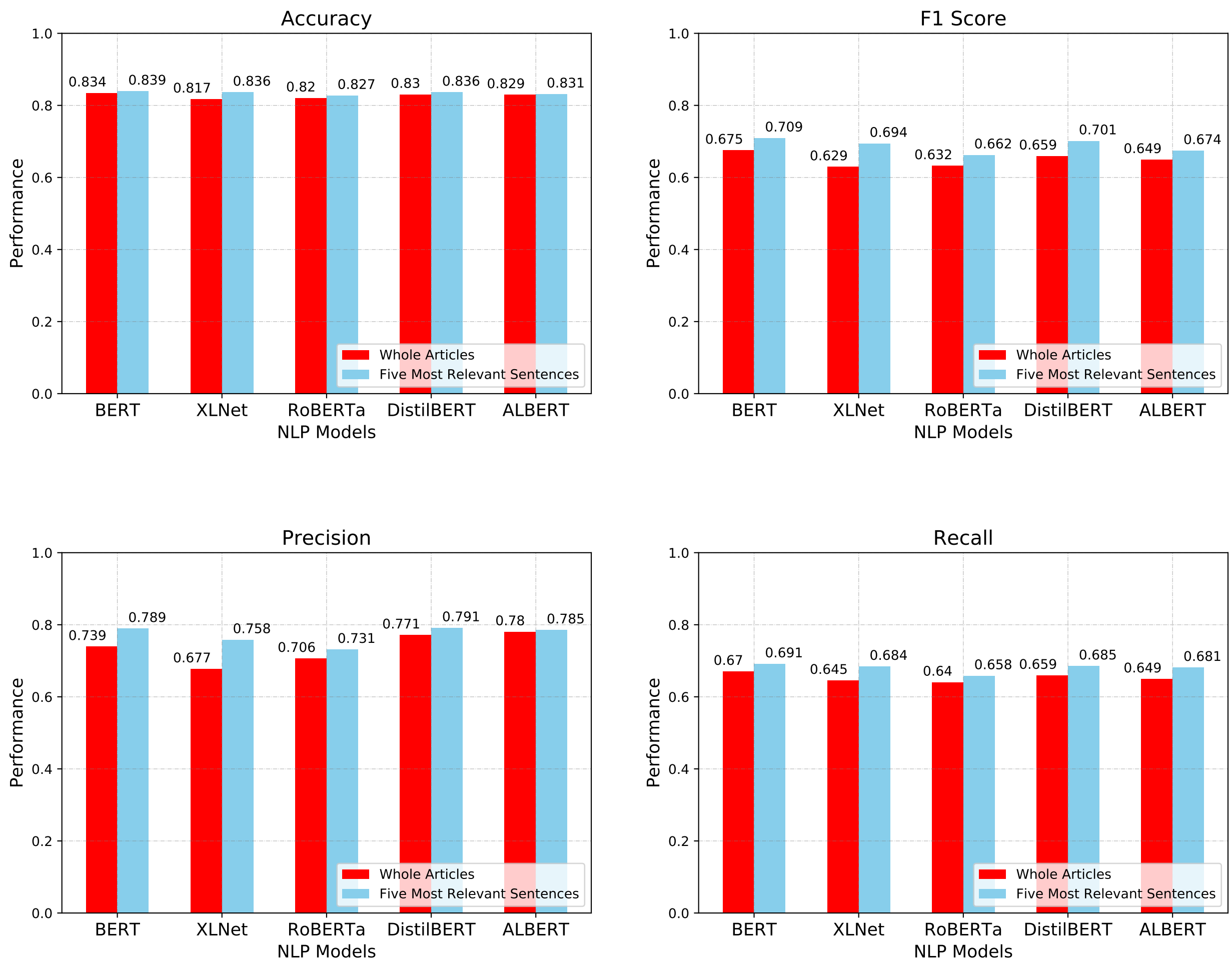


Figure 4.6: Performance of the WA method vs. the MRS method (K = 5)

Chapter 5

How Can Data Affect the Performance of NLP-based Misinformation Identification Models

5.1 Problem Definition

Misinformation can be identified based on news contents (fact checking), the credibility of the source of information, and patterns of information propagation on the Internet or in social networks [10]. The study of misinformation identification in this chapter is based on news contents and, furthermore, on natural language processing (NLP). Given a claim (statement), for example, "NASA has just confirmed earth has a new moon" [45], our model classifies the claim as true, partly true or false using a set of news articles whose contents are related to the claim. The set of related articles, collected from reputable sources, serves as the ground truth to assess the validity of the claim.

We started our research with the dataset provided by the "2019 Leaders Prize: Fact or Fake News? competition" [45]. This dataset contains 15,555 claims, each associated with a set of related articles, for a combined total of 64,974 articles. The claims and articles were published before November 2019 and cover many topics such as politics, finance, entertainment, sports and health care. (We name this set "general news dataset".) During a project that aimed to improve the performance of a NLP-based misinformation identification model, we came up with the following two research questions:

RQ.3a **Can the performance (e.g., accuracy, precision, recall, F1 score) of such a model be improved if the domain of knowledge is narrowed down to a specific area of interest?** We can define specific areas of interest such as the 2020 US election, the stock market, climate change, COVID-19, and curate datasets on each specific topic. A claim on the topic of climate change would then be evaluated using a model trained with a dataset containing only articles on climate change. We hypothesized that this method could result in higher performance than one that uses a general news dataset to evaluate a claim. As an analogy, consider a panel of medical experts taking questions from an audience. An oncologist would answer questions about cancer treatments better than a general practitioner.

RQ.3b **How do obsolete training or testing data affect the performance of such a model?** A general news dataset collected before November 2019 would not contain any information about the coronavirus pandemic. Therefore, if a COVID-19 related claim is evaluated using a model trained with the general news dataset mentioned above, the result may not be as good because the general news dataset is obsolete compared with the COVID-19 claim.

To answer the above questions, we conducted experiments on a news classification model that is based on Bidirectional Encoder Representation from Transformers (BERT), and evaluated the performance of the model in terms of accuracy, precision, recall and F1 score. We used two datasets: (i) the general news dataset mentioned above, and (ii) the COVMIS dataset containing claims and news articles on the topic of COVID-19 described in Chapter 3.

To answer RQ.3a, we compare the experimental results obtained from the general news dataset with those from the COVID-19 dataset. The results show that the performance of the model obtained from the COVID-19 dataset are significantly higher than those obtained from the general news dataset, confirming our hypothesis.

To answer RQ.3b, we conducted several experiments using either obsolete training data or obsolete testing data. Obsolete data refer to those extracted from the general news dataset. The objective is to quantify how obsolete training data affect teh classification performance. For example, the training data was from the general news dataset and the claims to be tested were from the COVID-19 dataset, or vice versa. The experimental results show that a temporal mismatch between the training and testing data significantly degrades the performance of the model. The results highlight the importance of keeping datasets used for misinformation identification up-to-date in order for the model to work effectively.

The remainder of the chapter structures as follows. Section 5.2 introduces the experiment setting, section 5.3 presents the experiment results and discussions. Finally, section 5.4 summarizes the chapter.

5.2 Experiment Setting

This section describes the datasets, performance metrics and NLP model used for experiments.

5.2.1 Datasets

We use two datasets in the experiments:

1. a general news dataset provided by "Leaders Prize: Fact or Fake News? competition" [45]. It covers a wide range of topics, such as politics, sports, finance, and health care.

2. COVMIS that consists of claims and news articles related to only the pandemic and COVID-19.

Each dataset has two parts: (i) a set of claims (statements) and (ii) multiple sets of articles, each associated with a claim. The set of articles related to a claim, collected from reputable sources, serves as the ground truth to assess the validity of the claim. Figure 5.1 illustrates an example of a claim with its related articles. Each claim is labelled as true, false or partly true, and associated with the following information: a unique ID (ID), the person or the source that made the claim (claimant), the date the claim was made (date), truth labelling of the claim (label), and the IDs of the related articles of the claim (related articles). Table 5.1 shows a sample of a claim and its associated information. Table 5.2 shows the statistics of the two datasets mentioned above, namely the total number of claims, total number of articles, the range of numbers of articles related to a claim, and the class distribution.

Table 5.1: Metadata of a claim

ID	Claim	Claimant	Date	Label	Related articles
76	You are 300-900 times more likely to die after getting the COVID-19 vaccine compared to the flu vaccine	Alex Berenson	2021-02-15	False	[10532,10535,10547, 10554,10567,10579]

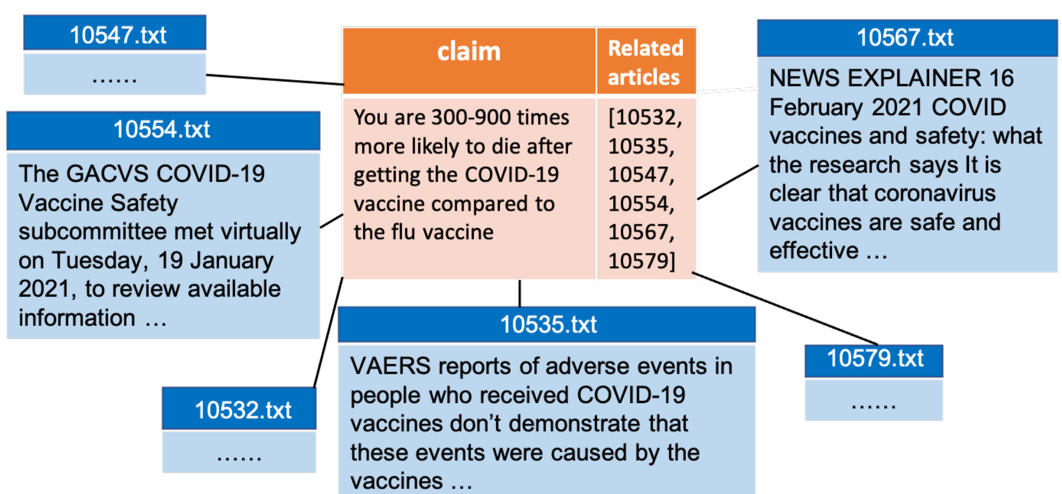


Figure 5.1: Sample of a claim and its related articles

Table 5.2: Dataset statistics. The three class labels are true (T), partly true (PT), and false (F)

	No. of claims	No. of articles	No. of related articles	Class distribution (T : PT : F)
General News Dataset [45]	15,555	64.974	2 - 66	11% : 41% : 48%
COVID-19 Dataset [?]	13,657	134,326	1 - 158	14% : 15.4% : 70.6%

5.2.2 Evaluation Metrics

We use accuracy, precision, recall and F1 score as the metrics to evaluate the performance of the misinformation identification model. The definitions of the metrics are provided in Section 4.3.3.

5.2.3 NLP Model Used

We use BERT [1] for the experiments in this study. This pre-trained model contains up to 16GB of vocabulary collected from Wikipedia and BookCorpus. There are two versions of BERT: the base version and the large version. Our study uses the base version, which contains 12 layers, 768 hidden units per layer, 12 self-attention heads, and 110 million parameters. In comparison, the large version has more encoding layers and parameters. For our purpose, the performance of the base version measured by the above metrics is not much different from that of the large version. The base version, however, runs faster than the large version and is simpler to set up. Therefore, we used the base version for this study. Formally, given a set of claims $C = \{c_1, c_2, \cdots, c_m\}$ and their related articles $A = \{\vec{a}_1, \vec{a}_2, \cdots, \vec{a}_m\}$, the output of BERT is $H = \{h_1, h_2, \cdots, h_m\}$. A softmax classifier is added on top of BERT to predict the probability of label $l \in \{0, 1, 2\}$: $p(l|h) = \text{softmax}(W \cdot h + b)$, where W is a matrix of trainable parameters, and the labels 0, 1 and 2 denote false, partly true and true, respectively. All the parameters from BERT as well as W are jointly fine-tuned by maximizing the log-probability of the correct label for each claim.

Note that there is a limitation of the BERT model, as mentioned in Chapter 4. The model can accept only 512 tokens as input. As a result, the set of articles related to a claim will be truncated once the model has input 512 tokens. However, the truncated data can be important for classification. Therefore, we use an information retrieval algorithm to extract from the set of articles related to a claim sentences that are the most relevant to the claim. The K most relevant sentences are then input into the model instead of whole articles. The procedure for extracting the K most relevant sentences from a set of related articles is described in Section 4.2.1. For the experiments in this chapter, we use K = 5. (An experiment in Chapter 4, Set II, shows no classification performance difference given by different values of K.)

5.3 Experiment Results and Discussions

We conducted two sets of experiments to answer the two research questions: RQ.3a and RQ.3b.

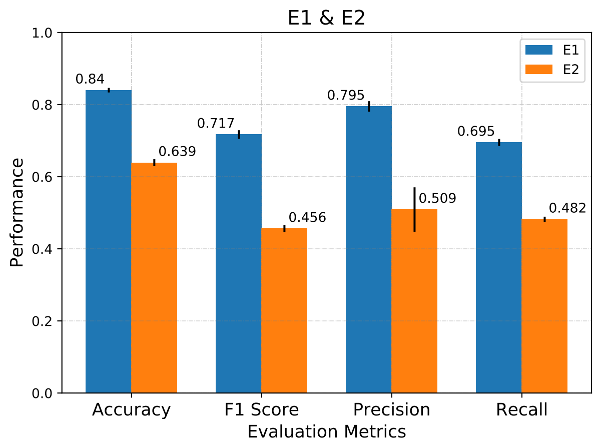
Set I was designed to answer the question of whether narrowing down the domain of knowledge improves the performance of the classification model. We performed two experiments in this set:

E1. We trained and tested the model using the COVID-19 dataset, which contains specialized domain knowledge related to COVID-19.

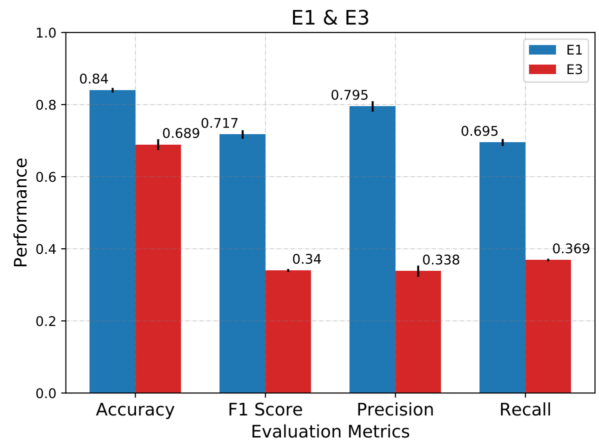
E2. We trained and tested the model using the general news dataset.

The purpose is to see if the COVID-19 dataset can give better performance than the general news dataset. In each experiment, 90% data of a set was used for training the model, and the remaining 10%, for testing.

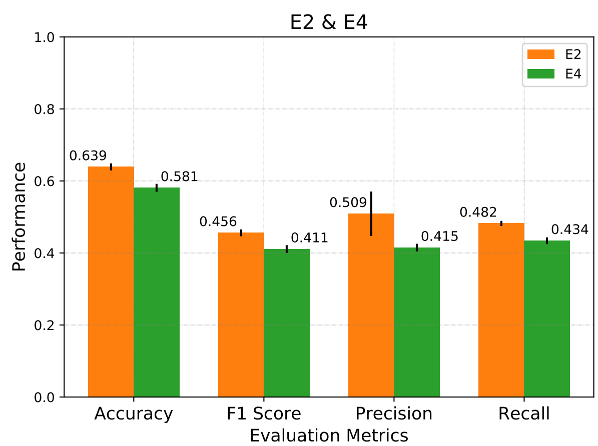
Set II is designed to quantify how outdated data can affect the performance of a misinformation identification model. We conducted three sets of experiments to answer this



(a) E1 & E2 comparison



(b) E1 & E3 comparison



(c) E2 & E4 comparison

Figure 5.2: Experimental result comparisons

question:

E3. In this set, the model was trained using the general news dataset (as in Experiment E2). We extracted randomly 10% of claims (and their related articles) from the COVID-19 dataset and used them for testing. Thus the training data (general news collected before June 2019) is outdated compared with the tested data (COVID-19 news published after November, 2020). We then compared the results from this set with those from Experiment E1, and expected that E3 would perform worse than E1 because E3 uses obsolete data for training.

E4. In this set, the model was trained using the COVID-19 dataset (as in Experiment E1). We extracted randomly 10% of claims (and their related articles) from the general news dataset and used them for testing. In this case, the tested data (general news collected before June 2019) is outdated compared with the training data (COVID-19 news published after November, 2020). We then compared the results from this set with those from Experiment E2, and expected that E4 would perform worse then E2 because E4 uses obsolete data for testing.

E5. In this set, we extracted randomly 10% of claims (and their related articles) from the COVID-19 dataset and used them for testing. The training data was a mix of COVID-19 and general news following a ratio of $g : (1 - g)$, where $g = \{0, 0.1, 0.2, \ldots, 0.5, 1\}$, resulting in a total of seven configurations. We expected that as g, the percentage of COVID-19 news used for training, increased from 0 to 1, the performance of the classification model would improve thanks to more and more up-to-date data (COVID-19 news published after November, 2020) being added to the training data.

Table 5.3 summarizes the configurations of the five experiments. Table 5.4 shows the performance of the experiments E1 to E4. Figure 5.2 illustrates the comparisons of the model performance in E1 to E4. Figure 5.3 shows the performance of the seven configurations in Experiment E5. All graphs were plotted with 95% confidence intervals.

Table 5.3: Experiment configurations

Experiment	Training	Testing
E1	COVID-19	COVID-19
E2	General News	General News
E3	General News	COVID-19
E4	COVID-19	General News
E5	Mix	COVID-19

Table 5.4: Experimental results

	E1	E2	E3	E4
Accuracy	0.8388279	0.6392274	0.688646	0.581028
F1 score	0.7089519	0.456040	0.340436	0.410775
Precision	0.789149	0.508782	0.337629	0.414775
Recall	0.691367	0.482305	0.368899	0.434011

5.3.1 Analysis of Set I

Figure 5.2(a) illustrates a comparison of E1 (COVID-19 dataset) and E2 (general news dataset). It shows that the accuracy of E1 is 20.0 percentage points higher than that of E2, 0.839 vs. 0.639. Similarly, the F1 score, precision and recall of E1 are improved by 25.3 (0.709 vs. 0.456), 28.0 (0.789 vs. 0.509) and 20.9 percentage points (0.691 vs. 0.482), respectively. The results demonstrate that, by narrowing down the knowledge domain to a specific topic (COVID-19 in this case), we can significantly improve the classification accuracy and other metrics.

Collecting a specialized dataset (e.g., COVID-19, stock markets, sports) is more labor and time consuming than a general news dataset. However by quantifying the performance difference as described above, we can see a big performance gap between using specialized datasets vs. general news datasets. This helps justify the additional cost of collecting specialized datasets.

5.3.2 Analysis of Set II

Figure 5.2(b) compares the results of E1 (up-to-date training data) and those of E3 (outdated training data). It shows that the accuracy, F1 score, precision and recall of E1 are significantly higher than those of E3, by 15.0, 36.9, 45.2, and 32.2 percentage points, respectively. The lower performance of E3 is the result of obsolete training data. The up-to-date training data enables E1 to achieve much higher performance than E3.

Figure 5.2(c) compares the results of E2 (up-to-date testing data relative to training data) and those of E4 (outdated testing data). The graph shows that E2 yielded higher accuracy, F1 score, precision and recall, by 5.8, 4.7, 9.4, and 4.8 percentage points, respectively. The lower performance of E4 results from a mismatch between the timelines of the training and testing data, with the testing data being outdated compared to the training data. The result would be optimal if the timeline of testing data matched that of the training data. This observation is validated further in Experiment E5.

In E5, the testing data was COVID-19. The training data is a mix of COVID-19 and general news with a ratio of $g : (1-g)$, where $g = \{0, 0.1, 0.2, \ldots, 0.5, 1\}$. As g, the percentage of COVID-19 news in the mix, increased from 0 to 1, more and more COVID-19 training data was added to the mix to match the timeline of the testing data. Consequently, the performance of the classification model improved as more COVID-19 training data was added to the mix, as shown in Figure 5.3. As the percentage of COVID-19 data g increases from 0 to 0.5, the accuracy, F1 score, precision and recall improve by 19.1, 22.3, 27.5, and 18.2 percentage points, respectively.

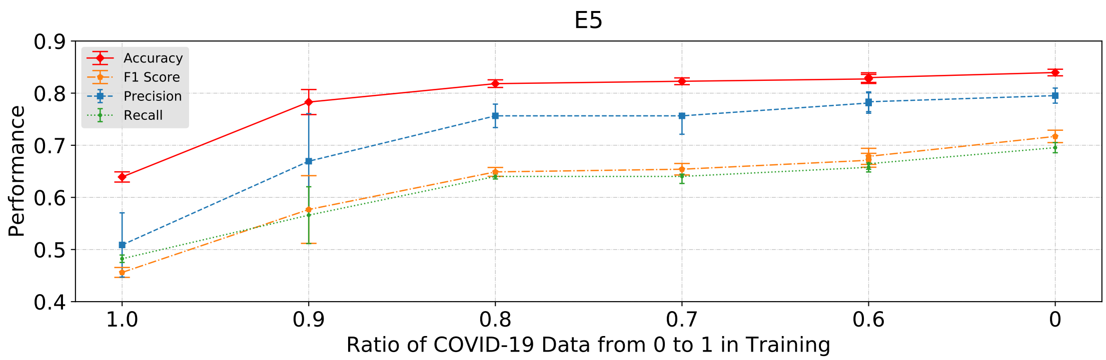


Figure 5.3: Results of experiment E5

5.4 Chapter Summary

In summary, the results of the above experiments indicate that

- the performance of a misinformation identification model can be improved significantly by narrowing down the knowledge domain to a specific topic such as COVID-19, the stock market, the 2020 US election, sports, or climate change.

- it is important to match the timeline of the training data with that of the testing data to obtain the best performance.

- training datasets used for classification tasks on *current news* must be updated regularly to provide the best performance.

Chapter 6

Conclusion and Future Work

6.1 Summary

Fake news is false information widely spread through media such as newspapers, magazines, television, radio, podcast, blogs, and social networks. While fake news has existed throughout human history, the popularity of social media has accelerated the speed and scope of misinformation propagation and exacerbated the harm caused by false information. Therefore, identifying misinformation is crucial to maintaining political, social, financial stability and democracy.

Our first contribution is the construction of the COVMIS dataset, a new large-scale, feature-rich, publicly available dataset for research on COVID-19 misinformation. COVMIS contains 14,384 claims, 134,320 related articles and many features associated with the claims, such as claimant, source, date and explanation. We provide a comprehensive analysis of the dataset to understand the data better.

Our second contribution is to evaluate several state-of-the-art transformer-based NLP models (i.e., BERT, DistilBERT, XLNet, RoBERTa, and ALBERT) using the COVMIS dataset. We evaluated the classification performance of the NLP models using the WA (whole articles) method and the MRS (most relevant sentences) method. Our experimental results show an improvement in terms of accuracy, F1 score, precision, and recall using the MRS method, compared to the WA method.

Our third contribution is to improve classification performance by narrowing the knowledge domain of the input data to a specific topic and to quantify the performance difference yielded by up-to-date and obsolete training data.

Our experimental results show that

- the performance of a misinformation identification model can be improved significantly by narrowing down the knowledge domain to a specific topic such as COVID-19, the stock market, the 2020 US election, sports, or climate change.
- it is important to match the timeline of the training data with that of the testing data to obtain the best performance.
- training datasets used for classification tasks on *current news* must be updated regularly to provide the best performance.

6.2 Future Work

Following are several potential research directions for future work:

- For the thesis, we curated a dataset named COVMIS that contains features such as claim, claimant, publication dates, news source, label, the country of where the claim is from, and related articles of each claim. In the future, we will enhance COVMIS by (i) collecting information on social user engagement and user networks to enable more research activities such as multi-model detection tasks, and (ii) updating the dataset periodically to reflect the progress of COVID-19 research, which is needed for real-time misinformation identification.
- To solve RQ.2, we extracted the most relevant sentences using an information retrieval algorithm to overcome the input limitation of 512 tokens for the five transformer-based

NLP models that we used (i.e., BERT, DistilBERT, XLNet, RoBERTa and ALBERT). For future work, we will evaluate a model named Longformer [152] that can accept up to 4096 input tokens to find out if it can improve the classification performance.

- To solve RQ.3, we applied the automatic fact-checking approach to identify COVID-19 misinformation. In the future, we will apply this approach to other knowledge domains such as stock markets or climate change for misinformation identification. Moreover, we will build a system for misinformation identification according to the topic of the claim. A claim (statement) to be classified will then be processed in two stages. In the first stage, a separate NLP-based classifier will place the claim into a specific category (e.g., COVID-19, stock markets, climate change, sports). The claim will then be evaluated using a misinformation identification model trained for that category to obtain the best performance.

Bibliography

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171–4186, June 2019.

[2] Nir Kshetri and Jeffrey Voas, "The economics of 'fake news'," *IT Professional*, vol. 19, no. 6, pp. 8–12, November 2017.

[3] Pew Research Center, "The future of truth and misinformation online," `https://www.pewresearch.org/internet/2017/10/19/the-future-of-truth-and-misinformation-online/`, October 2017.

[4] Hunt Allcott and Matthew Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–36, May 2017.

[5] Jessikka Aro, "The cyberspace war: Propaganda and trolling as warfare tools," *European View*, vol. 15, no. 1, pp. 121–132, June 2016.

[6] Duncan J. Watts, David M. Rothschild, and Markus Mobius, "Measuring the news and its impact on democracy," *Proceedings of the National Academy of Sciences*, vol. 118, no. 15, April 2021.

[7] Carlos Carvalho, Nicholas Klagge, and Emanuel Moench, "The persistent effects of a false news shock," *Journal of Empirical Finance*, vol. 18, no. 4, pp. 597–615, June 2011.

[8] Shimon Kogan, Tobias J Moskowitz, and Marina Niessner, "Fake news: Evidence from financial markets," `https://jhfinance.web.unc.edu/wp-content/uploads/sites/12369/2018/11/2019JHKoganMoskowitzNiessner_2018.pdf`, November 2018.

[9] Wen-Ying Sylvia Chou, Anna Gaysynsky, and Joseph N. Cappella, "We go from here: Health misinformation on social media," *American Journal of Public Health*, vol. 110, no. 3, pp. 273–275, October 2020.

[10] Xinyi Zhou and Reza Zafarani, "A survey of fake news," *ACM Computing Surveys*, vol. 53, no. 5, pp. 1–40, October 2020.

[11] Fabio Tagliabue, Luca Galassi, and Pierpaolo Mariani, "The 'Pandemic' of disinformation in COVID-19," *Springer Nature Comprehensive Clinical Medicine*, pp. 1–3, August 2020.

[12] "Misinformation fueling vaccine hesitancy, PAHO Director says," `https://www.paho.org/en/news/21-4-2021-misinformation-fueling-vaccine-hesitancy-paho-director-says`, April 2021.

[13] "Fake News Challenge," `http://www.fakenewschallenge.org/`, May 2017.

[14] "Talos targets disinformation with fake news challenge victory," `https://blogs.cisco.com/security/talos/talos-fake-news-challenge`, June 2017.

[15] Jingbo Shang, Jiaming Shen, Tianhang Sun, Xingbang Liu, Anja Gruenheid, Flip Korn, Adam D. Lelkes, Cong Yu, and Jiawei Han, "Investigating rumor news using agreement-aware search," *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, October 2018.

[16] Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel, "A simple but tough-to-beat baseline for the fake news challenge stance detection task," https://arxiv.org/abs/1707.03264, July 2017.

[17] Aswini Thota, Priyanka Tilak, Simrat Ahluwalia, Nibrat Lohia, Simeratjeet Ahluwalia, and Nibhrat Lohia, "Fake news detection: A deep learning approach," *Southern Methodist University Data Science Review*, vol. 1, no. 3, pp. 10, August 2018.

[18] Gaurav Bhatt, Aman Sharma, Shivam Sharma, Ankush Nagpal, and Balasubramanian Raman, "Combining neural, statistical and external features for fake news stance identification," https://doi.org/10.1145/3184558.3191577, November 2018.

[19] Bilal Ghanem, Paolo Rosso, and Francisco Rangel, "Stance detection in fake news a combined feature representation," *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pp. 66–71, November 2018.

[20] Andreas Hanselowski, Avinesh Pvs, Benjamin Schiller, and Felix Caspelherr, "Description of the system developed by team athene in the FNC-1," https://github.com/hanselowski/athene_system/blob/master/system_description_athene.pdf, June 2017.

[21] Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych, "A retrospective analysis of the fake news challenge stance-detection task," *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1859–1874, August 2018.

[22] Ali K Chaudhry, Darren Baker, and Philipp Thun-Hohenstein, "Stance detection for the fake news challenge: Identifying textual relationships with deep neural nets," https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2760230.pdf, June 2017.

[23] Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti, "Automatic stance detection using end-to-end memory networks," https://doi.org/10.48550/arxiv.1804.07581, April 2018.

[24] Heejung Jwa, Dongsuk Oh, Kinam Park, Jang Mook Kang, and Heuiseok Lim, "exBAKE: Automatic fake news detection model based on Bidirectional Encoder Representations from Transformers (BERT)," *Applied Sciences*, vol. 9, no. 19, pp. 1–9, October 2019.

[25] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimedia Tools and Applications*, pp. 11765–11788, January 2021.

[26] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier, "Tweetcred: Real-time credibility assessment of content on Twitter," https://arxiv.org/abs/1405.5490, May 2015.

[27] Hamid Karimi, Proteek Roy, Sari Saba-Sadiya, and Jiliang Tang, "Multi-source multiclass fake news detection," *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1546–1557, August 2018.

[28] Natali Ruchansky, Sungyong Seo, and Yan Liu, "CSI: A hybrid deep model for fake news detection," *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 797–806, September 2017.

[29] Niraj Sitaula, Chilukuri K. Mohan, Jennifer Grygiel, Xinyi Zhou, and Reza Zafarani, "Credibility-based fake news detection," https://doi.org/10.48550/arxiv.1911.00643, November 2019.

[30] Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu, "Jointly embedding the local and global relations of heterogeneous graph for rumor detection," https://doi.org/10.48550/arXiv.1909.04465, September 2019.

[31] Sardar Hamidian and Mona T Diab, "Rumor detection and classification for Twitter data," https://doi.org/10.48550/arxiv.1912.08926, November 2019.

[32] Ke Wu, Song Yang, and Kenny Q. Zhu, "False rumors detection on Sina Weibo by propagation structures," *IEEE 31st International Conference on Data Engineering*, pp. 651–662, April 2015.

[33] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang, "Prominent features of rumor propagation in online social media," *IEEE 13th International Conference on Data Mining*, pp. 1103–1108, December 2013.

[34] Juan Cao, Junbo Guo, Xirong Li, Zhiwei Jin, Han Guo, and Jintao Li, "Automatic rumor detection on microblogs: A survey," https://doi.org/10.48550/arxiv.1807.03505, July 2018.

[35] Naser Ahmadi, Joohyung Lee, Paolo Papotti, and Mohammed Saeed, "Explainable fact checking with probabilistic answer set programming," https://doi.org/10.48550/arxiv.1906.09198, June 2019.

[36] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang, "TabFact: A large-scale dataset for table-based fact verification," https://doi.org/10.48550/arxiv.1909.02164, June 2020.

[37] Nayeon Lee, Yejin Bang, Andrea Madotto, Madian Khabsa, and Pascale Fung, "Towards few-shot fact-checking via perplexity," https://doi.org/10.48550/arxiv.2103.09535, March 2021.

[38] Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen, "MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4685–4697, November 2019.

[39] Mohamed H. Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum, "Tracy: Tracing facts over knowledge graphs and text," *The World Wide Web Conference*, pp. 3516–3520, May 2019.

[40] Yixin Nie, Haonan Chen, and Mohit Bansal, "Combining fact extraction and verification with neural semantic matching networks," `https://doi.org/10.48550/arxiv.1811.07039`, November 2018.

[41] Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan, "FANG: Leveraging social context for fake news detection using graph representation," *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 1165–1174, October 2020.

[42] Nayeon Lee, Belinda Z. Li, Sinong Wang, Wen tau Yih, Hao Ma, and Madian Khabsa, "Language models as fact checkers?," `https://doi.org/10.48550/arxiv.2006.04102`, July 2020.

[43] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2931–2937, September 2017.

[44] "Opinion - Trump's lies vs. Obama's," `https://www.nytimes.com/interactive/2017/12/14/opinion/sunday/trump-lies-obama-who-is-worse.html`, December 2017.

[45] Leaders Prize: Fact or Fake News? Competition, "Documentation for the Data," `https://leadersprize.truenorthwaterloo.com/en/`, June 2019.

[46] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," `https://doi.org/10.48550/arXiv.1910.01108`, March 2020.

[47] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," `https://doi.org/10.48550/arXiv.1906.08237`, January 2020.

[48] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," `https://doi.org/10.48550/arXiv.1907.11692`, July 2019.

[49] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," `https://doi.org/10.48550/arXiv.1909.11942`, February 2020.

[50] Poynter, "International fact-checking network," `https://www.poynter.org/ifcn/`.

[51] Alyt Damstra, Hajo G. Boomgaarden, Elena Broda, Elina Lindgren, Jesper Strömbäck, Yariv Tsfati, and Rens Vliegenthart, "What does fake look like? A review of the literature on intentional deception in the news and on social media," *Journalism Studies*, pp. 1–17, September 2021.

[52] Xinyi Zhou, Atishay Jain, Vir V. Phoha, and Reza Zafarani, "Fake news early detection: A theory-driven model," *Digital Threats*, vol. 1, no. 2, June 2020.

[53] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, "Fake news detection on social media: A data mining perspective," *ACM Special Interest Group on Knowledge Discovery and Data Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, June 2017.

[54] Toluwase Victor Asubiaro and Victoria L. Rubin, "Comparing features of fabricated and legitimate political news in digital environments (2016-2017)," *Proceedings of the Association for Information Science and Technology*, vol. 55, no. 1, pp. 747–750, November 2018.

[55] Dinusha Vatsalan and Jeyakumar Samantha Tharani, "Understanding the strategies of creating fake news in social media," https://www.preprints.org/manuscript/202011.0369/v2, November 2020.

[56] Maria D. Molina, S. Shyam Sundar, Thai Le, and Dongwon Lee, "'Fake News' is not simply false information: A concept explication and taxonomy of online content," *American Behavioral Scientist*, vol. 65, no. 2, pp. 180–212, February 2021.

[57] Xinyi Zhou and Reza Zafarani, "Network-based fake news detection: A pattern-driven approach," *ACM Special Interest Group on Knowledge Discovery and Data Explorations Newsletter*, vol. 21, no. 2, pp. 48–60, December 2019.

[58] Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu, "Hierarchical propagation networks for fake news detection: Investigation and exploitation," https://arxiv.org/abs/1903.09196, March 2019.

[59] Francesco Ducci, Mathias Kraus, and Stefan Feuerriegel, "Cascade-LSTM: A tree-structured neural classifier for detecting misinformation cascades," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2666–2676, August 2020.

[60] Zilong Zhao, Jichang Zhao, Yukie Sano, Orr Levy, Hideki Takayasu, Misako Takayasu, Daqing Li, Junjie Wu, and Shlomo Havlin, "Fake news propagates differently from real news even at early stages of spreading," *European Physical Journal Data Science*, vol. 9, pp. 1–14, April 2020.

[61] "Detecting fake news online - Wikipedia," https://en.wikipedia.org/wiki/Detecting_fake_news_online#Cascade-based_fake_news_detection.

[62] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly, "Stop clickbait: Detecting and preventing clickbaits in online news media," *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 9–16, October 2016.

[63] Ágnes Veszelszki, "Linguistic and non-linguistic elements in detecting (Hungarian) fake news," *Acta Universitatis Sapientiae Communicatio*, vol. 4, no. 1, pp. 7–35, December 2017.

[64] Samantha Bradshaw, "Disinformation optimised: Gaming search engine algorithms to amplify junk news," *Internet Policy Review*, vol. 8, no. 4, pp. 1–24, December 2019.

[65] Emmanuel J. Genot and Erik J. Olsson, "The dissemination of scientific fake news," *The Epistemology of Fake News*, pp. 228–242, May 2021.

[66] Yoshiteru Ishida and Sanae Kuraya, "Fake news and its credibility evaluation by dynamic relational networks: A bottom up approach," *Procedia Computer Science*, vol. 126, pp. 2228–2237, July 2018.

[67] Diego Esteves, Aniketh Janardhan Reddy, Piyush Chawla, and Jens Lehmann, "Belittling the source: Trustworthiness indicators to obfuscate fake news on the web," *Proceedings of the 1st Workshop on Fact Extraction and VERification (FEVER)*, pp. 50–59, November 2018.

[68] Naveed Hussain, Hamid Turab Mirza, Ghulam Rasool, Ibrar Hussain, and Mohammad Kaleem, "Spam review detection techniques: A systematic literature review," *Applied Sciences (Switzerland)*, vol. 9, no. 5, pp. 1–26, February 2019.

[69] Bin Zhou and Maryland Baltimore, "Online review spam detection by new linguistic features," *iConference 2015 Proceedings*, pp. 1–5, March 2015.

[70] Hao Xue, Fengjun Li, Hao Xue, and Fengjun Li, "A content-aware trust index for online review spam detection," *31st IFIP Annual Conference on Data and Applications Security and Privacy*, pp. 489–508, June 2017.

[71] Guan Wang, Sihong Xie, Bing Liu, and Philip S. Yu, "Identify online store review spammers via social review graph," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 4, September 2012.

[72] Chunyuan Yuan, Wei Zhou, Qianwen Ma, Shangwen Lv, Jizhong Han, and Songlin Hu, "Learning review representations from user and product level information for spam detection," *Proceedings of the IEEE International Conference on Data Mining*, pp. 1444–1449, November 2019.

[73] Simone Leonardi, Giuseppe Rizzo, and Maurizio Morisio, "Automated classification of fake news spreaders to break the misinformation chain," *Information (Switzerland)*, vol. 12, no. 6, pp. 248, June 2021.

[74] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu, "Combating fake news: A survey on identification and mitigation techniques," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 3, May 2019.

[75] Vasu Agarwal, H. Parveen Sultana, Srijan Malhotra, and Amitrajit Sarkar, "Analysis of classifiers for fake news detection," *Procedia Computer Science*, vol. 165, pp. 377–383, November 2019.

[76] Chun Yuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu, "Early detection of fake news by utilizing the credibility of news, publishers, and users based on weakly supervised learning," `https://arxiv.org/abs/2012.04233`, December 2020.

[77] Arup Baruah, Kaushik Amar Das, Ferdous Ahmed Barbhuiya, and Kuntal Dey, "Automatic detection of fake news spreaders using BERT," *Conference and Labs of the Evaluation Forum 2020 Labs and Workshops*, vol. 2696, September 2020.

[78] Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu, "Unsupervised fake news detection on social media: A generative approach," *33rd AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 5644–5651, July 2019.

[79] Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov, "That is a known lie: Detecting previously fact-checked claims," *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pp. 3607–3618, July 2020.

[80] Georgios Karagiannis, Mohammed Saeed, Paolo Papotti, and Immanuel Trummer, "Scrutinizer: A mixed-initiative approach to large-scale, data-driven claim verification," *Proceedings of the Very Large Data Bases Endowment*, vol. 13, no. 12, pp. 2508–2521, August 2020.

[81] Milad Alshomary, Nick Düsterhus, and Henning Wachsmuth, "Extractive Snippet Generation for Arguments," *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1969–1972, July 2020.

[82] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha, "Detecting rumors from microblogs with recurrent neural networks," *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pp. 3818–3824, July 2016.

[83] Jing Ma, Wei Gao, and Kam-Fai Wong, "Detect rumors in microblog posts using propagation structure via kernel learning," *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 708–717, July 2017.

[84] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum, "Where the truth lies: Explaining the credibility of emerging claims on the web and social media," *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 1003–1012, April 2017.

[85] William Yang Wang, "'Liar, Liar Pants on Fire': A new benchmark dataset for fake news detection," *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 422–426, July 2017.

[86] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal, "FEVER: a large-scale dataset for fact extraction and VERification," *Proceedings*

of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 809–819, June 2018.

[87] Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H. Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B. Everett, Waleed Falak, Carl Gieringer, Jack Graney, Kelly M. Hoffman, Lindsay Huth, Zhenya Ma, Mayanka Jha, Misbah Khan, Varsha Kori, Elo Lewis, George Mirano, William T. Mohn IV, Sean Mussenden, Tammie M. Nelson, Sean Mcwillie, Akshat Pant, Priya Shetye, Rusha Shrestha, Alexandra Steinheimer, Aditya Subramanian, and Gina Visnansky, "Fake news vs satire: A dataset and analysis," *Proceedings of the 10th ACM Conference on Web Science*, pp. 17–21, May 2018.

[88] Craig Silverman, Strapagiel Lauren, Shaban Hamza, Hall Ellie, and Singer-Vine Jeremy, "Hyperpartisan Facebook pages are publishing false and misleading information at an alarming rate," `https://www.buzzfeednews.com/article/craigsilverman/partisan-fb-pages-analysis`, October 2016.

[89] Benjamin D. Horne and Sibel Adali, "This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news," `https://doi.org/10.48550/arXiv.1703.09398`, March 2017.

[90] Jeppe Norregaard, Benjamin D. Horne, and Sibel Adali, "NELA-GT-2018: A large multi-labelled news dataset for the study of misinformation in news articles," `https://doi.org/10.48550/arxiv.1904.01546`, April 2019.

[91] Maurício Gruppi, Benjamin D. Horne, and Sibel Adalı, "NELA-GT-2019: A large multi-labelled news dataset for the study of misinformation in news articles," `https://doi.org/10.48550/arXiv.2003.08444`, March 2020.

[92] Maurício Gruppi, Benjamin D. Horne, and Sibel Adalı, "NELA-GT-2020: A large multi-labelled news dataset for the study of misinformation in news articles," `https://doi.org/10.48550/arXiv.2102.04567`, February 2021.

[93] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu, "Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media," https://doi.org/10.48550/arXiv.1809.01286, March 2019.

[94] Enyan Dai, Yiwei Sun, and Suhang Wang, "Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, no. 1, pp. 853–862, May 2020.

[95] Limeng Cui and Dongwon Lee, "CoAID: COVID-19 healthcare misinformation dataset," https://arxiv.org/pdf/2006.00885.pdf, November 2020.

[96] Dimitar Dimitrov, Erdal Baran, Pavlos Fafalios, Ran Yu, Xiaofei Zhu, Matthäus Zloch, and Stefan Dietze, "TweetsCOV19 - A knowledge base of semantically annotated tweets about the COVID-19 pandemic," *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, October 2020.

[97] Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell, "A large-scale COVID-19 Twitter chatter dataset for open scientific research - An international collaboration," *Epidemiologia*, vol. 2, no. 3, pp. 315–324, August 2021.

[98] Xiaolei Huang, Amelia Jamison, David Broniatowski, Sandra Quinn, and Mark Dredze, "Coronavirus Twitter data: A collection of COVID-19 tweets with automated annotations," http://twitterdata.covid19dataresources.org/index, March 2020.

[99] Daniel Kerchner and Laura Wrubel, "Coronavirus Tweet IDs," https://doi.org/10.7910/DVN/LW0BTB, December 2020.

[100] Martin Mueller and Marcel Salathé, "Crowdbreaks: Tracking health trends using public social media data and crowdsourcing," *Frontiers in Public Health*, vol. 7, April 2019.

[101] Umair Qazi, Muhammad Imran, and Ferda Ofli, "GeoCoV19: A dataset of hundreds of millions of multilingual COVID-19 tweets with location information," *Special Interest Group on Spatial Information Special*, vol. 12, no. 1, pp. 6–15, June 2020.

[102] Sarah Alqurashi, Ahmad Alhindi, and Eisa Alanazi, "Large Arabic Twitter dataset on COVID-19," `https://doi.org/10.48550/arXiv.2004.04315`, April 2020.

[103] Zhiwei Gao, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki, "NAIST COVID: Multilingual COVID-19 Twitter and Weibo dataset," `https://doi.org/10.48550/arXiv.2004.08145`, April 2020.

[104] Rabindra Lamsal, "Coronavirus (COVID-19) geo-tagged tweets dataset," `https://dx.doi.org/10.21227/fpsb-jz61`, August 2020.

[105] Ibrahim Sabuncu; Zeynep Yurek, "Corona virus (COVID-19) Turkish tweets dataset," `https://dx.doi.org/10.21227/0wf0-0792`, August 2020.

[106] Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani, "ReCOVery: A multimodal repository for COVID-19 news credibility research," *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 3205–3212, October 2020.

[107] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex Wade, Kuansan Wang, Nancy Xin Ru Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier, "CORD-19: The COVID-19 open research dataset," `https://doi.org/10.48550/arXiv.2004.10706`, July 2020.

[108] Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty, "Fighting

an infodemic: COVID-19 fake news dataset," https://doi.org/10.48550/arXiv.2011.03327, May 2021.

[109] PolitiFact, "Fact-checking website that rates the accuracy of claims by elected officials and others on its truth-o-meter," https://www.politifact.com/.

[110] Snopes, "The definitive fact-checking site and reference source for urban legends, folklore, myths, rumors, and misinformation," https://www.snopes.com/.

[111] Agence France-Presse(AFP), "The news hub," https://www.afp.com/en/news-hub.

[112] Africa Check, "Sorting fact from fiction," https://africacheck.org/.

[113] POLYGRAPH.info, "Fact-checking website produced by Voice of America (voa)," https://www.polygraph.info/.

[114] Reuters, "Breaking international news views," https://www.reuters.com/.

[115] The Washington Post, "Breaking news, world, US, DC news and analysis," https://www.washingtonpost.com/.

[116] Google Fact Check Tools, "Search fact checks about a topic or person search fact checks from the web search," https://toolbox.google.com/factcheck/explorer.

[117] Media Bias/Fact Check, "Search and learn the bias of news media," https://mediabiasfactcheck.com/.

[118] Canadian Centre for Occupational Health and Safety, "An independent departmental corporation under schedule ii of the Financial Administration Act and is accountable to Parliament through the Minister of Labour," https://www.ccohs.ca/.

[119] ScienceDaily, "Your source for the latest research news," https://www.sciencedaily.com/.

[120] National Institutes of Health (NIH), "Turning discovery into health," https://www.nih.gov/.

[121] WebMD, "Better information. Better health," https://www.webmd.com/.

[122] Medical and health information, "Medical news and health news headlines posted throughout the day, every day," https://www.medicalnewstoday.com/.

[123] "Psychology of rumors: 6 reasons why rumors spread," https://medium.com/@So_Psych/the-psychology-of-rumors-f8cf1555ead2, September 2015.

[124] Vishal A. Kharde and S.S. Sonawane, "Sentiment analysis of Twitter data: A survey of techniques," *International Journal of Computer Applications*, vol. 139, no. 11, pp. 5–15, April 2016.

[125] Dietmar Gräbner, Markus Zanker, Günther Fliedl, and Matthias Fuchs, "Classification of customer reviews based on sentiment analysis," *Information and Communication Technologies in Tourism 2012*, pp. 460–470, May 2012.

[126] Xing Fang and Justin Zhan, "Sentiment analysis using product review data," *Journal of Big Data*, vol. 2, no. 1, June 2015.

[127] Samuel Brody and Noemie Elhadad, "An unsupervised aspect-sentiment model for online reviews," *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 804–812, June 2010.

[128] Mohammed Issa Al-Kharusi, Abubakar Idris Usman, and Jamilu Awwalu, "Application of Sentiment Analysis in Business Intelligence," *International Journal of Knowledge, Innovation, and Enterepreneurship*, vol. 3, no. 3, pp. 51 – 60, January 2015.

[129] Anita Yadav, C. K. Jha, Aditi Sharan, and Vikrant Vaish, "Sentiment analysis of financial news using unsupervised approach," *Procedia Computer Science*, vol. 167, pp. 589–598, March 2020.

[130] Hima Suresh and Gladston S., "An unsupervised fuzzy clustering method for Twitter sentiment analysis," *International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, pp. 80–85, October 2016.

[131] Noureddine Azzouza, Karima Akli-Astouati, Amira Oussalah, and Samy Ait Bachir, "A real-time Twitter sentiment analysis using an unsupervised method," *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics*, June 2017.

[132] Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov, "SemEval-2015 task 10: Sentiment analysis in Twitter," *Proceedings of the 9th International Workshop on Semantic Evaluation*, pp. 451–463, June 2015.

[133] Milagros Fernández-Gavilanes, Tamara Álvarez-López, Jonathan Juncal-Martínez, Enrique Costa-Montenegro, and Francisco Javier González-Castaño, "GTI: An unsupervised approach for sentiment analysis in Twitter," *Proceedings of the 9th International Workshop on Semantic Evaluation*, pp. 533–538, June 2015.

[134] Pedro Henrique Arruda Faustini and Thiago Ferreira Covões, "Fake news detection in multiple platforms and languages," *Expert Systems with Applications*, vol. 158, pp. 113503, November 2020.

[135] Hugo Queiroz Abonizio, Janaina Ignacio de Morais, Gabriel Marques Tavares, and Sylvio Barbon Junior, "Language-independent fake news detection: English, Portuguese, and Spanish mutual features," *Future Internet*, vol. 12, no. 5, May 2020.

[136] Gleb Kuzmin, Daniil Larionov, Dina Pisarevskaya, and Ivan Smirnov, "Fake news detection for the Russian language," *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, pp. 45–57, December 2020.

[137] Renato M. Silva, Roney L.S. Santos, Tiago A. Almeida, and Thiago A.S. Pardo, "Towards automatically filtering fake news in Portuguese," *Expert Systems with Applications*, vol. 146, pp. 113199, May 2020.

[138] Max Grusky, Mor Naaman, and Yoav Artzi, "Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies," *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 708–719, June 2018.

[139] Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer, "Variations of the similarity function of textrank for automated summarization," `https://arxiv.org/abs/1602.03606`, February 2016.

[140] Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein, "Learning-based single-document summarization with compression and anaphoricity constraints," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 1998–2008, August 2016.

[141] René Arnulfo García-Hernández, Romyna Montiel, Yulia Ledeneva, Eréndira Rendón, Alexander Gelbukh, and Rafael Cruz, "Text summarization by sentence extraction using unsupervised learning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5317 LNAI, pp. 133–143, October 2008.

[142] Mahmood Yousefi-Azar and Len Hamey, "Text summarization using unsupervised deep learning," *Expert Systems with Applications: An International Journal*, vol. 68, pp. 93–105, February 2017.

[143] Raj Dabre, Kehai Chen, Benjamin Marie, Rui Wang, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita, "NICT's supervised neural machine translation systems for the WMT19 news translation task," *Proceedings of the 4th Conference on Machine Translation*, vol. 2, pp. 168–174, August 2019.

[144] Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu, "Semi-supervised learning for neural machine translation," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 1965–1974, August 2016.

[145] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho, "Unsupervised neural machine translation," `https://doi.org/10.48550/arXiv.1710.11041`, February 2018.

[146] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato, "Unsupervised machine translation using monolingual corpora only," `https://doi.org/10.48550/arXiv.1711.00043`, April 2018.

[147] Zuchao Li, Hai Zhao, Rui Wang, Kehai Chen, Masao Utiyama, and Eiichiro Sumita, "SJTU-NICT's supervised and unsupervised neural machine translation systems for the WMT20 news translation task," *Proceedings of the 5th Conference on Machine Translation*, pp. 218–229, November 2020.

[148] Sourav Dutta, Jesujoba Alabi, Saptarashmi Bandyopadhyay, Dana Ruiter, and Josef van Genabith, "UdS-DFKI@WMT20: Unsupervised MT and very low resource supervised MT for German-Upper Sorbian," *Proceedings of the Fifth Conference on Machine Translation*, pp. 1092–1098, November 2020.

[149] Karen Spärck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, pp. 11–21, 1972.

[150] Dan Hendrycks and Kevin Gimpel, "Gaussian error linear units (GELUs)," `https://doi.org/10.48550/arxiv.1606.08415`, July 2020.

[151] Francisco J. Valverde-Albacete and Carmen Peláez-Moreno, "100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox," *PLoS ONE*, vol. 9, no. 1, January 2014.

[152] Iz Beltagy, Matthew E. Peters, and Arman Cohan, "Longformer: The long-document transformer," https://doi.org/10.48550/arXiv.2004.0515, December 2020.