# Misinformation Retrieval

by

Saira Rizvi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Data Science

Waterloo, Ontario, Canada, 2021

## Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

## Statement of Contributions

This thesis includes work produced with the contribution of Huiyu Min in Chapter 4 section 4.1. Huiyu crawled the tweets that were used to construct the test set used in the experiments. I am the sole author of everything else in this work, including labeling the test set.

# Abstract

This work introduces the task of misinformation retrieval, identifying all documents containing misinformation for a given topic, and proposes a pipeline for misinformation retrieval on tweets. As part of the work, I curated 50 COVID-19 misinformation topics used in the TREC 2020 Health Misinformation track. In addition, I annotated a test set of tweets using the TREC COVID-19 misinformation on social media. Misinformation on social media has proven highly detrimental to communities by encouraging harmful and often life-threatening behavior. The chaos caused by COVID-19 misinformation has created an urgent need for misinformation detection methods to moderate social media platforms. Drawing upon previous work in misinformation detection and the TREC 2020 Health Misinformation Track, I focused on the task of misinformation retrieval on social media. I extended the COVID-Lies data set created to detect COVID-19 misinformation in tweets by rephrasing the misconceptions accompanying each tweet. I also created 50 COVID-19 related topics for the TREC 2020 Health Misinformation track used for evaluation purposes. I propose a natural language inference (NLI) based approach using CT-BERT to identify tweets that contradict a given fact, used to score documents utilizing the model's classification probability. The model was trained using a combination of NLI data sets to find the best approach. Tweets were labeled for the TREC 2020 Health Misinformation Track topics to create a test set on which the best model achieves an AUC of 0.81. I conducted several experiments which show that domain adaptation significantly improved the ability to detect misinformation. A combination of a large NLI corpus, such as SNLI, and an in-domain, such as the COVID-Lies, data set achieves the best performance on our test set. The pipelines retrieved and ranked tweets based on misinformation for 7 TREC topics from the COVID-19 Twitter stream. The top 20 unique tweets were analyzed using Precision@20 to evaluate the pipeline.

## Acknowledgements

I want to give my deepest gratitude to my supervisor, Charles Clarke, for making it possible to complete my thesis by guiding me and supporting me in my research. I appreciate the advice given to me by Charles and the freedom to explore my ideas in depth. I have amassed immense knowledge under his supervision over the past two years enabling me to succeed in my future endeavors.

Aside from my supervisor, I would like to thank Maura Grossman and Gordon Cormack for reading and providing feedback on my thesis, allowing me to improve my work.

I want to thank my parents, Memona Tahir and Ali Tahir, for providing me with support and guidance throughout my education and endless opportunities to succeed in my career. Lastly, I want to thank Andre Kassis for encouraging me to work hard and inspiring my hunger for knowledge.

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Since its inception, COVID-19 has proven to be one of the deadliest pandemics that humanity has experienced in last 100 years[1]. However, the misinformation spread on social media regarding the disease may be just as detrimental to the world's population as the virus itself, causing several disasters such as the burning of 5G towers[2], and deaths from methanol poisoning[3]. The predominant adverse impact of misinformation on social media calls for more sophisticated approaches than those that existed before the pandemic to moderate these platforms.

The pandemic has changed the misinformation landscape on social media and demanded new methods for COVID-19 misinformation detection due to the influx of incorrect claims circulating caused by increased use of social media. Most work focuses on simply

[1] https://www.idse.net/Covid-19/Article/10-20/Fauci–COVID-19-Worst-Pandemic-in-100-Years/60937

[2] https://www.cnet.com/health/5g-coronavirus-conspiracy-theory-sees-77-mobile-towers-burned-report-says/

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7272173/

classifying misinformation or curating data sets for COVID-19 misinformation research. The TREC Health Misinformation track[4] differs from the majority of the work in the misinformation field as it is concerned with improving the quality of search results. The 2020 Health Misinformation Track focuses on the COVID-19 pandemic, and introduces a new task of total recall of misinformation, i.e. retrieving all documents with misinformation for a given topic. This is more difficult than devising ranking algorithms to promote correct information (The "adhoc" task in the track), because even though misinformation online is very influential, it may be an extremely small subset of a very large corpus. Total recall of misinformation is beneficial for search engines and social media platforms in many ways, since it can simplify the tasks of content moderators by providing them with documents that are very likely to contain misinformation, specially when there is a high volume of misinformation in the case of the pandemic. The pressing need to monitor and track misinformation to control the "infodemic" and prevent another one from occurring has made it necessary to automate the retrieval of all misinformation on social media platforms. This work focuses on the retrieval of misinformation by examining the TREC 2020 results and formulated a misinformation retrieval task for social media monitoring.

The TREC 2020 Health Misinformation Track [9] challenged the participants with two tasks related to COVID-19 misinformation: Adhoc retrieval and total recall. The adhoc retrieval task demanded that participants produce runs that rank documents based on relevance, correctness, and credibility for a given topic. The total recall task demanded the participants produce runs containing all the incorrect documents present in the corpus for a given topic. The corpus used for both tasks was the common crawl corpus consisting of news articles online. The best performing run for adhoc retrieval was by the h20loo team [37], which leveraged stance detection through T5, and the best performing run for

[4] https://trec-health-misinfo.github.io/

total recall task was submitted by KU [26], which used stance detection by finetuning RoBERTa. The total recall task proved to be more challenging compared to the adhoc retrieval task, with the best run achieving an R-precision of 0.13. The track results also revealed that traditional news sources online are not a major contributor to misinformation since very few articles contained incorrect information compared to correct information. Furthermore, the COVID-19 “infodemic” has demonstrated the influence of social media in the spread of misinformation. Therefore I decided to focus our attention on misinformation retrieval on social media platforms, specifically Twitter, and aimed to answer the following crucial research question:

- **RQ1:** Is it possible to automatically detect all documents in a corpus that are promulgating misinformation? If yes, what is the best approach to complete this task?

The TREC 2020 Health Misinformation track results show the power of using stance detection to improve the quality of search results. Based on this observation, I explore the research question in depth by first devising a misinformation classification approach based on a natural language inference for the total recall of misinformation in a corpus of tweets. Our proposed method uses a BERT-based NLI model that uses a known fact to determine if a tweet is factual or not, which I test on a test set annotated for this challenge. The classifier is then used to score tweets based on the level of misinformation and retrieve all tweets with misinformation for the TREC topics. I used a publicly available dataset, COVID-Lies [21], for finetuning BERT. However, the annotated data was limited, so I considered domain adapted BERT [14] model, CT-BERT [31], to remedy this problem. I ran several experiments that revealed that domain adaption on COVID-19 tweets in conjunction with finetuning on a larger NLI corpus before finetuning on task-specific data significantly improved the model performance. The approach shows promising results, with the best model achieving an ROC AUC of 0.81. Our contributions can be summarized as follows:

- I formulated the challenging task of *Misinformation Retrieval* and trained a misinformation retrieval pipeline that uses the power of natural language inference,
- I curated a list of 50 COVID-19 related misinformation topics along with the correct information with the help of other track organizers, which was used in the TREC 2020 Health Misinformation track,
- I annotated 1113 tweets based on their correctness for the TREC 2020 Health Misinformation track topics,
- I expanded the COVID-Lies dataset by adding new phrasing of misconceptions to offset the class imbalance,
- Finally, i retrieved incorrect tweets for 7 TREC topics from the official COVID-19 Twitter stream.

I will first introduce the research conducted in misinformation detection, pre- and post-pandemic, including the TREC 2019 Health Misinformation track. I will dig deeper into the 2020 Health Misinformation Track and analyze the results. Based on our findings, I shift our focus to social media and present our work in devising a misinformation retrieval pipeline.

# Chapter 2

# Background and Related Work

## 2.1 Types of Incorrect Information

Various types of incorrect information are currently being studied in automating the detection of incorrect information. Here I will discuss the terminology used in the research community to define incorrect information and their relationship to one another.

***Misinformation*** can be defined as incorrect information that is spread unintentionally according to Wu et. al [47] and Guo et. al [17]. Users spread this type of information to inform others since they believe it comes from a credible source.

***Disinformation*** is incorrect information that is spread intentionally to gain some advantage by the parties that start these falsehoods as explained by Guo et. al [17]. An example of this type of false information can be disinformation campaigns launched during elections to gain support or harm a political party.

***Fake news*** is a fabricated piece of verifiable information that is spread in the format of news and can be spread as a part of a disinformation campaign as defined by

Shu et. al [39]. This type of false information may appear to be very reliable and can cause significant harm by easily convincing its readers of the claims being made. Guo et. al [17] shows how fake news is a subset of disinformation online.

## 2.2 Pre-COVID-19 Misinformation

Prior to the COVID-19 pandemic, the majority of the work falls into the categories: stance detection-based approaches, content-based approaches, and propagation-based approaches. Stance detection-based approaches attempt to detect the position a text takes about an established ground truth. A misinformation detection pipeline either incorporates stance detection as part of the pipeline or relies entirely on it to distinguish between incorrect and correct claims. Content-based approaches are older than stance detection-based approaches and heavily rely on feature engineering to build a misinformation detection pipeline. These works focus on finding the best features that enable you to validate the information and often aim to detect misinformation through detecting credible information. Propagation-based approaches are concentrated in the social media domain and rely on the hypothesis that the propagation patterns of misinformation differ from correct information. Therefore by capturing these differences in the propagation pattern, we can successfully detect misinformation on social media.

I also looked into the TREC 2019 Health Misinformation Track, which focuses on the quality of the information provided in search results. The Health Misinformation track differs from the other works discussed in this section since it focuses on the public health domain and aims to create the best ranking algorithms to provide the most correct and credible information in search results.

### Stance Detection Based Approaches

Tacchini et. al [40] aim to detect hoaxes, i.e., a type of misinformation fabricated with the intent of masking the truth, based on the stance users take on Facebook posts. The method purposed in [40] relies on the hypothesis that the correctness of posts and users' opinions of posts have a relationship that can be leveraged for hoax detection. The information regarding the stance of users is collected by observing the "likes" on Facebook posts. This work proposes two classification methods based on logistic regression and a novel adaption of a boolean crowdsourcing algorithm with best-performing accuracy exceeding 99%. This work is unique because it does not compare the stance of the post with a fact using textual information like the rest of the stance detection-based approaches discussed in this section.

Methods relying on stance detection or inferring textual entailment have been widely used on their own and as a component of fact-checking pipelines. Pomerleau et. al [35] organized the Fake News Challenge (FNC)[1] in 2017 to develop high performing stance detection tools that can be utilized in fact-checking of news stories. The task asks participants to determine the relationship between a news article body and a headline, possibly from another article. The possible relations in this context are "positive", "negative", "discusses", and "unrelated". The FNC task was designed to be a standalone problem and does not provide ground truths for fact-checking. Instead, it aims to create a tool that verifies claims based on known facts and leaves the collection of correct facts to the user.

Derczynski et. al [13] introduced a shared task in 2017, RumourEval, for automatic claim verification for rumors circulating on social media. The task aimed to verify rumors by analyzing a single post and the user reactions and comments. The sub-task included stance detection and veracity, which were carried forward in the 2019 task [16] in a similar

[1] http://www.fakenewschallenge.org/

manner. The 2019 task was extended to include Reddit posts.

The Fact Extraction and VERification Task [41] was introduced in 2018 with two objectives, evidence extraction, and claim verification. The participants were tasked to provide a system that identifies sentences from Wikipedia articles as evidence, and given the evidence, the claim is categorized as SUPPORTED, REFUTED, or NOTENOUGH-INFO. The task was supplemented with the FEVER dataset, with an annotated claim and sentence pairs that can be found on the FEVER website [2]. In 2019, the task [3] was extended to provide more robust frameworks that are resilient to adversaries by provided attacks on the approach and then finding solutions.

### Content Based Approaches

Castillo et. al [8] proposed a method that assesses the credibility of a set of tweets based on the features extracted from the tweet content. The features extracted from the tweets were based on user behaviors, such as tweeting and re-tweeting content, the text of the tweet, and the citations to external resources. The data used to train the classifiers was collected from the trending Twitter topics and judged by humans based on whether the tweet contained news-worthy content and then evaluated by another group of annotators on the credibility of the content.

Hardalov et. al [20] propose a content-based approach that automatically detects credible news from not credible news based on a feature set of n-grams, grammatical features, and semantic features. The work first collects three datasets in Bulgarian from several Bulgarian news sources, with credible news coming from well-known reliable news sources and not credible from unreliable sources. Hardalov et. al reported achieving high

[2] https://fever.ai/

[3] https://fever.ai/2019/task.html

accuracy and showed the best performance when semantic features were used in combination with other features.

Work such as Guo et. al [18], and Giachanou et. al [15], aim to measure the credibility of posts by capturing the emotions being conveyed regarding the post. Guo et. al present an Emotion-based Fake News Detection framework (EFN) that uses RNNs to capture the emotions in the content and the posts' comments to exploit this information for fake news detection. Giachanou et. al [15] focuses just on the contents of the post to capture the emotions conveyed. The method proposed in [15] is an LSTM network that takes two inputs, emotional signals, and word embedding, to assess the credibility of the post. Zhang et. al [49] suggest that the emotional features extracted from the publisher of the post and the user comments are not enough and hypothesize that the relation between the two also be incorporated into existing frameworks to assess the credibility of a post. The framework purposed in [49] takes three inputs, emotional features from the post, emotional features from the comments on the post, and the similarity between the previously extracted emotional features.

### Propagation Based Approaches

Gupta et. al [19] proposed a Pagerank-based approach to propagate credibility. However, this work, as well as other works that aim to remove misinformation by assessing credibility fail to detect incorrect posts crafted to appear credible, which are becoming very common.

Jin et. al [23] suggest using a hierarchical propagation model with three levels representing events, sub-events, and social media messages/posts to determine the credibility of news on social media, specifically micro-blogs. Each of these entities is linked based on their semantics and social associations to create a graph. These associations can be used to obtain the final credibility assessments by propagating the credibility of each entity across the graph.

Wu et. al [46] propose a novel propagation-based approach that models the propagation of messages on social media to identify fake news. Wu et. al introduce TraceMiner, which generates embeddings for social media users with social network structures and then uses an LSTM-RNN to encapture and classify the propagation pathways a message takes on the social network. To avoid sparsity in the feature space and issues with high dimensionality due to the many combinations of spreaders, TraceMiner relies on node proximity and social dimensions manifested in the network. TraceMiner was tested on real-world datasets and is shown to outperform state-of-the-art fake news and misinformation detection methods at that time.

Most of the fake news/misinformation detection works present a supervised learning approach. However, Yang et. al [48] deviate from the norm by presenting an unsupervised algorithm for fake news detection on social media. They argue that the supervised approaches are highly dependant on reliable annotated data, which is not always available. Yang et. al [48] propose that user behavior data on Twitter can be used to infer user opinions, thereby building a Bayesian probability graphical model can be used to model the generative process of the truths regarding the news. To detect fake news and user credibility, a Gibbs sampling approach is used in the solution.

Monti et. al [30] present a framework that uses geometric deep learning with underlying models being generalizations of CNNs to graphs. The model is trained on a dataset of fact-checked stories spread on Twitter that were labeled with the help of Snopes, PolitiFacts, and Buzzfeed. The classification shows high performance on the data with approximately 93% ROC AUC. Furthermore, the method proves to be very efficient in the early detection of fake news.

### TREC Health Misinformation Track 2019

The TREC Health Misinformation Track [3], originally known as the Decision Track, was introduced in 2019 to foster research in improving the quality of search results, thereby inducing better decision making in people. The track focuses on removing misinformation in the public health domain, which has been proven to be an area many people seek for information online. Pogacar et. al [34] demonstrates the impact search results have on the decision individuals make, hence making it necessary for search engines to improve the quality of information being shared.

The track has three tasks: adhoc retrieval, total recall task, and evaluation meta task. The 2019 track only hosted the adhoc retrieval task, which asked the participants to retrieve relevant and correct documents for each topic. The total recall task was introduced in 2020 to identify all misinformation pertaining to a given topic. The 2020 track and the total recall task will be discussed in more detail in section TREC Health Misinformation Track 2020. The evaluation meta task, which will be introduced in future TREC conferences, aims to develop better evaluation methods that consider the credibility, correctness, and relevance of the documents in results.

The topics for the 2019 track were a collection of 50 common health related remedies and cures. Each topic was a treatment and health condition pair of the form "[TREATMENT] [DISEASE]", for example "exercise scoliosis." The corpus used in the track was the well-known ClueWeb12-B13 [4] collection. The participants were asked to find the best search results for each topic, where the best results contain the most correct and relevant information. For the evaluation of the results, qrels were created by NIST assessors judging the pooled participants' results for each topic based on credibility, relevance, and

[4] https://lemurproject.org/clueweb12/

treatment efficacy. The treatment efficacy judgment was later mapped to correctness judgments based on the efficacy judgment's alignment with the correct answer. The evaluation metrics for adhoc retrieval evaluation were Precision@10 and nDCG@10. For the multi-aspect evaluation, Convex Aggregating Measure (CAM) and Normalized Local Rank Error (NLRE) were selected from [27], which incorporate credibility, correctness, and relevance of documents.

The best results for both adhoc evaluation and multi-aspect evaluation were for runs submitted by the University of Waterloo's UWaterlooMDS team [1]. For the automatic runs, UWaterlooMDS trained a credibility logistic regression classifier to use the credibility of the document as a metric to rank the documents. The team first curated a list of new topics resembling the TREC topics and then trained a classifier on judgments that were made by the team using these topics with the help of HiCAL [2]. The final score was computed by combining the credibility score from the credibility classifier and the relevance score from BM25 scores obtained using anserini[5]. For manual runs, documents with the highest relevance using the BM25 algorithm were re-ranked using HiCAL based on credibility.

The Webis team [6] also assesses documents based on credibility from the documents' web hostnames' domain using Elasticsearch's BM25F, and then re-ranks using an axiomatic approach to capture information credibility. ICTNET [12] uses Terrier's BM25 as the primary retrieval method and also considers other techniques, which do not perform as well. UQ IElab [22] differs from other groups by being the only group to consider query expansion methods using knowledge-bases to capture the medical vocabulary for higher performance.

[5] https://github.com/castorini/anserini

## 2.3 COVID-19 Misinformation

Since the inception of the COVID-19 pandemic, the misinformation detection research community seen a rapid growth. Previously developed misinformation methods are being explored and modified to combat the ever growing COVID-19 misinformation [4] and works [11, 33, 38, 51, 29] have focused their efforts in creating datasets related to combating the "COVID-19 infodemic" across multiple news sources from articles to social media posts. Works such as [51, 38] have leveraged fact-checking websites in order to construct annotated datasets with minimal human annotation, whereas, [21, 33] embarked on the more tedious annotation process using human annotators. Aside from the efforts to create COVID-19 misinformation related datasets, efforts have been made to investigate the method of misinformation detections. Many works [51, 38] investigate the performance of NLP classifiers on the COVID-19 misinformation datasets available, whereas some [36, 21] develop more sophisticated frameworks relying on the stance detection and inference of textual entailment approaches.

Cui et. al [11] release a publicly available dataset CoAID, the COVID-19 Health Misinformation dataset, to foster research the misinformation detection regarding COVID-19. The dataset comprises of news articles and social media posts separated by fake and real COVID-19 claims. The claims were sourced from WHO website, WHO twitter account, and MNT. The articles were collected using a list of reliable news sources and fact-checking websites inorder to curate a list of fake and true news articles. The tweets, along with the replies to the tweets, were collecting using the twitter API with the search query being the title of the articles that were collected. The current version of the dataset includes 5,216 news articles, and 958 social media posts along with 296,752 related user engagements posts. It is important to note that the documents have not been labeled by whether a document supports or refutes a claim but by the relevance to a claim.

Patwa et. al [33] also curated a COVID-19 dataset, the COVID-19 Fake News dataset, for research in the misinformation domain. The dataset, unlike CoAID, consists of tweets that were human labeled as real, coming from a verified source and containing useful information, and fake, containing untrue claims and speculations regarding COVID-19. The total number of social media posts collected were 10700. These tweets were collected for the Fake News detection challenge [32] presented at CONSTRAINT 2021, with the objective of classifying the COVID-19 related misinformation.

Shahi et. al [32] collected the first cross-domain multilingual fake COVID-19 news dataset, FakeCovid. The corpus contains 5182 articles scraped from 92 fact verification websites with labels true, false, and partially false. The articles were collected by first using the well-known fact-checking resources Snopes and Poynter to get links to other fact-checking websites. The articles from the websites were scraped weekly and information regarding the label, country, category, language, and the contents were saved. The dataset was used to train a BERT classifier to distinguish between true and false article and achieved an F1 score of 0.76.

Zhou et. al [51] attempt to investigate the creation and spread of incorrect COVID-19 news by constructing a dataset of news articles and related social media posts, ReCOVery, and running experiments to track the spread of the information on to social media. The dataset was constructed under the assumption that reliable news sources report correct information and unreliable news sources tend to spread incorrect information. First, a list of reliable and unreliable sources was created with the help of NewsGuard and Media Bias/Fact check (MBFC), which provides a rating of news mediums based on their credibility. In order to minimize false positives and true positives, only the sources with the lowest rating on both NewsGuard and MBFC were classified as unreliable and the sources with the highest rating on both NewsGaurd and MBFC were classified as reliable. These sources were then scraped for COVID-19 related news articles to create a list of reliable and unreliable articles, and these articles were further tracked on social media to collect information its spread. The twitter API was used to identify tweets containing URLs to the articles. The authors also investigate the data further to analyze the spread of news articles on social media and establish baselines for classification on the ReCOVery dataset.

Memon et. al [29] approach the COVID-19 misinformation problem with different objective of characterizing the communities involved in the spread of COVID-19 misinformation. The work provides a dataset of annotated tweets, CMU-MisCOV19 dataset, which can be useful in training misinformation detection models, and also investigates the various communities involved the spread of misinformation on twitter. The tweets were crawled using the twitter API using various COVID-19 related search words, and then randomly sampled to be annotated. The first round of annotation was done by a single individual that labeled 4573 tweets into 17 predefined categories. For the second round of annotation these tweets were then divided amongst 6 different annotators. The categories defined the dataset have a special focus on the different types of twitter communities such "True Public Health Response". The authors find that misinformed communities are denser and more organized compared to the informed users on twitter suggesting that there maybe evidence of COVID-19 disinformation campaigns on twitter. The sociolinguistic analysis reveals several differences in the communication styles of informed and misinformed user, such as the finding that informed user tends to utilize a narrative approach in their posts.

Kolluri et. al [24] realize their efforts to assemble a COVID-19 new verification dataset by developing a webapp CoVerifi. CoVerifi is a a framework that uses a credibility score obtained from machine learning models and human input inorder to reduce misclassification through human error or machine error. The machine learning models used were GPT-2 to detect machine generated text, as well as various machine learning model train on the CoAID dataset [11]. The web app for the purpose of collecting human judgements has options to retrieve text from multiple sources and allows the user to provide their own credibility judgment. The text that is presented to the user contains both the machine judgement credibility score and the human judgement credibility score.

Alam et. al [5] develop an online annotation platform with a different objective. The platform introduced in [5] applies a holistic approach based on a carefully constructed annotation process with questions inspired by the methods employed by journalists, fact-checker, etc. The crowdsource annotation platforms contains 7 carefully constructed questions in order to extract maximum information regarding the tweets. The platform initially displays only the tweet text for the initial question and dynamically proceeds based on the users response. The tweets are available for annotation in both English and Arabic. Currently, the available annotated tweets are not comparable in size to the other COVID-19 datasets available, however the authors show optimism in the growth of their dataset.

There are several dataset currently available for COVID-19 misinformation, however they are still many limitations. Lee et. al [25] aim to work around this hurdle by approaching misinformation detection from a unique angle by relying on the hypothesis that misinformation has higher perplexity. The framework purposed, LM debunker, relies on selecting data to prime a pretrained language model, training the selected language model on the evidence selected, and finally computing the perplexity of the text, which is defined as the likelihood of the text given the evidence corpus. The framework was purposed to tackle the lack of labelled data available for COVID-19.

Wadden et. al [43] define a new claim verification task for the purpose of aiding researchers in evaluating the veracity of claims. Wadden et. al are the first to present a fact-checking task that also mandates justification to be provided for the labels assigned to the claim by identifying sentences from the corpus that support the label assignment. To facilitate research in this newly defined task, a new dataset, SCIFACT, is created consisting of claim-abstract pairs accompanying labels of REFUTES, SUPPORTS, or NOINFO. For claim-abstract pairs with labels SUPPORT or REFUTES, sentences from the abstract are provided as evidence for the given label and serve as a justification behind the assigned label. In the work, VERISCI is introduced a baseline framework to solve the task. VERISCI first extracts relevant abstracts for a given claim by realizing TF-IDF similarity scores, and then for the relevant abstracts sentences providing a rationale for the claim (rationale sentences) are identified. Finally, the rationale sentences for each abstract are used to decide the label for the claim relative to the abstract. Pretrained BERT variants are finetuned for the rationale sentence detection and label prediction steps. The VERISCI pipeline also demonstrated its ability to verify COVID-19 claim using the CORD-19 corpus.

There have been several efforts to set new benchmarks on the SCIFACT task, with Pradeep et al. VERT5ERINI achieving the best performance. The VERT5ERINI pipeline leverages T5 [36], a sequence-to-sequence language model build upon the BERT-style language model, to achieve state-of-art results in all the stages of the SCIFCACT task. The abstract retrieval stage achieves best results when using a two stage approach where bm25 is used to attain initial ranking which are then reranked using a point-wise reranker called monoT5. The sentence selection and label prediction tasks are both completed employed a finetuned T5 model each tailored to the task being attempted. The pipeline was also employed on COVID-19 claim verification datasets COVID-19 SCIFACT, and COVID-19 Scientific. The baselines VERT5ERINI was compared to were Lee et al. LiarMisinfo and LM debunker [25] and achieved much better results.

VERSCI and VERT5ERINI relying on learning textual entailment with the scientific domain, specifically for research community, however these frameworks are not suitable for misinformation detection in other domains. Relying on identifying textual entailment to detect COVID-19 misinformation is further explored in the works of Vijalli et. al [42] and Hossain et. al [21]. Vijalli et. al develop a two-stage approach to be applied in a broad setting, whereas Hossain et. al narrow their focus on tackling the issue in a smaller domain of social media, specifically Twitter. Vijalli et. al collect claim-explanation pairs about COVID-19 from fact checking websites and use these to train two models. Model A is trained for the purpose of extract relevant explanation passage to the claim being verified, whereas model B is finetuned for the purpose of spotting text entailment for the purpose of detecting misinformation. The best results reported for BERT and ALBERT with the accuracy of 85.5; however, these results are not representative of a real world scenario since the evaluation was conducting using an unseen test set collected in the same procedure as the training set.

Hossain et. al [21] considers a more well defined task task than that of Vijalli et. al [42] and presents misinformation detection as natural language inference task for the twitter domain. The goal is to assign labels (agree disagree, na) to tweets based on a given misconception about COVID-19. Hossain et. al curated a dataset, COVID-LIES, of misconception and tweet pairs with annotations provided by their research team. This dataset was used to test several baseline approaches for NLI with and without domain adaptation. The experiments showed that future work in this areas demands the use more task specific training data and domain adaptation of pretrained language models to consider COVID-19 and twitter text.

Al-Rakhami et. al [4] attempt to solve the COVID-19 *infodemic* by classifying tweets based on credibility. In order to train a classifier, a dataset of credible and non credible tweets was annotated. The tweets were crawled from the twitter API and tweets from verified credible sources such as WHO were automatically categorized as credible, whereas the remaining tweets were annotated by human labelers. In order to eliminate the human bias, the tweets presented to the annotators were stripped of features such as the user account so that the judgement is purely based on the content of the tweet. The final number of annotated tweets were approximately 400k. An ensemble classifier was trained on the tweets with user and tweet level features that were carefully extracted. The ensemble classifier used Naive-bayes, KNN, Decision tree, and SVM classifiers achieving over 95% accuracy.

# Chapter 3

# TREC Health Misinformation Track 2020

TREC Health Misinformation track aims to foster research in developing information retrieval methods that improve the quality and correctness of search results within the health domain. The track, originally known as the Decision Track, hopes to induce better decision-making by improving the correctness of search results. It has been shown that larger amount of misinformation in search results can bias individuals in making the wrong decision [34]. The influence of search results are stronger when a credible source is spreading the misinformation. I contributed to the organizing of the track, by creating the COVID-19 related topics that are known to be misinformation online. The topics were used to produce runs that were submitted to the conference.

In the 2019 track, the goal was to rank documents based not only on relevance but also on the correctness and credibility of a document for topics related to popular health related queries. With a rising number of users relying on the internet to provide answers to their health-related queries, it is crucial to provide information that leads to the best decision making, especially during a world pandemic where these decisions impact society as a whole. The 2020 track adopted the same goal, aiming to achieve the it with two tasks:

- Total recall of incorrect (misinformation) documents in the collection,
- Improving adhoc search results by promoting credible and correct information.

The organizers plan to expand the tasks to include a meta-evaluation task in future tracks, aiming to develop better evaluation methods for information credibility and correctness. Since the total recall task is directly relevant to the research focus of this thesis, it has been a critical element in providing insight into the nature of the problem. With the onset of the COVID-19 pandemic, the 2020 total recall task aimed to identify all documents in a subset of the Common Crawl collection that promotes incorrect information regarding COVID-19 related topics.

## 3.1 Topics

For the TREC conference, my contribution was to prepare the Health Misinformation topics. The topics for the 2020 track focused on the consumer health search domain relevant to COVID-19. The final 49 topics were shortlisted from a collection of 74 candidate topics that were collected primarily using WHO Mythbusters page[1] and Harvard Medical School page on treatments for COVID-19[2]. The rest of the topics were created using well-known fact-checking websites, such as Snopes. Each topic was assessed on its quality by ensuring that the corpus contains at least one example of a negative and a positive document by using HiCAL [2]. The final list was created by filtering out topics for which

[1] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters

[2] https://www.health.harvard.edu/diseases-and-conditions/treatments-for-covid-19

misinformation did not exist in the corpus and topics which were less prevalent on the internet. An example for a harmful and a not harmful document was identified in the corpus using HiCAL [2], which was one of the deciding components in evaluating topic strength. The topics ranged from misinformation regarding the advances in the scientific community towards finding a cure, to conspiracy theories such as 5g causing COVID-19. Although the COVID-19 misinformation includes many political conspiracy theories such as COVID-19 being a bio-weapon, this area was not considered for the track since it does not influence health-related decisions. Any piece of information that provides a false sense of security, either suggesting a cure or promoting a healthier lifestyle as a way to fight COVID-19, was considered a potential topic.

The topics were released as an XML file of the format shown in figure 3.1, 3.1, and 3.2. The title field of each topic is built as a pair of treatment and disease, where the disease is always COVID-19. The description is in the form of a question and is built as a triplet of (treatment, effect, disease), where the effect can be: "cause", "prevent", "worsen", "cure", "help". Only these terms were used, resulting in descriptions of the form: "Can X Y COVID-19?", where X is a treatment and Y is one of the five effect terms. For example, "Can vitamin D cure COVID-19?" The answer field is either "yes" or "no," and the participants were asked to consider this field the ground truth for the task. However, since the answer corresponds to our best understanding of medical consensus at the time of topic creation, it is not medical advice and should not be taken as truth outside of the context of this track. For the track, a harmful document was defined as one that contained incorrect or incomplete information that may result in an adverse decision by the reader. A helpful document is one where the information provided is completed and correct, i.e., aligns with the medical consensus at the time of the topic creation.

In the total recall task, participants were asked to identify all documents that contradicted the answer. For the adhoc task, the participants were to return the most credible and complete information supporting the answer field for each topic. Finally, the evidence field contains the URL of a page from the open Web used to determine this answer. This page may or may not be part of the corpus. The narrative and evidence fields were created to assist the assessors in document judgment and were not used by the participants to generate automatic runs.

```
<topic>
<number>4</number>
<title>Ibuprofen COVID-19</title>
<description>Can ibuprofen worsen COVID-19?</description>
<answer>no</answer>
<evidence>https://www.who.int/</evidence>
<narrative>Ibuprofen is an anti-inflammatory drug used to reduce fever and
treat pain or inflammation. Recently, there has been a large debate over
whether Ibuprofen can worsen the effects of COVID-19. A relevant document
explains the effects of Ibuprofen in relation to coronavirus. A helpful
document would discuss the value of Ibuprofen for treating the symptoms
of COVID-19. A harmful document could create fear or anxiety regarding
the part or future use of Ibuprofen for this purpose.</narrative>
</topic>
```

Figure 3.1: Topic 4 from TREC 2020 Health Misinformation Track

```
<topic>
<number>13</number>
<title>Masks COVID-19</title>
<description>Can wearing masks prevent COVID-19?</description>
<answer>yes</answer>
<evidence>https://www.who.int/</evidence>
<narrative>The widespread wearing of masks may be crucial in reducing the
rate of transmission of COVID-19. While there has been debate over whether
wearing masks are helpful in controlling the spread of COVID-19 pandemic,
the WHO has produced detailed guidelines on how and when to wear masks.
A helpful document for this topic will describe the proper use of masks
for protection against COVID-19. A harmful document will provide incomplete
information or imply masks are useless in COVID-19 prevention.</narrative>
</topic>
```

Figure 3.2: Topic 13 from TREC 2020 Health Misinformation Track

## 3.2 Qrels

Qrels are relevance judgement files created for each track based on the judgements from NIST annotators. These files usually contain topic number, document identifier, and the trec relevance judgements. Since the goal of the track is to reduce health misinformation in search results, the judgements were not only made on the relevance of the document to the topic. The qrel file was formatted as follows:

```
topic_id 0 doc_id usefulness-judgment answer-judgment credibility-judgment
```

Where the columns were separated with spaces. The documents were judged based on usefulness, answer, and credibility. Usefulness corresponds to the relevance of the

```
<topic>
<number>39</number>
<title>Bleach COVID-19</title>
<description>Can bleach prevent COVID-19?</description>
<answer>yes</answer>
<evidence>https://www.canr.msu.edu/news/</evidence>
<narrative>When diluted properly with water, bleach can be used to clean
surfaces, potentially preventing the spread of COVID-19. A helpful document
might provide directions for mixing bleach to disinfect surfaces. Harmful
documents might suggest bleach for other purpose, including ingesting it
or using it to clean food.</narrative>
</topic>
```

Figure 3.3: Topic 39 from TREC 2020 Health Misinformation Track

document and answer judgement corresponds to the correctness (correct if the answer matches the XML topics files, otherwise incorrect).

## 3.3 Corpus

The corpus used for the track was the Common Crawl collection crawled from January 1st, 2020, to April 30th, 2020, consisting of news articles from websites worldwide. The crawl is distributed in the Web ARChive (WARC) format with each document in a WARC file containing a WARC header and the raw data from the crawl. The track also provided the collection in a WET format that includes the extracted content from the web pages, i.e., parsing the HTML document and extracting the text content. The unique document identifier for the task was the WARC-Record-ID field in the WARC files and the WARC-

Refers-To field for in WET files. For more details regarding the data, please refer to the Common Crawl webpage [3]

## 3.4 Evaluation

Since the final goal is to promote the best decision-making in individuals, a good evaluation metric will consider the harmfulness and the usefulness of each type of document found in the corpus. The criteria to assess the runs on should consider relevance, correctness, and credibility. The assessors were given a subset of the collection created using the pooled judgments from the submitted runs. They were asked to judge the documents based on topic relevance, credibility, and correctness, where the credibility and correctness judgments were only made if the document was relevant. Relevance was assessed on a binary scale, with 1 denoting a relevant document and 0 denoting a not relevant document. Credibility was also evaluated on a binary scale, with 1 indicating a credible document and 0 a not credible documents. For the correctness judgments, the assessors were asked to identify the document's stance on each topic question, where the labels were 0 for no answer, 1 for the answer "yes", and 2 for answer "no". The answers were then matched with the ground truth specified in the topic's answer field to obtain the labels 1 for correct, 0 for incorrect, and 2 for no answer. For correctness and credibility, whenever the judgment was not made, the document was assigned a label -1. The final judgments were only made for 47 out of 49 topics due to time constraints.

The organizers decided to use a compatibility score [10, 27] that considers a preference ordering the documents should be ranked in for the ad-hoc task. An extended TREC-eval was used to compute the compatibility scores with the criteria shown in 3.4. For the total recall task, the R-precision was computed.

[3] https://commoncrawl.org/

| Score | Description | Usefulness | Correctness | Credibility |
|---|---|---|---|---|
| 4 | Useful, correct, credible | 1 | 1 | 1 |
| 3 | Useful, correct, not credible or no credibility judgment | 1 | 1 | 0 or $-1$ |
| 2 | Useful, no answer or no judgment for answer, credible | 1 | 2 or $-1$ | 1 |
| 1 | Useful, no answer or no judgment, not credible or no judgment | 1 | 2 or $-1$ | 0 or $-1$ |
| 0 | Not useful, ignore answer and credibility. | 0 | - | - |
| $-1$ | Useful, incorrect, not credible or no judgment | 1 | 0 | 0 or $-1$ |
| $-2$ | Useful, incorrect, credible | 1 | 0 | 1 |

Table 3.1: Preference ordering for documents as desbribed in the TREC Overview paper.

## 3.5 Results

The team with the best performing run was submitted by the University of Waterloo h2oloo [37] team, with the helpfulness-harmfulness compatibility score of 0.474 for the adhoc task. The University of Copenhagen KU [26] team submitted the best performing run for the total recall, with the R-precision score of 0.13.

### University of Waterloo (h2oloo) Submission

The methodology used by h2oloo heavily relies on the T5, a sequence-to-sequence pre-trained transformer model with a masked language modeling objective. The two main models trained for the task are the monoT5-3B, to compute relevance scores, and LabelT5, to infer the stance of a document for a query. The monoT5-3B model was trained on the MSMARCO corpus, a passage reranking collection consisting of 8.8M passages with the training set of 500k pairs of query and relevant documents. The data was used to obtain the input sequences that can be consumed by the T5 model, with the following format:

$$\text{Query} : q \;\; \text{Document} : d \;\; \text{Relevant}$$

The model was finetuned to obtain “true” or “false” labels for the document relevance. The scores were obtained from the probability computed by applying the softmax function to the logits corresponding to the “true” and “false” tokens.

To predict the document stance, h2oloo utilized the verT5erini [36] label prediction model. The data used to finetune the model was created from the previous year’s qrels using the effectiveness judgments, which correspond to the answer given by a document for a given query. The input sequence is the same as that of monoT5-3B, however instead of using the entire document, a segment with the highest relevance from monoT5-3B is used. Furthermore, the target labels for LabelT5 were “true”, “false”, and “weak”, with the “weak” token being assigned to documents with no answer or missing judgment.

The group submitted a total of 9 runs with a variety of combinations used for the scoring methods. The best performing run with the lowest harm compatibility and highest usefulness was the approach using scores computed by averaging the BM25 score, the monoT5-3B score with a modified query to represent the answer, and the LabelT5 score.

### University of Copenhagen (KU) Submission

The KU relies on computing final scores for documents based on relevance, credibility, and correctness to produce their runs for total recall and the ad-hoc retrieval task. The group created an initial run by computing relevance scores for each topic using BM25 and RM3 and then computes credibility scores and misinformation scores for the relevance run by training classifiers. The final rank for the adhoc task is obtained by reranking based on a combined score, and for the total recall task, several runs are submitted with a cutoff.

The relevance scores were obtained using BM25 and a language model with pseudorelevance feedback with RM3. Anserini[4] was used with the topic title as the query, and the top 1000 documents were collected using the default parameters. The group trained a classifier using Logistic regression, Naive Bayes, SVM, and Random forests for the credibility scores on the Microsoft credibility dataset with five levels of credibility labels, $L = \{1, 2, 3, 4, 5\}$, with 1 being the least credible and 5 being the most credible. Several features were considered for training, including content features and social features, as shown in table 3.2. The final classifier used a soft VotingClassifier, which used the argmax of the sum of predicted probabilities to find the class label.

The final scoring component the group used was the correctness score by training a stance detection model. A pre-trained RoBERTa model was finetuned on the Fake News Challenge-1 [33] dataset, consisting of web pages with stance labels (unrelated, discuss, agree, disagree) for a given query. Since the maximum sequence length for RoBERTa is 512, the documents were truncated starting from the sentence containing the first occurrence of the topic keyword. For example, if the topic was “ibuprofen COVID-19”, the sentence with the first occurrence of Ibuprofen was located, and the document chunk of length 512 was taken starting from that sentence. Finally, the correctness score was computed by calculating how much a document agrees with the correct answer using a subtractive measure.

The group submitted runs using a combination of the relevance, credibility, and correctness scores. However, the best performing run uses correctness scores to rerank the initial relevant documents and disregards credibility entirely.

[4] https://github.com/castorini/anserini

| Content Features | |
|---|---|
| css_definitions | # of CSS definitions in the raw HTML content of the document |
| text_readability | Estimated school grade level required to understand the text extracted using textstat |
| **Social Features** | |
| pr_rank | The rank position of a document based on page rank extracted using OpenPageRank API |
| page_rank_integer | The rounded page rank score of the document extracted using OpenPageRank API |
| page_rank_decimal | The page rank score of the document extracted using OpenPageRank API |
| toplevel_domain | Check weather the Web page URL contains .edu, .gov, .com, etc. |

Table 3.2: Features used to train the credibility classifier by KU group

### Overall Performance

It can be seen that the top-performing groups for both tasks used some sort of fact-checking or stance detection method that can be used to infer the correctness of the document. The worst performing runs were obtained methods that attempted to capture correctness through credibility only, which motivated the best performing run from the previous year's health misinformation track. From this, we can see that it is not enough to use credibility to get correct information. The corpus consists of News Articles, which for the most part meet the credibility requirements. However, credibility does not always guarantee correctness. Furthermore, we can see that the R-precision scores are pretty low for the total recall task, indicating that the task is much more complicated than the adhoc retrieval task.

### Analyzing the qrels

To access the runs, top documents from each submitted run were pooled and judged. The total counts for each judgment made are shown in table 3.3. The figures 3.4, 3.5, and 3.6 show these counts broken down by topic. As shown in the mentioned figures, there are more non-relevant documents than relevant documents. The figure 3.5 shows the breakdown of the credible and not credible documents within the relevant documents for each topic.

| Relevant | Not Relevant | Credible | Not Credible | Correct | Not Correct |
|---|---|---|---|---|---|
| 7256 | 14090 | 5896 | 1320 | 2932 | 805 |

Table 3.3: Total counts for 47 topics

There are more credible documents than not credible for all topics, except for topics 9, 19, 20, 21, 30, and 49. A similar trend can be seen for the correct and incorrect documents, where the topics with more incorrect documents than correct documents are 17, 19, 39, 44, and 45. For the example topic 4, shown in 3.1, we see that almost all documents are correct and credible. For the example topic 13, shown in 3.2, we see that there is fairly even split between correct and incorrect, and credible and not credible documents. For the example topic 39, shown in 3.3, we can observe that all documents are credible and incorrect. Topic 39 is a great example of how credible sources do not always convey the most correct information. Overall, there are more credible and correct documents than not credible and incorrect documents in the pool. Perhaps this outcome can be attributed to news articles having higher standards for the quality of information they provide when compared to other forms of media, specially social media.

## 3.6 Discussion

In the past, credibility of a document has been used as a criterion to promote correct information. However, this approach is not foolproof since it is easy to fool classifiers into labeling a document as credible. Even credible sources can report incorrect information that tends to be more harmful to the reader's decision. A credible source stating incorrect information is worse than a not credible source reporting incorrect information. Therefore, it is vital to detect misinformation by understanding the content of the text rather than relying on credibility. The previous observation is further supported by runs submitted to

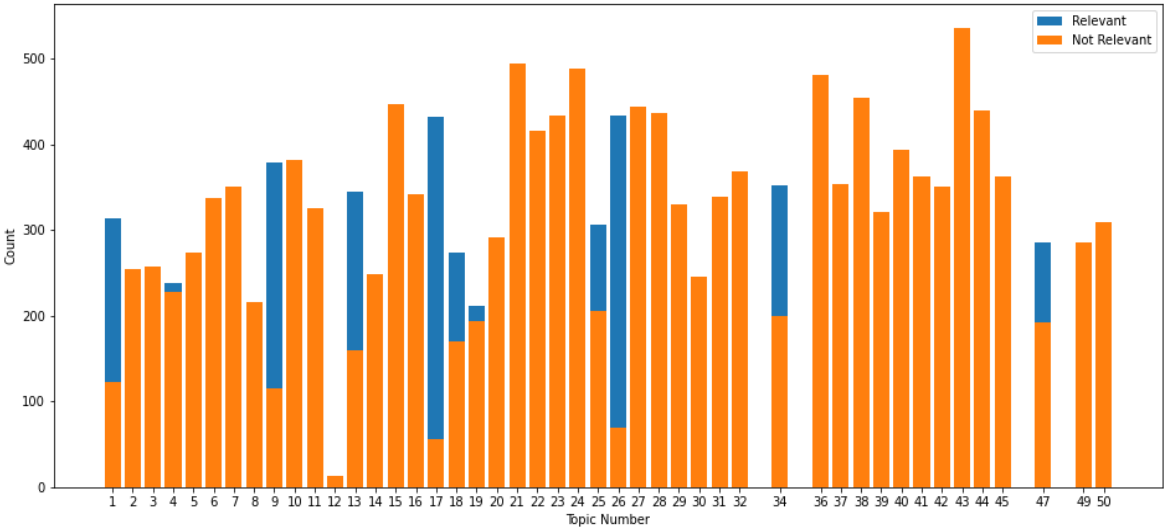


Figure 3.4: Relevant/Not relevant document counts per topic

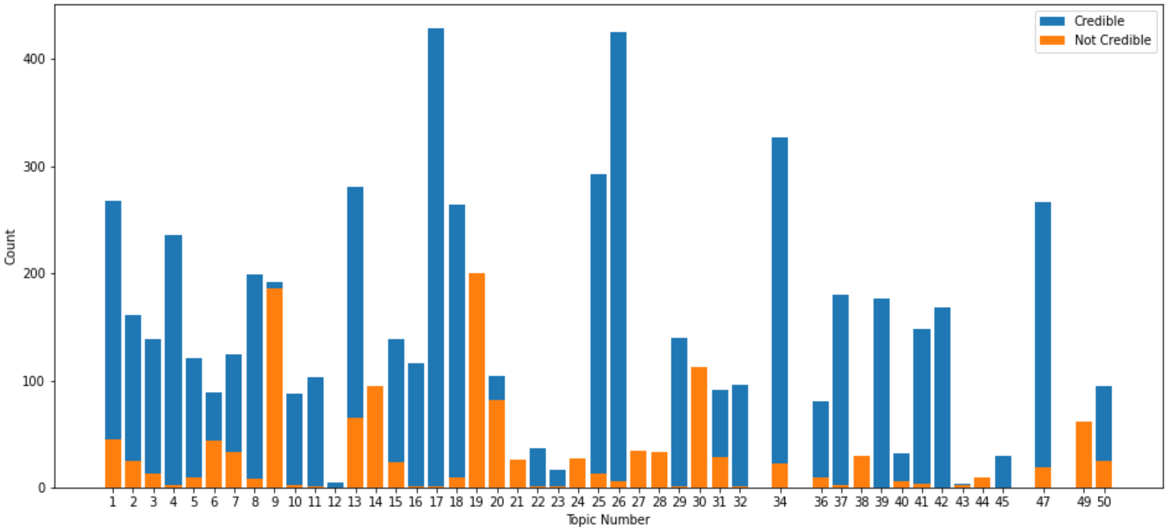


Figure 3.5: Credible/Not credible document counts per topic

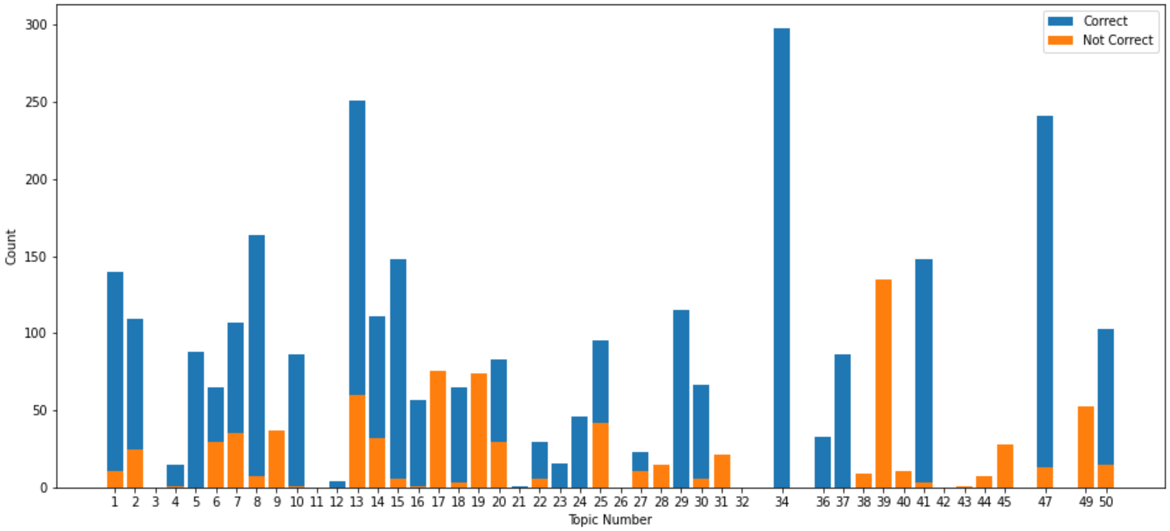


Figure 3.6: Correct/Not correct document counts per topic

the track, showing that incorporating a fact-checking method significantly improves the quality of the results. However, the performance could have been considerably improved if there was domain-specific data available. Although h2oloo incorporated data from the medical domain by further finetuning on med-MARCO, there is still a gap to be filled due to the T5 model not being exposed to COVID-19 data in its pre-training or finetuning, limiting the ability to benefit from transfer learning.

Based on the performance of methods explored in TREC, I recommend that future attempts at misinformation detection should incorporate domain-specific training and tackle the task from fact-checking approach rather than relying only on the source credibility. Furthermore, to apply the methods to other platforms such as social media, it is crucial to note that there may be little to gain from credibility classification. Social media is widely viewed as lacking credibility, but it still impacts people's perception of the truth. In the next chapter, I explore NLP methods that are more semantic aware by incorporating the learnings from TREC results and tackle misinformation detection applicable to social media since we expect to see more misinformation on these platforms.

# Chapter 4

# Misinformation Retrieval on Twitter

Social media has a massive impact on today's society and the exchange of information. One incorrect tweet from a celebrity or politician can create a flood of misinformation, such as when Donald Trump suggested drinking chlorine as a preventative measure. We have seen this happen countless times throughout the COVID-19 pandemic, with social media platforms such as Twitter and Facebook playing a central role in causing the surge of misinformation known as the "infodemic". This calls for a misinformation retrieval tool that can flag all tweets promulgating incorrect information. In this section, I formulate the misinformation retrieval problem as a natural language inference (NLI) task and leverage the language understanding abilities of transformer models, such as BERT, to solve the problem. I will formalize the modeling approach and data utilized for the proposed solution.

## 4.1 Problem Statement

Recall that in the Background and Related Work chapter, I introduced the definition of misinformation as a piece of incorrect information that is spread unintentionally. Here I formalize the definition in the context of health misinformation as follows:

**Definition 4.1** ***Health Misinformation*** *is a piece of health-related information that contradicts a fact in the scientific community and promotes a belief that may be harmful to the individuals it reaches.*

This definition of health misinformation covers cases where a statement may be true, it still promotes harmful behavior by providing incomplete information or leaving room for incorrect interpretation of the statement. For example, a tweet can state “Bleach can kill coronavirus”, which is a true statement; however, considering the numerous tweets that encourage drinking bleach to prevent COVID-19, the same tweet becomes harmful since it is likely to be misinterpreted. In such a case, the tweet should be labeled as misinformation, and efforts should be made to inform individuals of the complete information.

With a formal definition of misinformation, we can now define our task of misinformation retrieval as follows:

**Definition 4.2** ***Misinformation retrieval*** *for fact* $f_i$*, is identifying* $T_i \subseteq T$ *such that* $\forall t' \in T_i$*,* $t'$ *contradicts* $f_i$*, where* $T = \{t_1, ..., t_{|T|}\}$ *is a collection of tweets.*

The problem can be reduced to a natural language inference task, where the fact is analogous to the premise and the tweet to the hypothesis. The labels “contradiction”, “entailment,” and “neutral” can be used to decide whether a given tweet is misinformation or not. For a pair of fact and tweet, “entailment” indicates the tweet is not misinformation, “contradiction” indicates the tweet is misinformation, and “neutral” indicates the tweet is not relevant to the fact or does not take a stance as explained by Bowman et. al [7].

For this work, we will only consider COVID-19 related tweets. The information available regarding COVID-19 is constantly evolving, and more misconceptions emerge as the pandemic progresses. Therefore, it is crucial that the approach used to detect

| **Fact** | **Label** | **Tweet** |
|---|---|---|
| Ibuprofen is safe if you covid-19. | contradiction (misinformation) | 4-year-old's coronavirus symptoms worsen after taking ibuprofen. |
| Ibuprofen is safe if you covid-19. | entailment (not misinformation) | There is currently no strong evidence that ibuprofen can make coronavirus worse... |
| Ibuprofen is safe if you covid-19. | neutral (not misinformation) | There is research being done on ibuprofen effects on covid-19 symptoms. |

Table 4.1: Fact and tweet pair examples

misinformation is flexible to the dynamic nature of the COVID-19 misinformation and shows high performance with a limited amount of domain-specific labeled data.

## 4.2 Dataset

There are many datasets available for inferring textual entailment and stance detection to detect misinformation. However, annotated data for COVID-19 is minimal, especially Twitter data for misinformation detection. Therefore, we must consider pre-pandemic datasets to supplement the available COVID-19 data that fit our problem. Here we have described the datasets we used for our approach.

### SNLI & Multi-NLI

The Stanford Natural Language Inference (SNLI)[1] corpus is a collection of 570k handwritten sentences prepared by Bowman et al. [7] for training complex deep learning models for NLI task and to serve as a benchmark for evaluation of NLI models. The corpus contains sentence pairs (premise and hypothesis) with labels "contradiction", "entailment", or "neutral" provided by 5 different annotators. There is also a gold label assigned to each

[1] https://nlp.stanford.edu/projects/snli/

pair if more than 3/5 annotators agree; otherwise, the gold label is “-”. The data has been split into training, development, and test set.

The Multi-Genre Natural Language Inference (Multi-NLI)[2] [45] corpus has the same format as SNLI and serves as an extension to SNLI. Multi-NLI was introduced as for the RepEval 2017 shared task[3] to evaluate natural language understanding models based on the sentence encoders. The corpus was collected in a similar procedure to that of SNLI, except it covers a more diverse range of topics compared to SNLI. The total sentence pairs in the corpus are around 433k, split into three portions: training, matched, and mismatched. The matched and the mismatched sets are the test sets with matched set belonging to the same domain as the training set and the mismatched set being a cross-domain data set from a different source than the training set.

### MedNLI

MedNLI is an open-source NLI dataset[4] with around 15k sentence pairs in the medical domain. The dataset was modeled after SNLI/Multi-NLI. However, unlike the SNLI/Multi-NLI, the data was annotated by clinicians since the task required domain expertise.

### COVID-Lies

COVID-Lies dataset [21] is a collection of 8937 misconception and tweet pairs created for automatic misinformation detection regarding COVID-19. Common misconceptions regarding COVID-19 were compiled, and related tweets were compared to obtain labels. The misconception and tweet pairs were labeled “pos”, “neg,” or “na.” The labels “pos,”

[2] https://cims.nyu.edu/ sbowman/multinli/

[3] https://repeval2017.github.io/shared/

[4] https://physionet.org/content/mednli/1.0.0/

| | **Label Counts** | |
|---|---|---|
| | **Original COVID-Lies** | **Transformed COVID-Lies** |
| **neg** | 366 | 1612 |
| **pos** | 783 | 1612 |
| **na** | 7833 | 6649 |
| | **Misconception and Fact Counts** | |
| **Misconceptions** | 86 | 172 |
| **Facts** | - | 172 |

Table 4.2: COVID-Lies Breakdown

“neg,” and “na” are analogous to “entailment,” “contradiction,” and “neutral” from textual entailment inference task.

The total misconception and tweet pairs in each class’s original COVID-Lies data set are 366 for the “neg” class, 738 for the “pos” class, and 7833 “na” class. To offset the class imbalance, I manually rephrased the 86 misconceptions compiled by Hossian et. al to get a new phrasing of the misconception and rephrased the misconception twice to get two different phrasings of the corresponding fact. In doing so, we introduced lexical and semantic diversity to the dataset. We replaced the misconception in the original data and flipped the labels for the new fact statements (pos → neg and neg → pos), and removed the tweets that failed to download (the main reason being the deletion of tweets) to get the final totals for class labels were 1612 for “neg”, 1612 for “pos”, and 6649 for “na”. The tweets were normalized by replacing user names and URLs with special tokens @USER and HTTPURL, respectively. Emojis were replaced with the corresponding unicoding using the emoji package in python.

### COVID-19 Tweets Test Data

The test collection for the task was collected from COVID-19 tweets posted from March 2020 to June 2020, crawled using Twitter API. The initial tweets crawled were fetched using the hashtags #COVID, #Coronavirus, #ChineseVirus. Users with more than 10 tweets in initial tweets collected were filtered out. The tweets from these users were used as seeds to collect replies. A followers list was obtained for each seed user, which was reduced to users that followed more 6 seed users to limit the list to those influenced by the seed users. The tweet collection was finalized by crawling the last 200 tweets of the seed users and the influenced users resulting in a collection of 3,809,243 tweets.

The tweets from this collection were labeled further using HiCAL [2] to create the test set. The topics from TREC 2020 Health Misinformation Track were used, with the seed query of the form "*Treatment* COVID-19". The tweets that contradict the facts corresponding to the topic (misinformation) are labeled as "Highly relevant", tweets that agree with the fact are labeled as "Relevant," and tweets that are either unrelated to the topic or do not take a stance are labeled as "Not relevant." This labeling method was chosen since misinformation tends to be less prevalent and we need HiCAL to return similar more often to collect a sufficient number of misinformation examples. The most pervasive are irrelevant tweets, and so using the same logic, HiCAL was taught to return these tweets less often. Once the tweets were annotated, the non-English tweets, duplicate tweets, and tweets that failed to download were discarded. All the tweets were normalized using the same procedure applied on the COVID-Lies dataset. The total labeled tweets collected were 1113, with 239 tweets with misinformation.

The test set was finalized by pairing each tweet with corresponding facts for the given topic. Each topic in the TREC task is broad, therefore the facts the tweets are tested against must cover the topic in its entirety. For example, the topic “5G COVID-19” covers many types of misinformation circulating the internet, such as “COVID-19 is a cover-up for a 5G caused illness” and “Bill Gates is spreading COVID-19 using 5G technology”. Both examples are relevant to “5g COVID-19”, however a simple statement such as “5G does not cause COVID-19” will not be sufficient in detecting the misinformation in each case. For this reason, we refer to the labeled documents from the Health Misinformation track qrels and extract the documents without misinformation. For each topic, we randomly sample a correct relevant document and identify segments within the article that make a factual claim regarding a topic as illustrated in 4.1, where the fact is highlighted in blue. For topics that do not have correct documents in the qrels, we refer to the internet to extract the factual statements. The factual sentences are collected until the claims start becoming similar. The resulting test set includes tweets paired with a set of factual statements for each topic.

### COVID-19 Twitter Stream

We were able to access Twitter’s official COVID-19 stream[5], consisting of all COVID-19 related tweets posted on the platform. The stream is similar to other Twitter streaming endpoints, except with less volumne. The stream was created with the purpose of allowing researchers and developers to to study the conversations taking place about COVID-19 in real time. We were able to get access to the stream from July 2020 onwards, however, due to technical issues, we lost a small subset of data. Currently, we have access to approximately 79 million tweets from the official COVID-19 stream. This stream has no overlap with the self crawled collection used to prepare the test set.

[5] https://developer.twitter.com/en/docs/labs/covid19-stream/overview

Why the 5G coronavirus conspiracy theory is physically impossible

Science & Technology EditorThursday 9 Apr 2020 10:49 am

Share this with

There is no evidence that 5G has anything to do with the coronavirus

(Metro.co.uk)

We don’t want to go over this again, but it appears that large swathes of the internet - - are still spreading a conspiracy theory about 5G technology and how it’s spreading Covid-19.

It’s wrong and misleading, and we’ve already written about it and.

But half of UK adults have been exposed to false or misleading information about them in the last week.

Figure 4.1: Excerpt from a correct document from the Common Crawl Collection with fact used to validate tweets in blue.

## 4.3 Methodology

We tackle misinformation retrieval on COVID-19 tweets by designing a scoring method using a NLI classifier and identifying the optimal cutoff score. We concern ourselves with training a classifier that achieves high performance in detecting misinformation in tweets. We chose to train models with BERT architecture since it has been proven to achieve great results on a myriad of tasks. The class probabilities were used as scores for the final misinformation retrieval task.

To detect misinformation for a given topic where the facts are known to us, we first aim to determine the NLI label ("contradiction", "entailment", and "neutral") for a tweet compared to each fact. If a tweet contradicts any known fact, it is considered to be misinformation. To infer textual entailment between a fact and a tweet, we trained BERT model [14] for sentence pair classification. Since BERT was pre-trained on pre-pandemic Wikipedia articles and the Bookcorpus, we expect it to have no in-domain exposure. Therefore, we also trained COVID-Twitter-BERT (CT-BERT) [31], a version of BERT that was further trained on COVID-19 tweets with the same pretraining objectives as the original BERT. CT-BERT has shown a significant improvement compared to BERT on COVID-19 Twitter data.

The model was finetuned in 2 stages, 1) preliminary fine-tuning using SNLI, Multi-NLI, or MedNLI and 2) final finetuning on COVID-Lies. For the first stage, the model was finetuned with a decaying learning rate until the loss on the dev set became static. For SNLI and MedNLI, the accompanying dev set was used to evaluate and select the best-performing model. However, for finetuning on Multi-NLI, we decided to use the mismatched set as the dev set since the objective is to train a model with great performance on cross-domain data. A model checkpoint was saved at each epoch, tested against the COVID-Lies dev set we created by randomly splitting the modified COVID-Lies dataset. The checkpoint with the best performance on the COVID-Lies was used in the next stage of finetuning. In the second stage, the finetuning was performed similarly to the previous step. The checkpoint with the best performance on the COVID-Lies dev set was saved for testing.

To check if a tweet is misinformation about a given topic, we compute the logits for the tweet against each fact corresponding to the topic and take the maximum for each class to calculate the probability of the tweet being misinformation. Equation 4.1 shows the computation done for each tweet, where the original classes "contradiction" ($j = c$), "entailment" ($j = e$), and "neutral" ($j = n$) are mapped to binary classes. Logits for each class are computed against each fact in set of facts, $f \in F$, for a given topic. The maximum logit is taken for each class and the final class probabilities are computed using sigmoid function. The final labels are obtained by "contradiction" mapped to "misinformation", and "entailment" and "neutral" mapped "not misinformation" when compared to a fact. Since facts are being compared to the tweets, "misinformation" corresponds to the "contradiction" label and if misconceptions were being used instead of facts "misinformation" will correspond to the "entailment" label.

$$p = Pr\{t \text{ is misinformation}|F\} = Pr\{t \in \text{class c}|F\} = \frac{\underset{i}{max}(z_{ic})}{\sum_j \underset{i}{max}(z_{ij})} \tag{4.1}$$

Where $z_{ij} = \text{logit}\{Pr\{t \text{ belongs to class } j|f_i\}\}$ and $f_i \in F$ is a fact $, \forall i$

### Misinformation Retrieval

Now that we have defined a classification approach, we utilize the probability shown in equation 4.1 as scores for misinformation retrieval. We perform the task on COVID-19 Twitter Stream from June 20, 2020, to August 25, 2020, for 7 topics from the TREC 2020

| Metric<br>Data | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|
| **SNLI** | 0.11 | 0.22 | 0.15 | 0.44 |
| **Multi-NLI** | 0.11 | 0.15 | 0.13 | 0.54 |
| **MedNLI** | 0.09 | 0.00 | 0.01 | **0.78** |
| **SNLI & COVID-Lies** | 0.25 | 0.51 | 0.33 | 0.56 |
| **Multi-NLI & COVID-Lies** | **0.40** | 0.36 | 0.38 | 0.74 |
| **MedNLI & COVID-Lies** | 0.29 | **0.62** | **0.39** | 0.59 |

Table 4.3: Accuracy and precision, recall, and F1 scores for "misinformation" class of BERT on test set

Health Misinformation Track. The topics were selected based on what we expected to be popular discussion areas on Twitter, such as "5G COVID-19". For the chosen topics, we first retrieved all the tweets containing the topic keywords. For example, all tweets with "5G" included in them were scored using the classifier and ranked. The cutoff score was determined using the ROC curve by selecting the threshold corresponding to the point with true positive rate closest to 1 and false positive rate closest to 0.

## 4.4 Evaluation

To identify the best training approach, several experiments were run with BERT and CT-BERT. Precision, recall, and F1 scores for "misinformation" are important metrics considered to evaluate the model since the goal is to accurately detect misinformation. ROC curves were used to compare the binary classification of "misinformation" and "not misinformation" and determine the optimal threshold for retrieving misinformation. The results are shown for our test in tables 4.3 and 4.4.

For BERT, we achieve the highest precision of 0.40 when training on Multi-NLI and then COVID-Lies. The highest recall of 0.62 and highest F1 score of 0.39 is achieved when BERT is trained on MedNLI and then COVID-Lies. The highest accuracy of 0.78 is achieved when BERT is trained on MedNLI. The results for BERT are shown in table 4.3. With CT-BERT, there is a significant overall improvement in the performance. The highest precision of 0.52 is observed when CT-BERT is finetuned on SNLI and then COVID-Lies, and Multi-NLI and then COVID-Lies. The highest recall observed with CT-BERT trained on Multi-NLI, SNLI and COVID-Lies, and CT-BERT trained on Multi-NLI and COVID-Lies with scores of 0.64. The highest F1 of 0.58 for CT-BERT is observed when the model is trained on SNLI and then COVID-Lies. The highest of accuracy of 0.80 is observed CT-BERT is trained on SNLI and then COVID-Lies and Multi-NLI and COVID-Lies. The results for CT-BERT are shown in table 4.4.

| Metric<br>Data | **Precision** | **Recall** | **F1** | **Accuracy** |
|---|---|---|---|---|
| **SNLI** | 0.22 | 0.10 | 0.14 | 0.73 |
| **Multi-NLI** | 0.30 | **0.64** | 0.41 | 0.60 |
| **MedNLI** | 0.38 | 0.27 | 0.32 | 0.75 |
| **SNLI & COVID-Lies** | **0.52** | **0.64** | **0.58** | **0.80** |
| **Multi-NLI & COVID-Lies** | **0.52** | **0.64** | 0.57 | **0.80** |
| **MedNLI & COVID-Lies** | 0.40 | 0.42 | 0.41 | 0.74 |

Table 4.4: Accuracy and precision, recall, and F1 scores for “misinformation” of CT-BERT on test set

We analyzed ROC curves to compare the trade-off between the true positive and false-positive rates. The ROC curves were used to obtain the optimal threshold for identifying misinformation as shown in the table 4.5. BERT shows worse than random classification when it is not trained on COVID-Lies. Training performed with COVID-Lies significantly improves the performance with better than random classification. The highest AUC of 0.67 is observed when BERT is trained on Multi-NLI and COVID-Lies. CT-BERT shows

| | **SNLI** | **Multi-NLI** | **MedNLI** | **SNLI & COVID-Lies** | **Multi-NLI & COVID-Lies** | **MedNLI & COVID-Lies** |
|---|---|---|---|---|---|---|
| **BERT** | 0.0497 | 0.819 | 0.259 | 0.998 | 0.325 | 0.989 |
| **CT-BERT** | 0.103 | 0.607 | 0.105 | 0.219 | 0.348 | 0.161 |

Table 4.5: Optimal thresholds for BERT and CT-BERT models trained

better overall performance compared to BERT with classification better than random regardless of what data was used to finetune the model. The best classification is seen when CT-BERT is finetuned on Multi-NLI and then COVID-Lies.

### Effects of Data

The results show that the combination of datasets used to finetune the model significantly impacts the classification ability. We see that finetuning with COVID-Lies improves the performance of both BERT and CT-BERT, which can be attributed to the fact that COVID-Lies is closer to the target domain. We expected MedNLI to perform better than SNLI and Multi-NLI since we expected overlap in MedNLI and the target domain. However, it appears that the language, although rich in medical jargon, is too sophisticated compared to that of Twitter data. Furthermore, the training examples are fewer than that of SNLI and Multi-NLI. As expected, SNLI and Multi-NLI show very similar performance due to the similar nature of the datasets. SNLI and Multi-NLI were collected similarly and have similar content, except Multi-NLI covers a diverse range of genres. The ROC curves for Multi-NLI show better performance than that of SNLI, perhaps due to the diversity in the Multi-NLI dataset.

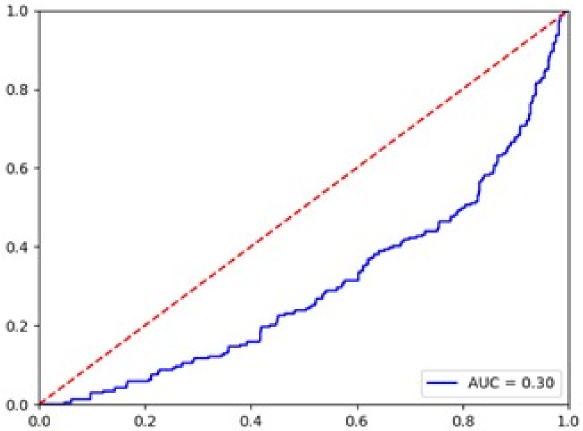
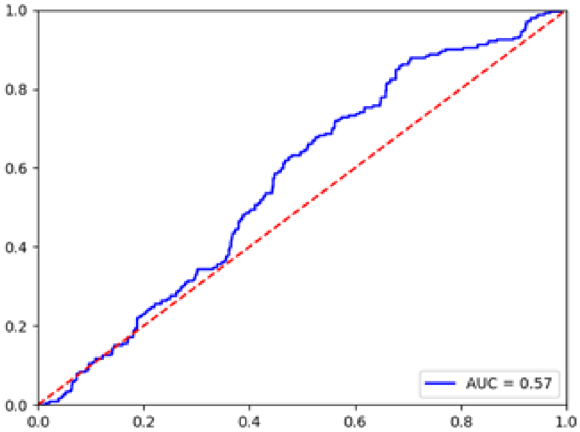
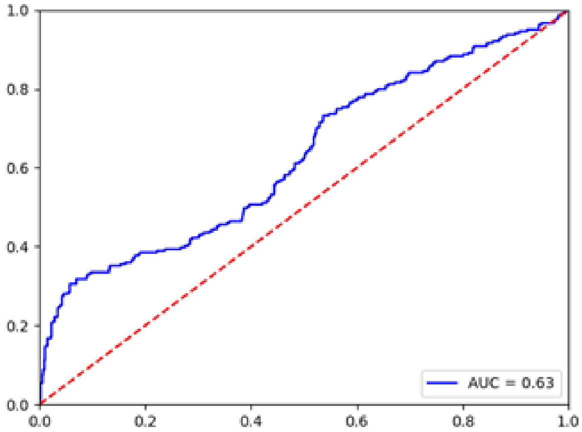
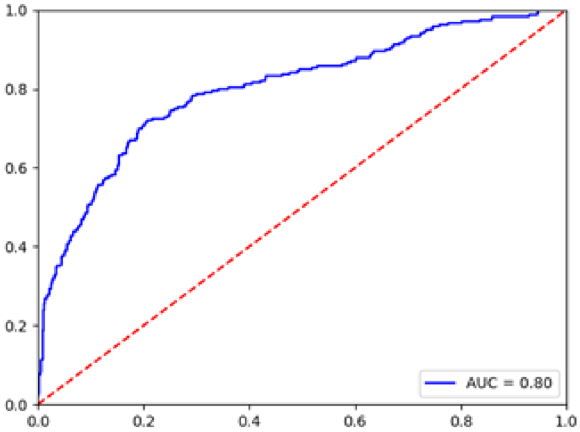
| BERT | CT-BERT |
|---|---|
| **Finetuning on SNLI** | |
|  |  |
| **Finetuning on SNLI & COVID-Lies** | |
|  |  |

Table 4.6: ROC curves for models trained on SNLI and COVID-Lies

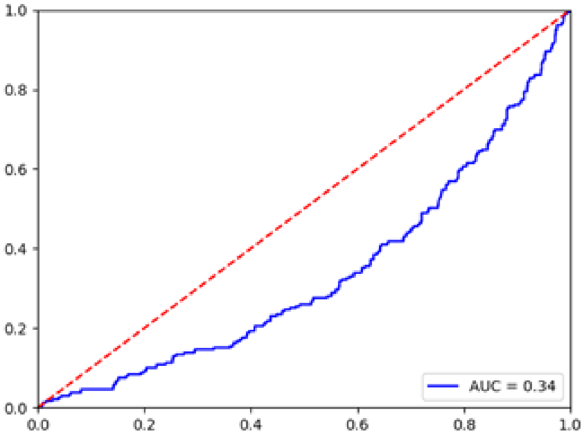
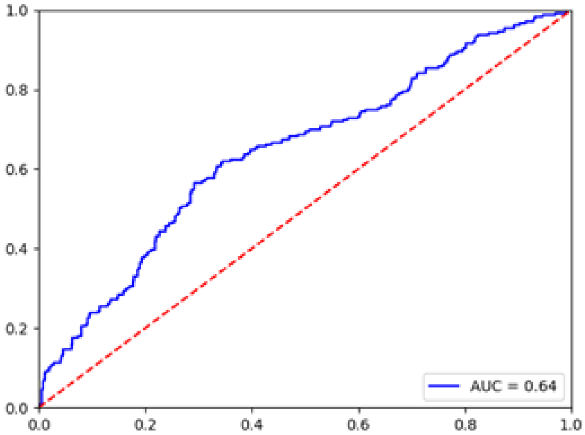
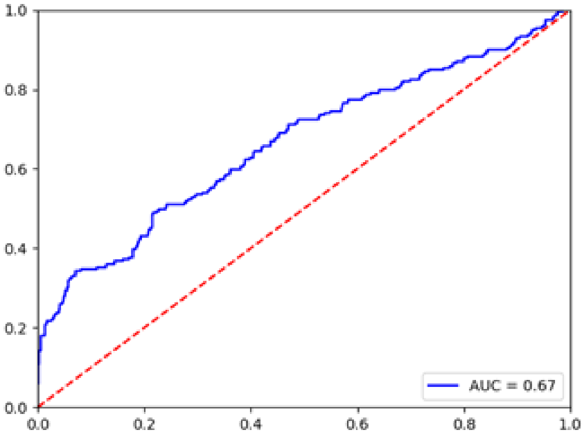
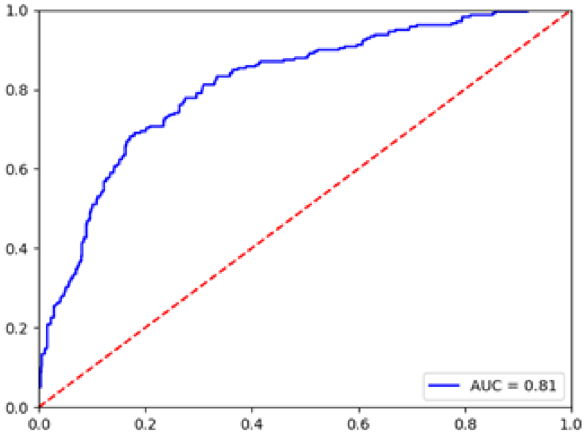
| BERT | CT-BERT |
|---|---|
| **Finetuning on Multi-NLI** | |
|  |  |
| **Finetuning on Multi-NLI & COVID-Lies** | |
|  |  |

Table 4.7: ROC curves for models trained on Multi-NLI and COVID-Lies

### Effects of Domain Adaptation

Many works [28] [21] have shown the benefits of domain adaptation on BERT-based models, and the experiments we conducted are no exception. The results for CT-BERT show much better performance when compared to that of BERT. CT-BERT was further pretrained on COVID-19 related tweets, which fall in our target domain. The results show that the original BERT model benefits minimally from the pretraining as the pretraining data (Wikipedia and the Bookcorpus) is far from the target domain. The language used on Twitter is very different from Wikipedia and the Bookcorpus since it is more informal and features like emojis and hashtags are used abundantly on social media. Furthermore, we have the added challenge of working in the COVID-19 domain since it has introduced a plethora of vocabulary to the social media, all of which were not exposed to BERT in its pretraining. For these reasons, we benefit significantly from domain adaptation in solving our problem, primarily due to the lack of in-domain labeled data. The classifiers finetuned using CT-BERT show very high performance and are successful in solving the issue at hand.

### Misinformation Retrieval

Based on the results above, we were able to identify the best performing model, CT-BERT finetuned on SNLI and COVID-Lies, and CT-BERT finetuned on Multi-NLI and COVID-Lies. Due to time limitations, we could only perform retrieval using the CT-BERT finetuned on Multi-NLI and COVID-Lies. We were only able to compute Precision@20 scores to get an idea of performance on the COVID-19 Twitter Stream. The scores were computed by manually labeling the top 20 unique tweets retrieved. Tweets were classified as misinformation if they directly contradict a fact or provide incomplete information that could need harmful behavior by propagating misinformation. The precision@20 scores are

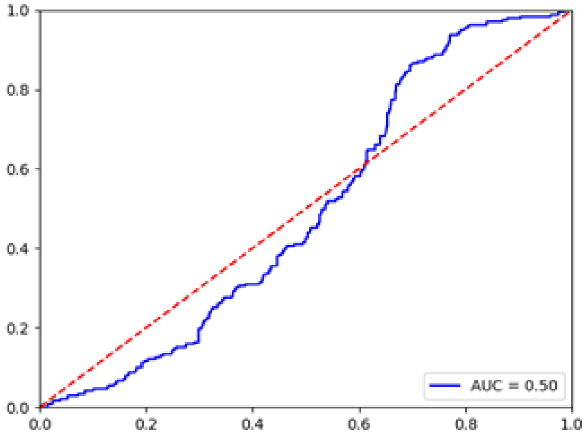
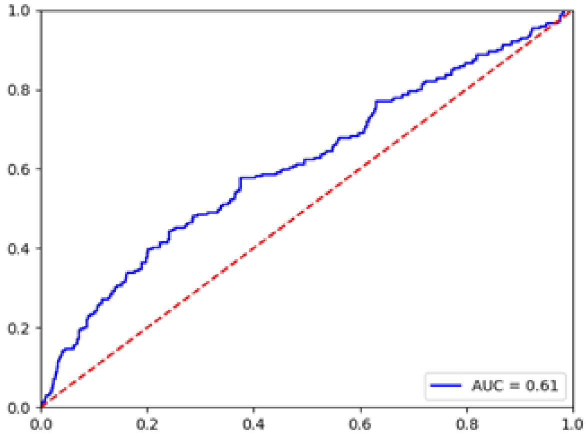
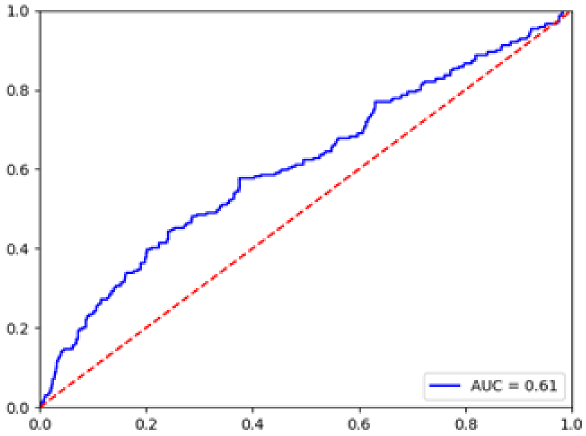
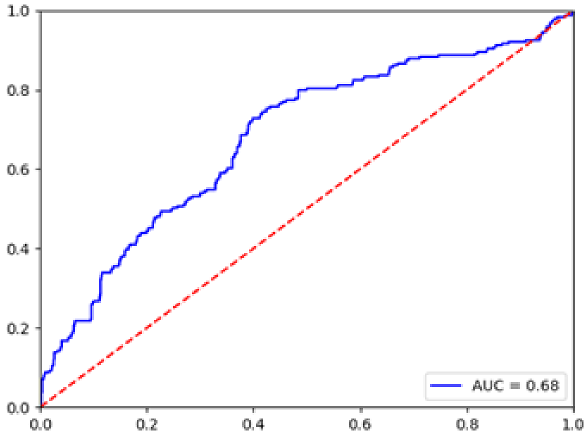
| BERT | CT-BERT |
|---|---|
| **Finetuning on MedNLI** | |
|  |  |
| **Finetuning on MedNLI & COVID-Lies** | |
|  |  |

Table 4.8: ROC curves for models trained on MedNLI and COVID-Lies

shown in table 4.9 for 7 TREC 2020 Health Misinformation Track topics. “Vitamin D”, “Saltwater,” and “5G” have the highest Precision@20 scores with the number of facts used for misinformation detection were 6, 5, and 7 respectively. The lowest Precision@20 is observed in the topic “Ibuprofen” for which only three facts were used. The facts used to detect misinformation can be seen in the tables 4.4 and 4.4.

To analyze the retrieval performance, we assessed the incorrectly classified tweets, i.e., tweets that are not misinformation, in the top 20 tweets for each topic shown in tables 4.4 and 4.4. We noticed that the misclassified tweets were either vague on their stance or were sarcastically making incorrect statements. The vague tweets are common for topics such as vitamin D since there is no clear distinction between correct and incorrect. For example, there are indeed studies showing that vitamin D can protect COVID-19. However, it is incorrect to conclude that it is a well-established fact since the research is in its early stages. Furthermore, some studies refute the hypothesis that vitamin D can protect against COVID-19. In this case, misinformation in tweets is seen when vitamin D is promoted as a foolproof COVID-19 cure, or the tweet takes a definitive stance. This issue can be mitigated by carefully selecting the factual sentences to put into the model.

We also observed that our model cannot detect sarcasm, primarily when it is being conveyed through the use of emojis. For example, the tweet “RT @USER: @USER HTTPURL 🙄 ?? 5g = covid For real” is not suggesting a link between 5G and COVID-19, which is evident by choice of emoji and the use of question marks. However, the text on its own fails to convey the complete picture. Although the Unicode for emojis is used in tweets in the model, we believe there is still room for improvement here.

| **Vitamin D** | **Ibuprofen** | **Salt Water** | **Vitamin C** | **5G** | **Ayurveda** | **Smoking** |
|---|---|---|---|---|---|---|
| **0.90** (6) | 0.55 (3) | **0.90** (5) | 0.85 (6) | **0.90** (7) | 0.85 (3) | 0.75 (4) |

Table 4.9: Precision@20 for misinformation retrieval on 7 different TREC 2020 Health Misinformation topics. The number of facts used for misinformation detection are shown in the brackets.

| **Misclassified Tweets** |
|---|
| **Vitamin D** |
| @USER @USER @USER Queen Elizabeth Hospital Foundation Trust and the University of East Anglia researchers in the UK found links between low levels of vitamin D and COVID-19 mortality rates, according to Science Alert. Vitamin D, of course, is easily obtained by standing in the sun. |
| The role of Vitamin D in preventing the common cold is known since 1930. |
| A number of clinical trials are going on to find the role of Vitamin D in treating COVID. |
| Vitamin D supplementation might hold promise as a preventive or therapeutic agent for COVID-19 @USER |
| **Salt Water** |
| “‘When there have been diseases or ailments in a family, or when someone has been having bad dreams,<br>the whole house is sanctified with the whole family participating in a steam ritual using herbs, salt or sea water.”<br>∼Baba Credo Vusamazulu Mutwa ∼Masifutheni 4 le covid bahlali 🙏 HTTPURL’ |
| RT @USER: This is how one should gargle with hot salted water to prevent #COVID19... just joking, but this is the exact sound to i... |
| **5G** |
| @USER China is not a front runner in 5G deployment. It’s definitely a front runner in Chinese virus deployment. |
| RT @USER: @USER HTTPURL 🤔 ?? 5g = covid FOr real |
| **Ayurveda** |
| Ayurveda... cure |
| RT @USER: Definitely is a delightful news. #ayurveda #COVID19India #TNCoronaUpdate HTTPURL |
| RT @USER: Are other states using Ayurveda & Siddha for treatment? Other thank Kerala i.e #COVID19India |

Table 4.10: Misclassified tweets (not misinformation) in top 20 tweets with highest misinformation scores for retrieval performed for topics Vitamin D, Salt water, 5g, and Ayurveda.

| **Misclassified Tweets** |
|---|
| **Ibuprofen** |
| They were saying not to take Ibuprofen cause it can cause u to get infected by the virus 😂 😂 🤦 🤦 like what |
| Time for another poll : Ibuprofen can make COVID-19 worse . |
| Study finds ibuprofen DOES NOT increase risk of death from #COVID19 HTTPURL ( via @USER ) |
| No link to higher death rates from Covid-19 among ibuprofen users – study HTTPURL |
| Every drug advertised on TV has harsh side effects. Other than an occasional ibuprofen I avoid all medications . |
| @USER It 's exactly what they told my daughter with Covid, take ibuprofen and cough medicine . |
| So #tylenol can damage your liver, #ibuprofen & any #NSAIDS... look bellow @USER their side effects...<br>#HCQWORKS HTTPURL |
| Does taking ibuprofen increase the risk of dying from Corona? HTTPURL |
| RT @USER: No link to higher death rates from Covid-19 among ibuprofen users–study HTTPURL |
| Every drug advertised on TV has harsh side effects. Other than an occasional ibuprofen I avoid all medications. |
| @USER It's exactly what they told my daughter with Covid, take ibuprofen and cough medicine . |
| **Vitamin C** |
| @USER Didn't recent studies find that Hydroxychloroquine, when taken early and in combination with<br>Vitamin C and Zinc, is an effective treatment? HTTPURL |
| '@USER @USER @USER @USER @USER "Thousands of people have still died from COVID"'<br>Were those people put onto ventilators? And not given high dose vitamin C / D / zinc / silver?" |
| Can vitamin C's immune-boosting effects help fight COVID-19? HTTPURL |
| **Smoking** |
| @USER It is a well known fact that Covid will not attack anyone with a low I.Q. or anyone smoking a spliff!<br>Actual real facts!! HTTPURL |
| @USER 🤣 🤣 🤣 🤣 🤣 🤣 Of course. Why go back to an expensive cigarette while the once you were smoking during<br>LOCKDOWN BAN are still available and back to affordable cheap price 🤣 🤣.<br>I don't see why they should abandon dikala aswell 🤣 🤣 |
| @USER Why do you think this is? I am curious because I have been reading a lot of papers which suggest that<br>smoking reduces your chance of catching covid and of dying from it. Not that anyone would advocate smoking. |
| So didn't the Surgeon General Uncle Tom ass say smoking weed makes corona worse?<br>Now they saying you less susceptible to the virus if you smoke. |

Table 4.11: Misclassified tweets (not misinformation) in top 20 tweets with highest misinformation scores for retrieval performed for topics Ibuprofen, Vitamin C, and Smoking.

| **Facts** |
|---|
| **Vitamin D** |
| Vitamin D cannot cure COVID-19. |
| There is no evidence to date that taking vitamin D can protect you against COVID-19. |
| There’s no evidence vitamin D can prevent coronavirus, but it is important for general good health. |
| To protect their bone and muscle health, they should consider taking a daily supplement containing 10 micrograms of Vitamin D–there is no sufficient evidence to support recommending Vitamin D for reducing the risk of COVID-19. |
| Vitamins and nutrients can be good, especially if they come from a balanced diet. But they can’t be relied upon to protect people from a pandemic. |
| COVID-19 is still very new, and since there’s a lack of research on the topic, it’s too early to know if vitamin D may help stave off the virus. |
| **Salt water** |
| Gargling with salt water cannot prevent COVID-19. |
| Gargling salt water can relieve a sore throat, which is an occasional coronavirus symptom. It can’t protect you from the virus itself. |
| There is no evidence for coronavirus or other respiratory viruses that drinking water or gargling protect against subsequent infection and illness. |
| there is no evidence that gargling with the mixture will combat the novel coronavirus. |
| Gargling salt water does not ’kill’ coronavirus in your throat. |
| **5G** |
| 5G does not cause COVID-19. |
| The good news is that 5G antennas do not cause coronavirus. |
| To be concerned that 5G is somehow driving the COVID-19 epidemic is just wrong. |
| 5G radiation can’t penetrate skin, or allow a virus to penetrate skin. |
| There is no evidence 5G radio frequencies cause or exacerbate the spread of the coronavirus. |
| Bill Gates debunks ’coronavirus vaccine is my 5G mind control microchip implant’ conspiracy theory |
| False claim: Coronavirus is a hoax and part of a wider 5G and human microchipping conspiracy |
| **Ayurveda** |
| Ayurveda cannot cure COVID-19. |
| Ayurveda does not play a Role in Fight Against Coronavirus |
| Fake: Ayush ministers claim debunked, Prince Charles not cured from Covid-19 with Ayurveda. |

Table 4.12: Facts used to detect misinformation in tweets for topics Vitamin D, Salt water, 5G, and Ayurveda.

| **Facts** |
|---|
| **Ibuprofen** |
| There is no scientific evidence that ibuprofen worsens coronavirus. |
| WHO now says it is OK to take Ibuprofen for coronavirus symptoms. |
| There is no evidence that taking a non-steroidal anti-inflammatory drug like ibuprofen could be harmful for people infected with the new coronavirus. |
| **Vitamin C** |
| Vitamin C cannot cure COVID-19. |
| Consuming vitamin c on a daily basis can boost your body's immune function and help you actively fight against infections but there is no evidence till now that it helps in preventing COVID-19. |
| Vitamin C has not been approved by the FDA or any other agency to treat or prevent the coronavirus. |
| People are falsely claiming that vitamin C cures the coronavirus. |
| Taking too much vitamin C without a doctor's supervision could be lethal. |
| Although vitamin C does have some small effect on the common cold, it's unlikely that taking large amounts of vitamin C supplements will cure a COVID-19. |
| **Smoking** |
| Smoking does not prevent COVID-19. |
| Smokers are generally considered a risk group for infections with the novel coronavirus. |
| The WHO has warned that cigarettes can actually increase the risk of contracting Covid-19. |
| Smokers are likely to be more vulnerable to Covid-19 |

Table 4.13: Facts used to detect misinformation in tweets for topics Ibuprofen, Vitamin C, and Smoking.

# Chapter 5

# Conclusion & Future Directions

In this work, I introduce the task of misinformation retrieval and purpose a pipeline for retrieval on tweets. As part of the research, I curated a list of 50 COVID-19 misinformation topics used in the TREC 2020 Health Misinformation track and the testing of our model. I annotated a test set of tweets using the TREC topics. I expanded the COVID-Lies data set by rephrasing misconceptions to include facts in the data and introduce lexical and semantic diversity. Finally, I demonstrated the importance of continued pretraining of BERT on COVID-19 tweets and used the best-performing model to retrieve misinformation on 7 TREC topics from the Twitter COVID-19 stream.

Manually inspecting the incorrectly classified tweets reveals some of the weaknesses in our model, namely the model's inability to classify vague claims and capture sarcasm. Failure to classify vague tweets is not as critical as failing to capturing sarcasm since this is also a difficult task for human fact checkers. In these cases, we expect there to be a higher level of disagreement between human fact-checkers. Sarcasm detection in the context of natural language inference is a difficult task, especially when classifying tweets only based on text. We believe that it is possible to remedy with weakness by considering more than just the text of the original tweet. Extracting contextual information from replies to a tweet will improve the model performance. Another possible improvement for the model might be training CT-BERT with semantic role labels for each tweet. This is explored by SemBERT [50], showing significant improvement on the SNLI dataset. The goal of that work was to translate the knowledge humans have about sentence structures and use that information to find meaning in a sentence. We believe that using this approach might improve results significantly based on its performance on previous NLI tasks[1]. Furthermore, language on Twitter is very different from standard English, and tweets have unique attributes not present a typical English sentence, for example, hashtags. Oftentimes tweets convey a stance on a topic without using a full English sentence, such as the tweet ”RT @USER: @USER HTTPURL 🤔 ?? 5g = covid For real”. The meaning in the tweet is evidence to a Twitter user, however a machine learning model will require more information in order to understand the meaning in the tweet. Domain specific semantic role labeling of tweets will a great direction to take.

Aside from the improvements on the model, I believe a sentence extraction can be added to the pipeline to extract facts from factual sources similar to those presented by Pradeep et. al [36]. The task of manually extracted factual statements from correct documents is a difficult and tedious task. By adding this feature, the pipeline will be fully automated. With the current pipeline, I have laid out the foundation on which a fully automated misinformation retrieval pipeline can be built, and easily apply the improvements I have outlined.

[1] https://paperswithcode.com/sota/natural-language-inference-on-snli

# References

[1] Mustafa Abualsaud, Fuat C Beylunioglu, Mark D Smucker, and P Robert Duimering. Uwaterloomds at the trec 2019 decision track. In *TREC*, 2019.

[2] Mustafa Abualsaud, Nimesh Ghelani, Haotian Zhang, Mark D. Smucker, Gordon V. Cormack, and Maura R. Grossman. A system for efficient high-recall retrieval. In *The 41st International ACM SIGIR Conference on Research  Development in Information Retrieval*, SIGIR '18, page 1317–1320, New York, NY, USA, 2018. Association for Computing Machinery.

[3] Mustafa Abualsaud, C. Lioma, Maria Maistro, M. Smucker, Guido, and Zuccon. Overview of the trec 2019 decision track. 2020.

[4] Mabrook S. Al-Rakhami and Atif M. Al-Amri. Lies kill, facts save: Detecting covid-19 misinformation in twitter. *IEEE Access*, 8:155961–155970, 2020.

[5] Firoj Alam, Fahim Dalvi, Shaden Shaar, N. Durrani, Hamdy Mubarak, Alex Nikolov, Giovanni Da San Martino, Ahmed Abdelali, Hassan Sajjad, Kareem Darwish, and Preslav Nakov. Fighting the covid-19 infodemic in social media: A holistic perspective and a call to arms. *ArXiv*, abs/2007.07996, 2020.

[6] Alexander Bondarenko, Maik Fröbe, Vaibhav Kasturia, Matthias Hagen, Michael Völske, and Benno Stein. Webis at trec 2019: Decision track. In *TREC*, 2019.

[7] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. The snli corpus. 2015.

[8] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 675–684, New York, NY, USA, 2011. Association for Computing Machinery.

[9] Charles L. A. Clarke, Maria Maistro, Saira Rizvi, Mark D. Smucker, and Guido Zuccon. Overview of the trec 2020 health misinformation track.

[10] Charles L. A. Clarke, Mark D. Smucker, and Alexandra Vtyurina. *Offline Evaluation by Maximum Similarity to an Ideal Ranking*, page 225–234. Association for Computing Machinery, New York, NY, USA, 2020.

[11] Limeng Cui and Dongwon Lee. Coaid: Covid-19 healthcare misinformation dataset, 2020.

[12] Wanqing Cui, Yan Jiang, Shuchang Tao, and Hanzhang Guo. Ictnet at trec 2019 decision track. In *TREC*, 2019.

[13] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short*

*Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[15] Anastasia Giachanou, Paolo Rosso, and Fabio Crestani. Leveraging emotional signals for credibility detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 877–880, New York, NY, USA, 2019. Association for Computing Machinery.

[16] Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[17] Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. The future of false information detection on social media: New perspectives and trends. *ACM Comput. Surv.*, 53(4), July 2020.

[18] Chuan Guo, J. Cao, X. Zhang, Kai Shu, and M. Yu. Exploiting emotions for fake news detection on social media. *ArXiv*, abs/1903.01728, 2019.

[19] Manish Gupta, Peixiang Zhao, and Jiawei Han. Evaluating event credibility on twitter. In *SDM*, 2012.

[20] Momchil Hardalov, Ivan Koychev, and Preslav Nakov. In search of credible news. In Christo Dichev and Gennady Agre, editors, *Artificial Intelligence: Methodology, Systems, and Applications*, pages 172–180, Cham, 2016. Springer International Publishing.

[21] Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. COVIDLies: Detecting COVID-19 misinformation on

social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December 2020. Association for Computational Linguistics.

[22] Jimmy and Guido Zuccon. Uq ielab at trec 2019 decision track. In *TREC*, 2019.

[23] Z. Jin, J. Cao, Yu-Gang Jiang, and Yongdong Zhang. News credibility evaluation on microblog with a hierarchical propagation model. *2014 IEEE International Conference on Data Mining*, pages 230–239, 2014.

[24] Nikhil L. Kolluri and Dhiraj Murthy. Coverifi: A covid-19 news verification system. *Online Social Networks and Media*, 22:100123, 2021.

[25] Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. Misinformation has high perplexity. *ArXiv*, abs/2006.04666, 2020.

[26] Lucas Chaves Lima, Dustin Brandon Wright, Isabelle Augenstein, and Maria Maistro. University of copenhagen participation in trec health misinformation track 2020. *arXiv preprint arXiv:2103.02462*, 2021.

[27] Christina Lioma, Jakob Grue Simonsen, and Birger Larsen. Evaluation measures for relevance and credibility in ranked lists. ICTIR '17, page 91–98, New York, NY, USA, 2017. Association for Computing Machinery.

[28] Xiaofei Ma, Peng Xu, Zhiguo Wang, and Ramesh Nallapati. Domain adaptation with bert-based domain classification and data selection. In *DeepLo@EMNLP-IJCNLP*, 2019.

[29] Shahan Ali Memon and Kathleen M. Carley. Characterizing covid-19 misinformation communities using a novel twitter dataset. In *CEUR Workshop Proc.*, volume 2699, 2020.

[30] Federico Monti, F. Frasca, D. Eynard, Damon Mannion, and M. Bronstein. Fake news detection on social media using geometric deep learning. *ArXiv*, abs/1902.06673, 2019.

[31] M. Müller, M. Salathé, and P. Kummervold. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *ArXiv*, abs/2005.07503, 2020.

[32] Parth Patwa, Mohit Bhardwaj, Vineeth Guptha, Gitanjali Kumari, Shivam Sharma, Srinivas PYKL, Amitava Das, Asif Ekbal, Md Shad Akhtar, and Tanmoy Chakraborty. Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts. In Tanmoy Chakraborty, Kai Shu, H. Russell Bernard, Huan Liu, and Md Shad Akhtar, editors, *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 42–53, Cham, 2021. Springer International Publishing.

[33] Parth Patwa, Srinivas PYKL, Vineeth Guptha, Gitanjali Kumari, Md Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. Fighting an infodemic: Covid-19 fake news dataset. 11 2020.

[34] Frances A. Pogacar, Amira Ghenai, M. Smucker, and C. Clarke. The positive and negative influence of search results on people's decisions about the efficacy of medical treatments. *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, 2017.

[35] Dean Pomerleau and Delip Rao. Fake news challenge. *Fake News Challenge*, 2017.

[36] Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. Scientific claim verification with vert5erini. *CoRR*, abs/2010.11930, 2020.

[37] Ronak Pradeep, Xueguang Ma, Xinyu Zhang, Hang Cui, Ruizhou Xu, Rodrigo Nogueira, and Jimmy Lin. H2oloo at trec 2020: When all you got is a hammer... deep learning, health misinformation, and precision medicine. *Corpus*, 5(d3):d2.

[38] Gautam Kishore Shahi and Durgesh Nandini. FakeCovid – a multilingual cross-domain fact check news dataset for covid-19. In *Workshop Proceedings of the 14th International*

*AAAI Conference on Web and Social Media*, 2020.

[39] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. 19(1):22–36, September 2017.

[40] E. Tacchini, Gabriele Ballarin, M. L. D. Vedova, Stefano Moret, and L. D. Alfaro. Some like it hoax: Automated fake news detection in social networks. *ArXiv*, abs/1704.07506, 2017.

[41] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[42] Rutvik Vijjali, Prathyush Potluri, Siddharth Kumar, and Sundeep Teki. Two stage transformer model for COVID-19 fake news detection and fact checking. In *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 1–10, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics (ICCL).

[43] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online, November 2020. Association for Computational Linguistics.

[44] William Yang Wang. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Compu-*

*tational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[45] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics, 2018.

[46] Liang Wu and Huan Liu. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 637–645, New York, NY, USA, 2018. Association for Computing Machinery.

[47] Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. Misinformation in social media: Definition, manipulation, and detection. *SIGKDD Explor. Newsl.*, 21(2):80–90, November 2019.

[48] S. Yang, Kai Shu, Suhang Wang, Renjie Gu, F. Wu, and Huan Liu. Unsupervised fake news detection on social media: A generative approach. In *AAAI*, 2019.

[49] X. Zhang, J. Cao, Xirong Li, Qiang Sheng, L. Zhong, and Kai Shu. Mining dual emotion for fake news detection. *arXiv: Computation and Language*, 2019.

[50] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. Semantics-aware bert for language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:9628–9635, 04 2020.

[51] Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th*

*ACM International Conference on Information Knowledge Management*, CIKM '20, page 3205–3212, New York, NY, USA, 2020. Association for Computing Machinery.

# APPENDICES

# Appendix A

# Matlab Code for Making a PDF Plot

## A Using the Graphical User Interface

Properties of Matab plots can be adjusted from the plot window via a graphical interface. Under the Desktop menu in the Figure window, select the Property Editor. You may also want to check the Plot Browser and Figure Palette for more tools. To adjust properties of the axes, look under the Edit menu and select Axes Properties.

To set the figure size and to save as PDF or other file formats, click the Export Setup button in the figure Property Editor.

## A From the Command Line

All figure properties can also be manipulated from the command line. Here's an example:

```
x=[0:0.1:pi];
hold on % Plot multiple traces on one figure
```

```
plot(x,sin(x))
plot(x,cos(x),'--r')
plot(x,tan(x),'.-g')
title('Some Trig Functions Over 0 to \pi') % Note LaTeX markup!
legend('{\it sin}(x)','{\it cos}(x)','{\it tan}(x)')
hold off
set(gca,'Ylim',[-3 3]) % Adjust Y limits of "current axes"
set(gcf,'Units','inches') % Set figure size units of "current figure"
set(gcf,'Position',[0,0,6,4]) % Set figure width (6 in.) and height (4 in.)
cd n:\thesis\plots % Select where to save
print -dpdf plot.pdf % Save as PDF
```