FACT EXTRACTION AND VERIFICATION PIPELINE FOR COVID-19 RELATED USER POSTS IN SOCIAL MEDIA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

ORKUN TEMİZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

JUNE 2022

[SAMPLE 1]
Approval of the thesis:

FACT EXTRACTION AND VERIFICATION PIPELINE FOR COVID-19 RELATED USER POSTS IN SOCIAL MEDIA

Submitted by ORKUN TEMİZ in partial fulfillment of the requirements for the degree of **Master of Science in in Information Systems Department, Middle East Technical University** by,

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname : Orkun Temiz

Signature : _____

ABSTRACT

FACT EXTRACTION AND VERIFICATION PIPELINE FOR COVID-19 RELATED USER POSTS IN SOCIAL MEDIA

Temiz, Orkun

MSc., Department of Department of Information Systems

Supervisor: Prof. Dr. Tuğba Taşkaya Temizel

June 2022, 45 pages

Social media has become a prevalent platform for consuming and sharing information online. The vast amounts of information, shared easily and rapidly by social media, have increased the demand for fact-checking. Misinformation threatens not only the reputation of individuals and organizations but also society. When the COVID-19 pandemic broke out, the concerns around misinformation, which threatens public health and society, have significantly increased. In this thesis, a new zero-shot fact extraction and verification pipeline for user posts related to COVID-19 against the medical articles is proposed. The pipeline comprises preprocessing of user posts, claim extraction, document retrieval, evidence selection, and verdict assignment components. The proposed pipeline not only labels the claim but also presents the related evidence set extracted from the pipeline regarding the claim, which gives interpretable results for the society about the claim. Also, it does not need to see previously labeled posts unlike numerous supervised studies in the literature instead; it uses the zero-shot capabilities of existing models. The proposed pipeline obtains on-par and stable performance compared with the state-of-art supervised techniques for classifying raw user posts (CoAID) and rumors collected from social media (COVID-19 Rumors Dataset). On the other hand, it achieves superior performance in detecting new emerging misinformation topics.

Keywords: Data Mining, Fact Checking and Verification System, Natural Language Processing, Information Retrieval, Social Media

ÖZ

COVID-19 İLE İLGİLİ SOSYAL MEDYA GÖNDERİLERİNDE ÖNERME ÇIKARMA VE DOĞRULAMA MODEL HATTI

Temiz, Orkun

Yüksek Lisans, Bilişim Sistemleri Bölümü

Tez Yöneticisi: Prof. Dr. Tuğba Taşkaya Temizel

Haziran 2022, 45 sayfa

Günümüzde sosyal medya, çevrimiçi bilgi tüketmek ve paylaşmak için yaygın bir platform haline geldi. Sosyal medya tarafından kolay ve hızlı bir şekilde paylaşılan büyük miktardaki bilgi, doğrulama talebini artırdı.Yanlış bilgi sadece kişi ve kuruluşların itibarını değil, toplumu da tehdit etmektedir. COVID-19 pandemisi patlak verdiğinde halk sağlığını ve toplumu tehdit eden yanlış bilgilendirme konusundaki endişeler de önemli ölçüde artmıştır. Bu tezde, COVID-19 ile ilgili kullanıcı gönderileri için bunları tıbbi makalelere karşı değerlendiren yeni bir sıfırdan gerçek çıkarma ve doğrulama hattı önerildi. İşlem hattı, kullanıcı gönderilerinin ön işlemesini, talep çıkarma, belge alma, kanıt seçimi ve karar atama bileşenlerini içerir. Önerilen model hattı sadece iddiayı etiketlemekle kalmamakta, aynı zamanda iddiaya ilişkin model hattından çıkarılan ilgili kanıt setini de sunup bu iddia hakkında toplum için yorumlanabilir sonuçlar da vermektedir. Ayrıca önerilen model hattının, literatürdeki çok sayıda denetimli çalışmanın aksine önceden etiketlenmiş gönderileri görmesine gerek yoktur ve mevcut modellerin sıfır atış yeteneklerini kullanmaktadır. Ham kullanıcı gönderilerini (CoAID) ve sosyal medyadan toplanan söylentileri (COVID-19 Rumors Dataset) içeren veri setleri kullanılarak sınıflandırma konusunda son teknoloji denetimli tekniklerle karşılaştırma yapıldığında eşit ve istikrarlı performans elde edilmektedir. Öte yandan, ortaya çıkan yeni yanlış bilgilendirme konularını tespit etmede önerilen model karşılaştırılan modellere nazaran üstün performans elde etmektedir.

Anahtar Sözcükler: Veri Madenciliği, Gerçek Çıkarma ve Doğrulama Sistemi, Doğal Dil İşleme, Bilgi Çıkartma, Sosyal Medya

To Second Chances

ACKNOWLEDGMENTS

First and foremost, I would like to thank my supervisor Prof. Dr. Tuğba Taşkaya Temizel for her valuable time, wisdom, guidance, support, patience, and generosity. Her efforts have been crucial not only in my thesis research but also in my academic journey. I am immensely grateful that I was a part of her research team that pushed me to learn more and go beyond. I am grateful to everyone who taught me something or influenced me, especially professors in the Informatics Institute at Middle East Technical University, that lead me to this position. I am thankful to the examining committee members for their valuable feedback and suggestions.

I would like to acknowledge the support of TUBITAK for granting me 2210/C scholarship in my master studies.

Finally, I would like to thank my family and friends for their support. I also want to thank Neşe Küçükbalcı for her support and contributions in the data retrieval process.

TABLE OF CONTENTS

ABSTRACT iv
ÖZ v
DEDICATION vi
ACKNOWLEDGMENTS vii
TABLE OF CONTENTS viii
LIST OF TABLES x
LIST OF FIGURES xi
LIST OF ABBREVIATIONS xii
CHAPTER 1
1. INTRODUCTION 1
1.1. Research Questions 2
1.2. Contributions of the Study 3
1.3. Organization of the Thesis 3
CHAPTER 2
2. RELATED WORK 5
2.1. Claim Extraction 5
2.2. Document Retrieval & Evidence Retrieval 6
2.3. Fact Checking 8
2.4. The Research Gap in the Literature 9
CHAPTER 3
3. METHOD 11
3.1. Preprocessing 12
3.2. Claim Extraction 12
3.3. Keyword Extraction & Enhancement 13
3.4. Document Retrieval 13
3.5. Evidence Selection 14

3.6. Textual Entailment 17
3.7. Heuristic Verdict Assignment 18
CHAPTER 4
4. EXPERIMENTS 19
4.1. Datasets 19
4.2. Results 20
4.2.1 First Evaluation Scheme 21
4.2.2 Second Evaluation Scheme 22
4.3. Ablation Studies 25
4.3.1 Study 1 26
4.3.2 Study 2 26
4.3.3 Study 3 27
CHAPTER 5
5. DISCUSSION & FUTURE WORK 29
CHAPTER 6
6. CONCLUSION 33
REFERENCES 37

LIST OF TABLES

Table 1: The test results of the CoAid (Tweets + News Titles), which was split according to user postdates.. 22
Table 2: The results of the CoAid (Tweets + News Titles), COVID-19 Rumors. 23
Table 3: The results of the models on the CoAID “Claim” Posts Only 25
Table 4: The results of the ablation study for Natural Language Inference Model 25
Table 5: Ablation study results showing the effect of summarization and MeSH terms CoAID dataset 25
Table 6: The results of the third ablation study 26

LIST OF FIGURES

Figure 1: The proposed system pipeline. 11
Figure 2: An example outcome for a given user post and its retrieved evidence pairs....15
Figure 3: An example outcome for a given user post and its retrieved evidence pairs....16
Figure 4: The testing results of the proposed pipeline and the baseline models.....23
Figure 5: The KL divergence results for different total cluster numbers for given dataset 24

LIST OF ABBREVIATIONS

API	Application Programming Interface
BART	Denoising Autoencoder For Pretraining Sequence-To-Sequence Models
BERT	Bidirectional Encoder Representations from Transformers
BIOBERT	Biological Bidirectional Encoder Representations from Transformers
Bi-LSTM	Bidirectional long-short term memory
CNN	Convolutional Neural Networks
FEVER	Fact Extraction and Verification
MeSH	Medical Subject Headings
NE	Named Entities
NER	Named Entity Recognition
NLI	Natural Language Inference
PEGASUS	Pre-training with Extracted Gap-sentences for Abstractive Summarization
QA	Question Answering

CHAPTER 1

INTRODUCTION

Social media has become the leading platform for creating and sharing information, ideas, interests, and other forms of narration using the World Wide Web. The tremendous amount of information and user posts increased the necessity of fact-checking. The spread of misinformation directly threats individuals and organizations as well as public health. For example, during the COVID-19 pandemic, the spread of misinformation and its prevention has become one of the main concerns. It is well-known that social media can propagate misinformation quickly, easily distorting trustworthy claims about different health topics. In the context of the COVID-19 pandemic, misinformation disseminates faster than the virus [1]; therefore, it has been called infodemic [2]. To cope with the spread of misinformation, fact-checking organizations such as Snopes [3], and Politifact [4] have increased their activities. However, they have failed to timely respond to COVID-19 misinformation as they have referred to domain experts and journalists to analyze data to debunk false and misleading information and manual preparation of each claim's response takes a significant amount of time. Social media platforms engage in content moderation to cope with this situation. In 2016, Facebook started an action against posts disproved by fact-checkers. In 2020, Twitter took similar actions to manage the spread of misinformation that could endanger public health related to the COVID-19. Twitter gave access to developers and researchers by providing a specific COVID-19 streaming endpoint [5] and an academic version of its API, which allows them to study misinformation and hoaxes in user posts better. YouTube removes the videos containing COVID-19 misinformation and limits the recommendation of antivaccination content to deal with the problem. However, for the social media platforms, the problem is still the same; detecting misinformation takes time, and information spreads very fast [6, 7]. Thus, detecting misinformation early, ideally at the time of submission of a post, is becoming more and more critical. When misinformation becomes viral, users might become less likely to change their beliefs even when the misinformation is debunked. Hence, there is a strong need for an automatic, almost real-time fact-checking solution to detect misinformation as well as to verify given information. In this thesis, as a solution, I propose a zero-shot learning

approach for misinformation classification and verification of COVID-19 claims in user posts using peer-reviewed scientific articles. This approach also facilitates the identification of new emerging health topics in user posts on time. The proposed method is scalable for other health topics, although it was only applied to COVID-19 datasets in this thesis. I used two different datasets, one of which contains misinformative and informative tweets of COVID-19, collected from fact-checking websites and reliable organizations [8], and the other comprises claims and raw tweets collected by using the claims as search terms [9].

I built a fact extraction and verification system, which labels a claim present in a user post either "*SUPPORTS", "NOT ENOUGH EVIDENCE" or "REFUTES"* and presents the related evidence set extracted from the pipeline regarding the claim. Giving related evidence about the claim makes the fact-checking process depend on evidence-based and contextualized analysis, which society comprehends. The links to the documents of the retrieved evidence sources are also presented, so users can inspect them for further information.

1.1. Research Questions

This thesis seeks to answer the following main research question:

RQ: How is it possible to develop a method for fact-checking and verifying user posts against published peer-reviewed articles?

This primary research question is divided into the following sub research questions:

RQ1: Is it possible to map an informal medical claim to formal medical articles in social media?

It is hypothesized that verified medical articles can be used as a single source of truth, and medical claims in social media can be mapped to the evidence found in articles. Consequently, if the evidence and claim are entailed with each other, this concludes that the claim is correct or vice versa. To explore this, user posts which include medical claims are analyzed.

RQ2: Can an evidence-based fact-checking method without direct supervision for newly emerging medical claims perform on-par/better results than state-of-art supervised models?

It is known that supervised models must learn from the past data, but if newly emerging topics are considered, there is a high probability that there will not be enough data to train the supervised models. In addition, medical articles from respected platforms could be considered as the ground truth, therefore, to decide whether a claim is "true" or "false," it should be fact-checked against an article as in the processes in the

independent fact-checking platforms (e.g., Snopes). It is hypothesized that the evidence-based fact-checking pipeline with a zero-shot based approach might perform better in newly emerging medical claims/topics than state-of-art supervised models in fact classification.

RQ3: Is it possible to improve medical document retrieval performance by using MeSH (Medical Subject Headings) tree structure for query enhancement?

It is hypothesized that expanding the search query with the medical terms and synonyms of those medical words (e.g., common flu vs. influenza) might increase the document retrieval performance. In addition to the search query, keywords found in the MeSH tree are used to explore this situation.

1.2. Contributions of the Study

The contributions of the study are as follows:

- Proposing a pipeline that utilizes zero-shot capabilities of the existing models to fact-check user posts, including medical claims.
- Utilizing CORD-19 [10] dataset together with the MedrXiv [11] and Pubmed [12] to keep up with the latest medical information for medical fact-checking.
- Mapping informal user posts to scientific medical articles and fact check these posts against the related evidence found in these articles.
- Determining the medical keywords in the claim using MeSH for enhancing document retrieval and textual entailment performance.
- Using text simplification and transformation to map formal evidence to informal user claims to improve the textual entailment performance.
- Providing a method that requires no supervision for checking newly emerged medical claims.
- Providing evidence-based fact-checking linked to a medical article/sources empowers users to make up their minds in light of explicit references related to the claim.

1.3. Organization of the Thesis

The organization of the thesis is as follows. Chapter 2 reviews the previous studies and explains how the thesis fits in the literature. Chapter 3 presents the data source of choice, used datasets, the methodology of collecting the data,

and the methodology of specific analysis methods and provides the results. Chapter 4 summarizes the results and provides further discussions. Chapter 5 concludes the thesis, mentions limitations and points to directions for future work.

CHAPTER 2

RELATED WORK

In this chapter, related studies are given, and later they are explained in terms of how these studies fit in the literature. It is possible to group related work under three categories: claim extraction, evidence retrieval, and fact-checking. These categories are discussed in the following sections.

2.1. Claim Extraction

Previous research has studied extracting claims from various textual sources. In the literature, this task is called argumentation mining or claim extraction, which can be considered as the emerging field of natural language processing under sentence classification tasks. Although argumentation mining seems like a single task, the requirements and use-case scenarios differ substantially. Past studies has tried to retrieve claims from social media posts [13], news articles/ paragraphs [14, 15], Wikipedia [16, 17], and articles and essays [18]. However, due to a lack of data, there is significantly less emphasis given to the biomedical domain compared to the other domains. In the past studies, the claim retrieval and extraction task is mainly handled by feature or rule based methods [19, 20] and simple machine-learning methods [21], on the other hand recent argumentation mining techniques use state-of-art deep learning architectures or transfer learning methods (e.g., [22, 23, 24, 25]).

Thorne et al. [26] built the FEVER dataset which includes of claims which includes factual information from Wikipedia fact-checked and labeled by the community. However, fact-checking scientific (medical) claims requires high domain profession rather than common wisdom from a shared community. Similarly, PUBMEDRCT [27], a dataset of scientific expression that sampled sentence groups were annotated according to their presence in the background, introduction, method, result, and conclusion sections of the article. Although the task is different from the argument mining, the main task of both are sentence classification, yet this dataset was not expertly annotated, although it requires specific domain knowledge. Consequently, for claim extraction tasks, especially in the medical domain, there are limited datasets annotated by domain experts (see [19, 21, 28]). Other than these, different techniques

have been studied for claim extraction task. In Yuan et al. [20], the authors used feature-based (rule-based) claim extraction methods, and in Yu et al. [29], the authors covered multiple deep learning architectures. Li et al. [30] studied a Bi-LSTM architecture which uses triplets to retrieve claims from raw text. In Arslan et al. [31], a feature-based method that uses part-of-speech (POS) tags to assign a check-worthiness score to a sentence was proposed, and they extracted claims from the raw sentences with a predetermined threshold.

It is known that deep learning techniques require a significant amount of data to work well, and therefore datasets with scale are required to train deep learning models. Maria et al. [21] created the CoreSC dataset includingş 270 hundred articles, which composing physical and bio chemistry topics. Dasigi et al. [28] created a dataset which compose 75 articles for extracting statement using medical articles from PubMed. Stateli et al. [19] introduced a dataset with evaluations for claims and additions using the complete copy of 30 articles which retrieved from the computer science area. Nevertheless, these datasets are still small to train a deep learning model for augment mining on medical text. Since this task is complex, an expert contribution is highly needed to provide these annotations. Furthermore, most of the datasets in the literature are restricted to predefined domains such as politics, computer linguistics, and chemistry. As a consequence, feature-based ClaimBuster API is used in the proposed pipeline to extract claims, although it is trained on general election debates in user posts from social media. In order to push the development of augmentation mining task, Conference and Labs of the Evaluation Forum (CLEF) organized a series of challenges [32]. The organization has been active since 2018 and encourages the proposal of a multilingual approach to this task. Throughout the history of this organization, they organized challenges related to the identification and verification of claims in political debates in six different languages, and recently, they have added COVID-19 related infodemic posts to the challenges. This dataset is not considered in the scope of this thesis since they announced that recently.

2.2. Document Retrieval & Evidence Retrieval

Document retrieval is a task that identifies the relevant documents from sources (e.g., Wikipedia, Pubmed). In other words, it is a task that aims at matching a query against a collection of unstructured or structured documents.

In the earlier studies of the document retrieval task, mention-based methods are used. Considering most of the claims have contained named entities as a subject and object in the sentences, Hanselowski et al. and Tuhin et al. [33, 34] have concentrated on the importance of named entities (NE) to direct the document search to specific results. Hanselowski et al. [33] studied a method which consists of mention retrieval, article search, and filtering. In the study, mention extraction part depends on a simple constituency parser [35]. According to the results of the constituency parser, noun

phrases in a claim are labeled as potential entities. Also, all words near the main verb are considered as potential entity mentions. The candidate article search component uses an external search API to retrieve the potential entity mentions. Candidate filtering component filters the entity mentions, which are not part of the claim. The methodology of the work presented in [33] is also followed by [36, 37]. Tuhin et al. [34] used a similar approach, apart for NE (Named-entity) recognition authors have used the Google search API along with a dependency parser for improving the scope of the documents that are retrieved. In this study, the authors also resolve the disambiguation between the different meanings of a same word.

Another line of study for document retrieval tasks is, using keyword-based methods. Atanasova et al. [38] presented a model that has three stages and relies on the Neural Semantic Matching Network (NSNM), which can be considered as another type of ESIM [39]. In the document retrieval task, authors have utilized a key-value matching approach which depends on one-to-one matching, article, and singularization. Then, all documents which do not include direct information are included to the document list. The remaining documents are sorted and picked out using Neural Semantic Matching and a limitting value. The study of Luken et al. [40] aimed to retrieve POS tags, dependencies, etc., by utilizing the CoreNLP parser [41] for key-phrase identification.

Apart from those methods, document retrieval is directly tied to the task of evidence retrieval, and standard practices such as BM25 or cosine similarity indexes in the word vector space can be used as baseline approaches for retrieving the related documents. BM25 indexing is also utilized in the document retrieval module of the proposed pipeline for indexing the documents and retrieving the related top related ones only.

Unlike document retrieval, evidence retrieval is also a important piece in fact-checking task. In other words, it is essential to select and retrieve the pieces of evidence from the large and noisy documents. These evidence sentences explain why a claim has been evaluated as reliable or not. For this task, different approaches are adopted in the literature; Thorne et al. [26] utilized a (Term Frequency – Inverse Document Frequency) TF-IDF based method which is similar to their document retrieval approach from Wikipedia. UCL [41] used a logistic regression model on top of a chosen set of features and attributes. Apart from those deep learning based approaches adopted for evidence retrieval, Enhanced Sequential Inference Model (ESIM) [42] has been used for evidence retrieval [33, 43]. ESIM utilizes the co-attention mechanism on top of two BiLSTMs to detect a related evidence based on a given sentence. Recent works proposed sentence similarity models for this task. Atanasova et al. [38] and Soleimani et al. [44] used BERT-based model for retrieving related evidence collections from raw articles. A zero-shot-based approach is proposed for this task in the proposed pipeline, which uses the BERT-based question-answering model. It is also shown that using BERT based question-answering model as evidence retrieval achieves comparable results.

2.3. Fact Checking

Machine learning methods are used to identify misinformation, and they generally make use of (1) content, (2) derived features from content, social network, or author of the post, or (3) hybrid features (combining the first two). Claims present in posts and metadata are typically encoded using convolutional neural networks (CNNs) or recurrent neural networks (RNNs) [10, 45]. Methods such as support vector machines (SVM) or multi-layer perceptron (MLP), applied on smaller datasets, often use handcrafted or derived features, including a bag of words and other lexical features, e.g., LIWC. With the advent of deep learning, there has been a significant development in the field of text classification and, thereby, fake news classification. Especially transformers-based models (BERT, XLNET, ROBERTA) and their ensembled versions are widely used. Different from machine learning models, such as neural networks, that output a single value to predict the veracity of a claim, Zhang et al. [46] used a Bayesian-learning based approach, which outputs a distribution that can model both the prediction and its uncertainty in parallel. In this way, they model the inability to represent the uncertainty in the prediction. Also, when the past studies on the fact-checking topic are analyzed, it can be seen that the most of the studies were done on text classification. However, apart from text which subject to fact-checking, misinformation consumers tend to trust the posts/ news when images supplement the text. For this purpose, Ghai et al. [47] developed deep-learning-based image forgery detection to address the problem of image manipulation.

Few studies used Twitter-specific features in addition to word and lexical features like follower count, verified account info etc. [48, 49]. Elhadad, Li, and Gebali [50] compared ten supervised machine learning algorithms with seven feature extraction techniques to detect COVID-19 misleading textual content, where logistic regression, decision tree, and neural network methods produced the best results. AlRakhami and Al-Amri [51] proposed a stacking-based ensemble-learning model by integrating six machine learning algorithms to detect misinformation in Twitter posts using tweet-level and user-level features. In this study authors applied the model to a dataset of tweets collected from 15 January to 15 April 2020 using relevant keywords about COVID-19. The results showed that the proposed ensemble-learning model had better performance compared to single machine-learning-based models. As we see in the COVID-19 infodemic, there are a lot of contractual user posts on social media according to medical information. Wang et al. [52] proposed an ensemble model that looks for social media posts as whole (images, texts, and hashtags) instead of focusing on textual components to detect anti-vaccine content propagating through social media.

Similarly, due to the complexity of medical and public health issues in addition to disagreements that get exacerbated by misinformation, it is often challenging to be both accurate and factual. This difficulty is compounded by the rapid evolution of knowledge regarding the disease. For example in COVID-19 pandemic as researchers

learn more about the virus, statements that seem to be true may turn out to be false in the future, and vice versa. When only historical data (i.e., previous posts or syntactic features) are used, misinformation detection is becoming limited in evaluating pressing and rapidly evolving issues like COVID-19 [53]. In this context, supervised models may not perform well in identifying new misinformation topics since they were not trained on them. To counterbalance the lack of data in newly emerged topics, zero-shot learning can be used to predict data instances without requiring any explicit training since each class to predict is associated with a semantic prototype that reflects the essential features of the task. This work aims to detect misinformative user posts related to COVID-19 by proposing a pipeline that uses zero-shot capabilities of existing transformers-based deep learning models (e.g., BERT trained on SQUAD corpus, BERT trained on NLI corpus, etc.).

2.4. The Research Gap in the Literature

A fact-checking approach that can identify new misinformation topics from social media posts in the health domain such as COVID-19 while giving the related scientific evidence needs to meet the following requirements in a pipeline:

1) Being able to react to new emerging claims, 2) Being able to retrieve relevant documents from a regularly updated document collection such as MEDLINE, 3) Selecting textual evidence sentences that can support or refute the claim, 4) Being able to establish a link between informal and formal texts to relate claims present in user posts with the evidence obtained from the scientific articles 5) Being able to predict the claim's veracity based on the evidence collection.

Recent related works have advanced the field by partially addressing several aforementioned pipeline requirements with numerous models and datasets [1, 26, 54, 55, 56]. For example, one line of work that includes FEVER [26] and SciFact [55] realizes the third and fifth requirements but fails to address the second one fully as it works only with a static document collection (Wikipedia or CORD-19, respectively). Multi-FC [56] successfully handles the first, second, and fifth requirements but not the third one. Because it checks real-world claims collected from fact-checking websites, evidence-based documents, and other meta information, but it does not provide evidence sentences supporting or refuting the claims in the output. Most of the existing studies do not implement the fourth requirement since the domain, and the level of formality are always assumed to be the same between the documents comprising evidence and claim.

It is believed that a greater deal of coverage of the fact-checking requirements can be achieved through the proposed pipeline (See Figure 1), and this coverage might be used to battle misinformation on social media effectively.

CHAPTER 3

METHOD

This thesis proposes a system comprising five major components: claim extraction, query enhancement, document retrieval, evidence selection, and textual entailment, as shown in Figure 1. This section presents the details of each component used in the pipeline with justifications. I also conducted an extensive ablation study to show the impact of each component in Section 4.

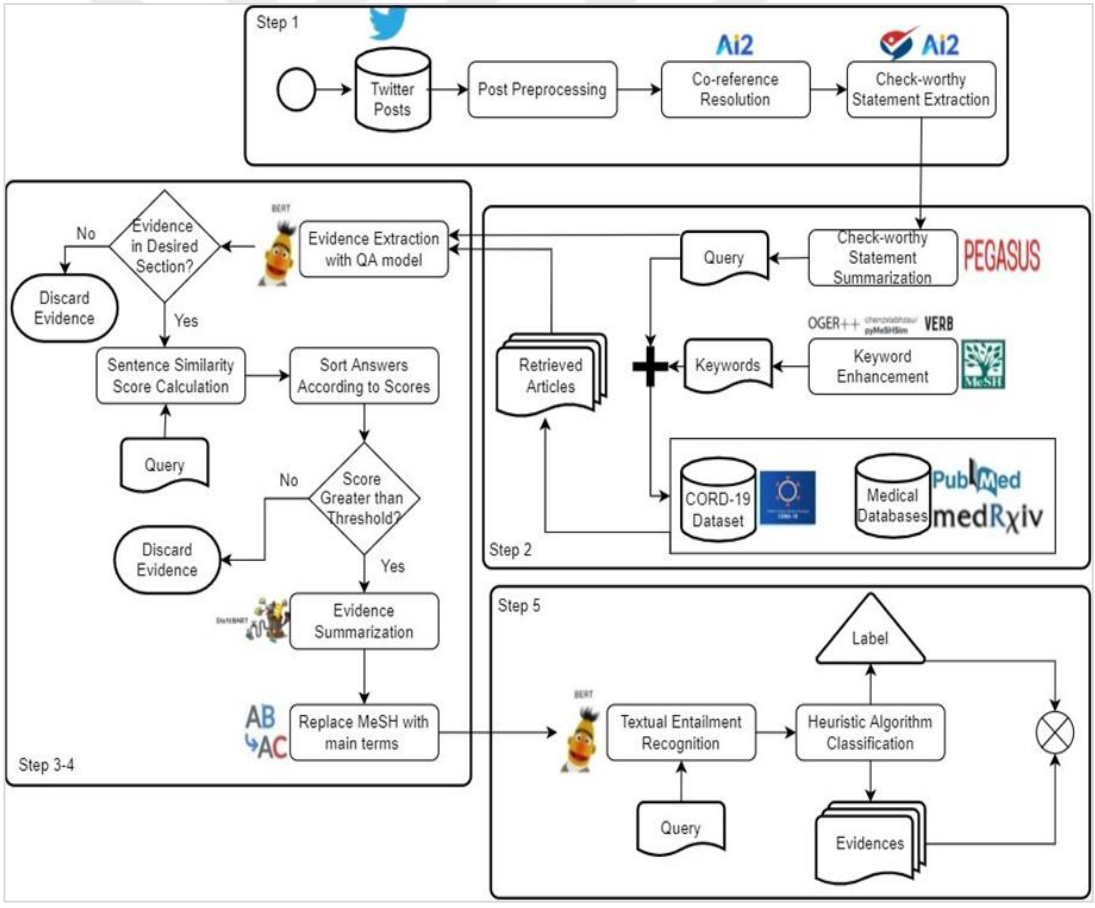

Figure 1: The proposed system pipeline.

3.1. Preprocessing

Firstly, the text is stripped out from special characters and irrelevant text (e.g. hashtags, URLs, emojis, mentions, images, etc.), which do not contribute to the claim itself to simplify user posts. Also, hashtag characters (#) are removed from the text, leaving the words without hashtag characters since they can be meaningful and used to form the part of a claim. Hashtag words are not removed directly to preserve the meaning of the posts. However, hashtag words were further processed by a constituency parser to differentiate hashtag chunks from hashtag words in a sentence.

3.2. Claim Extraction

To determine whether a user post includes a claim and select the claimant statement from the text, external ClaimBuster API [31] is used with modifications. Given a sentence, ClaimBuster gives it a score between 0.0 and 1.0. Consequently, the more likely the sentence contains check-worthy factual claims as the score increases. The lower the score, the more non-factual, subjective, and opinionated the sentence is. ClaimBuster's score is based on a classification and scoring model. The model was trained using past general election debates labeled by human coders. To fine-tune the threshold score for the domain in which the pipeline is performing, 100 random tweets are selected from the CoAID dataset, then sentences comprising the claim are manually labeled. Then according to the recall score, the best performing threshold is selected. As a result, the check-worthiness detector of the ClaimBuster is used with a threshold score of 0.35 whereas the default value was 0.5. In parallel, each post is parsed with the AllenNLP Constituency Parser [57] to identify sentence and noun phrase structures. If the outcome from the ClaimBuster API is not a noun phrase (i.e., hashtag chunks) or a question, it is labeled as a claim and eligible for fact-checking. In order to identify interrogative sentences on the raw user posts, certain patterns at the beginning of a sentence, e.g. (auxiliary verbs+ subjects such as; "do you," "can you," and WH-words such as who, what) and the presence of a question mark at the end of the sentence is searched. Also, the co-reference problem is handled before extracting the claim from the post to preserve the subject and objects of the claim. For this purpose, a pre-trained co-reference resolution model, which uses SpanBERT [58] embeddings [59], is used. For instance, a tweet comprising two sentences, "*Can regularly rinsing nose with saline help prevent infection with the new coronavirus? No. There is no evidence that this protected people from infection with new coronavirus*." the first one is labeled as "not-check worthy" since it is an interrogative question where as the second sentence comprises a claim, and "*this*" keyword is replaced with "*rinsing your nose with saline*" by using a co-reference resolution model (See step 1 in *Figure 1*).

3.3. Keyword Extraction & Enhancement

In this phase, search keywords used for the document retrieval model are determined from the claim and they are used supplementary to search query. Since the document retrieval module uses OKAPI BM25 [60] for indexing results, selecting appropriate keywords for the document retrieval module is important. Initially, medical keywords are retrieved by tokenizing the query using SciBERT [61], a pre-trained language model based on BERT and trained on papers from the corpus of Semantic Scholar [62]. After tokenization, stemming is applied to these keywords. Finally, each stem is checked against the National Library of Medicine's MeSH (Medical Subject Headings) vocabulary to determine whether it corresponds to any medical terms in the literature. If the match is found, it is added to the keyword set; otherwise, it is discarded from the keyword set.

In order to enhance search keywords, MeSH terms are used (e.g., 2019-nCov for COVID-19) from the NLM by searching Qualifier, Descriptor Terms, and Supplementary Concept Record Terms for the keyword via MeSH Tree Structure. To further enhance the keywords, OGER++ [63] and PyMeshSim [64] are used. OGER++ is a hybrid system for named entity recognition and concept recognition (linking), which combines a dictionary-based annotator with a corpus-based disambiguation component for medical subject headings. The pyMeSHSim recognizes bio-NEs using MetaMap, which produces Unified Medical Language System (UMLS) concepts in the natural language process. In order to identify medical terms successfully, these two methods are used as complementary to the query-based approach. Also, the verbs dependent/linked with the medical terms present in a query using Dependency Parser [65] are included in keywords (See step 2 in *Figure 1*).

3.4. Document Retrieval

In the document retrieval module, documents in Kaggle COVID-19 Open Research Dataset [10] are used. The Allen Institute for AI (AI2), in line with other partners at The White House Office of Science and Technology Policy (OSTP), the National Library of Medicine (NLM), the Chan Zuckerburg Initiative (CZI), Microsoft Research, and Kaggle, coordinated by Georgetown University's Center for Security and Emerging Technology (CSET), together released the initial version of CORD-19. This resource compose of large and growing collection of publications and preprints on COVID-19 and previous coronaviruses such as SARS and MERS. CORD-19 is a resource of over 500,000 scholarly articles, including over 200,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses, and it is updated regularly. Besides the CORD-19 dataset, MedRxiv and PubMed databases are used as supplementary sources to further enhance the results of the document retrieval module. To retrieve articles efficiently, Anserini indexing [66], which uses OKAPI BM25 for indexing the paragraphs of the articles (e.g., article id + paragraph id) is applied. Query

and keywords are fed in a concatenated way to the Anserini (which holds the indexes of the documents in CORD-19 dataset), MedRxiv, and Pubmed. The articles published before the post's publication date are taken into account to prevent any data leakage and bias. If no postdate or claim date is present, all the results are retrieved (See step 2 in *Figure 1*).

3.5. Evidence Selection

In this section, sentences related to the check-worthy statement as potential evidence to the claim itself are selected from the retrieved article paragraphs. In order to select the relevant parts, BIOBERT [67] question answering model is employed instead of the sentence similarity model even though the claim is not a question. Although sentence similarity models were highly used in FEVER (Fact Extraction and Verification) [26, 68] tasks, with the help of QA models, it is speculated that the search may grasp the nuance and semantic meaning of the query better than sentence similarity models. In addition, the query, which was formed by the content of user posts, is written informally. Therefore, finding similar sentences with the query may not give the desired outcomes. In line with the proposed approach, Google also employs BERT question answering model for its searches [45]. In the proposed pipeline as a QA model, BIOBERT, based on BERT architecture, trained in the biomedical corpora is used.

Since BERT can handle a maximum of 512 tokens at once, paragraphs are split into chunks of 512 words with 10% overlapping words. In the consequent splits, a portion of the words are overlapped with each other to preserve the semantic meaning of the splits. Then QA model is applied to all chunks, and the answer with the highest score is labeled as the final answer. After that, to ensure completeness, sentences including the answer and its preceding one are retrieved. Then the full answers are sorted with the Universal Sentence Encoder according to their similarity score with the query [69], and evidence with a 0.4 and above confidence score is labeled as gold evidence. This score is determined by a brute-force approach, we score between 0.2-0.8 are evaluated and best performing one as taken as threshold for confidence score. The upper bound of the number of answers returned is set to the number of retrieved documents. As a rule of thumb, the number of retrieved documents from the CORD-19 document is set to 10. For the documents coming from the MedrXiv and Pubmed platforms, there is no explicit limit for the number of documents retrieved. However,

Query: CDC recommends people shave their facial hair to prevent coronavirus.

	Lucene ID	Evidence	Confidence	Title/Link	Publish Date
1	m5jpsxc4.00017	Wearing medical headgear does not offer additional protection, but it might help reduce unintentional hand-face contact and makes putting on the protective equipment without contamination easier. Make sure that FFP2 / 3 masks sit tightly ; bearded men may require shaving.	0.731351	Coronavirus disease 2019 (COVID-19): update for anesthesiologists and intensivists March 2020	24/03/2020
2	uu8ff703.00002	Keeping all this in mind, we recommend that hygiene rules be very strictly adhered to : nails cut as short as possible, hair tied back (it too can be contaminated with the virus) and avoidance of eyelash extensions. It would also be good to shave beards , taking into account the sebum secretion in beard hair ; however, this could be a problem for those who need to maintain beards for religious purposes.	0.712648	Observations about sexual and other routes of SARS-CoV-2 (COVID-19) transmission and its prevention	30/05/2020
3	skkpkqw6.00005	Therefore, the use of viral filters and closed suctioning, airflow changes, and negative pressure air pollution in closed environments, with the probability of infection, is recommended (Malhotra et al. 2020). Thus, people in contaminated environments should use a mask and, if there is no mask, should use face masks (Bowdle and Munoz-Price 2020). Facial hair can provide a way for SARS-CoV-2, to penetrate (Malhotra et al.	0.664152	The role of environmental factors to transmission of SARS-CoV-2 (COVID-19)	15/05/2020
4	ch8hpw62.00005	Qualitative mask-fit testing should ideally be performed in advance, as correct face mask and size are needed to ensure a proper seal. Facial hair at the face-mask interface promotes seal leakage and may decrease protection. 4 We strongly recommend shaving facial hair. Gloving : Although not needed to be sterile, always use extended-cuff gloves.	0.58742	Common breaches in biosafety during donning and doffing of protective personal equipment used in the care of COVID-19 patients	14/04/2020

Figure 2 An example outcome for a given user post and its retrieved evidence pairs. The query stands for the check-worthy claim extracted from the user post. Lucene ID stands for the unique article id and the paragraph of that article. The Evidence column shows the retrieved evidence sentences from the paragraph. Confidence stands for the similarity score between the evidence and the check-worthy claim. Title/Link indicates the article title and url of that article. Publish Date is the publication date of the article.

since evidence are sorted according to its similarity with the query, limiting the number of retrieved documents is insignificant. It is finally ended up with a query, its answers, and the confidence scores between the query and the answers.

Query: Claim that coronavirus stays in the throat 4 days before reaching the lungs and that gargling lukewarm water with vinegar and salt can kill it.

	Lucene ID	BERT-SQuAD Answer with Highlights	Confidence	Title/Link	Publish Date
0	nl55rm8o.00012	Further randomized studies need to be carried out to evaluate the effectiveness of these alternatives in management of SARS-Cov2. Gargling with naturally occurring sea water or salt solution containing chloride ions (Cl−) inhibit a number of viruses, including the coronavirus from the throat mucosa resulting in a reduction of the viral infection period with an average of 2. 5 days and a reduction in viral shedding .	0.692914	Dry Taps? A Synthesis of Alternative "Wash" Methods in the Absence of Water and Sanitizers in the Prevention of Coronavirus in Low-Resource Settings	24/06/2020
1	p7zor88v.00003	Possible Beneficial Role of Throat Gargling in the Coronavirus Disease Pandemic A randomised trial study in Japan showed that throat gargling with tap water 3 times a day significantly reduced the incidence of upper respiratory tract infection (UTRI) by 36 %.	0.644386	Possible Beneficial Role of Throat Gargling in the Coronavirus Disease Pandemic	03/06/2020
2	q5vit1dk.00013	More than one-third participants agreed about eating citrus fruits and gargling salt water can help prevent infection. According to WHO gargling warm or saltwater and consuming citrus fruits will not kill novel-corona virus [27].	0.64246	Indian communitys Knowledge, Attitude & Practice towards COVID-19	09/05/2020
3	2ngwjp3r.00016	CONCLUSIONS : Gargling might be effective in preventing febrile diseases in children. The results of this study and the site of action of gargling suggest that gargling might prevent viral upper respiratory tract infections .	0.630907	Gargling for Oral Hygiene and the Development of Fever in Childhood: A Population Study in Japan	05/01/2012
4	2ngwjp3r.00015	Gargling for Oral Hygiene and the Development of Fever in Childhood : A Population Study in Japan BACKGROUND : Fever is one of the most common symptoms among children and is usually caused by respiratory infections. Although Japanese health authorities have long recommended gargling to prevent respiratory infections, its effectiveness among children is not clear .	0.596628	Gargling for Oral Hygiene and the Development of Fever in Childhood: A Population Study in Japan	05/01/2012
5	c2nkp1gq.00012	The technique of nasopharyngeal wash to prevent the virus from inhabiting and replicating in the nasal and pharyngeal mucosa has been suggested to be useful in reducing symptoms, transmission, and viral shedding in cases of viral acute respiratory tract infections. In rapid systematic review, we found studies showing some improvement in prevention and treatment of upper respiratory tract infections. We postulate that hypertonic saline gargles and nasal wash may be useful in prevention and for care of patients with COVID-19 .	0.593045	Nasopharyngeal wash in preventing and treating upper respiratory tract infections: Could it prevent COVID-19?	04/05/2020

Figure 3 An example outcome for a given user post and its retrieved evidence pairs. The query stands for the check-worthy claim extracted from the user post. Lucene ID stands for the unique article id and the paragraph of that article. The Evidence column shows the retrieved evidence sentences from the paragraph. Confidence stands for the similarity score between the evidence and the check-worthy claim. Title/Link indicates the article title and url of that article. Publish Date is the publication date of the article.

Since the ultimate aim of this pipeline is to map formal text, the answer retrieved from the article to informal text, the query, it is desired to avoid scientific or experimental sections as answers because they are too technical and hard to interpret for end-users. Instead, it is desired to have simple and direct information from the answers retrieved from the article sections. Therefore, only the abstract, introduction, discussion, result,

or conclusion sections of an article are considered. In an otherwise situation, the retrieved answer is discarded (See steps 3-4 in *Figure 1*). *Figure 2* and *3* show the outcome of the model in HTML format for different user posts and retrieved evidence pairs.

Figure 2 shows an example of the program output, where all the returned evidence supports the claim stated in the query (check-worthy statement extracted from user posts), whereas mixed results are encountered in *Figure 3*, where only some of the articles support the claim in the query while the remaining support the opposite. This phenomenon has been frequently observed in the COVID-19 pandemic since it was a novel coronavirus (later named SARS-CoV-2), and the preventive measures/treatments concerning the virus have constantly been changing over time. Hence, although articles with the publication date before the claim date itself are considered in user posts, contradicting articles are encountered among those. For example, "*WHO had previously said there was not enough medical evidence to support members of the public wearing a mask unless they were sick or around people with the coronavirus.*" Then June 8, 2020 -- The World Health Organization has changed its stance on wearing face masks during the COVID-19 pandemic and said, "*WHO advises that governments should encourage the general public to wear masks where there is widespread transmission, and physical distancing is difficult, such as on public transport, in shops or in other confined or crowded environments.*" [70, 71].

3.6. Textual Entailment

Initially, a preprocessing step was conducted to improve the performance of entailment scoring. Since entailment scores between a formal (sections from the scientific article) and informal text (user posts) are measured, it is required to bring them to a similar level in terms of writing style. As a solution for this situation, summarization models are utilized. For the query side, a claim located in a user post might include irrelevant text and numerous noisy characters; to emphasize the claim rather than the opinion, text simplification is made on the check-worthy claim. For this purpose, PEGASUS trained on CNN dataset [72], a state-of-art abstractive summarization model trained on C4 and HugeNews is utilized. With the usage of this model, the claim included sentences rewritten more formally and simpler.

Similarly, for summarizing the answers (retrieved article sections) and shortening the full answer text, in other words to emphasize more the critical information in the text, DistilBART trained on the CNN dataset [73] is employed. In this way, again, it is aimed to bring both informal query and formal answers to a similar level by rewriting the texts while improving the entailment performance between them.

In addition, a dictionary of MeSH terms of the medical words is used to expand and standardize the term list (e.g., replace all nCov-2019, sars-cov-2 with COVID-19). It is speculated that standardizing nouns used in claims and evidence and using the same

noun for all possible variations of medical terms in the comparison will improve the entailment model's performance. This speculation has been experimented with in the ablation study in section 4.3.2.

Finally, for the entailment model, BERT model trained on bio-medical PubMed corpus [74] including 3.1B words/ 21GB of textual data, fine-tuned on MNLI [75] is used. The model is trained on various medical articles with various topics found in Pubmed Corpus. It is speculated that it might perform well compared to a model trained on general-purpose corpus as it can identify medical words, thus can better understand relations between two given texts (See step 5 in *Figure 1*).

3.7. Heuristic Verdict Assignment

After calculating the entailment score between all the candidate answers to the query, the final verdict, which will be one of Support, Contradict, or Neutral (Not Enough Info or *NEI*), is determined according to a heuristic. If there is no answer, with a similarity score between the query and evidence higher than 0.4, it is assigned as Neutral. If the "neutral" score between the query and answer is higher than 0.5, it is counted as a Neutral vote. For the other cases, the contradiction and entailment scores between the query and the answer are considered. If the entailment score is higher than the contradiction score, then one vote is added for the Support label. If the contradiction score is higher than the entailment score, one vote is added for the Contradict label. In the end, a majority vote is taken between the Support and Contradict label votes to determine the final verdict. However, if the votes for the Support and Contradict labels are equal, the average of those votes is considered a tie-breaker, and the highest is taken as the final verdict. If the outcome is Neutral, it means no information supports or contradicts the claim in the medical articles. This is an expected outcome since not every claim on social media can be verified using health articles, such as in the case of a post including "*Shoes can spread COVID-19*" sentence (See step 5 in *Figure 1*). The given threshold scores are determined with a brute force approach by taking independent runs for both similarity and entailment thresholds. Values from 0.1 to 0.9 with a 0.1 increment have been experimented separately, and the best performing threshold for those is determined as a threshold score.

CHAPTER 4

EXPERIMENTS

4.1. Datasets

I have conducted the experiments using two different datasets of user posts.

- CoAID (COVID-19 heAlthcare mIsinformation Dataset) includes variety of misinformation regarding with the COVID-19 pandemic, those misinformation include fake news on various websites and known social platforms, along with users' social engagement (comments) about such news and posts. CoAID dataset comprises 4,251 news, 296,000 related comments, 926 social media posts regarding to COVID-19. Here, manually labeled ground truth claims were used as search queries to automatically retrieve related tweets and label them accordingly with the corresponding claim's final verdict. The retrieved social media posts and news include attributes such as; user ID, tweets, replies, favorites, retweets, and location. In total, 10,439 tweets about fake news articles, 141,652 tweets about true news articles, 484 tweets about fake claims, and 8,092 tweets about true claims were obtained. In the experiments, I have used the tweets including user posts only (henceforward called "Claims") or the tweets comprising "news titles" (henceforward called "News"). An example of a user post sharing a news title is as follows: "Look at this and please share" and the title of the news shared in the tweet is "New flu drug drives drug resistance in influenza viruses".

- COVID-19 Rumor Dataset include manually labeled 6,834 data (4,129 rumors from news and 2,705 rumors from tweets) with sentiment and stance labels. The true status of the rumors is retrieved from fact-checking websites. Moreover, it composes 32,750 reposts, which includes news rumors and 34,847 retweets for Twitter rumors, in addition to a manually labeled stance feature. However, while the rumors were collected from various topics, the number of posts, including the rumors about medical claims, is lower than the CoAID dataset.

4.2. Results

To the best of my knowledge, there is no unsupervised domain-specific method proposed for fact extraction and verification of informal medical texts (e.g., user posts, tweets, news titles in tweets) in the literature. On the other hand, there are approaches for extraction and verification of claims, using Wikipedia, for the challenge dataset FEVER (Fact Extraction and Verification task). In the FEVER task, a given factual claim includes one or more entities (resolvable to Wikipedia pages), and the system is required to extract textual evidence (sets of sentences from Wikipedia pages) that supports or refutes the claim. Using this evidence, the claim is labeled as Supports, Refuses, or NotEnoughInfo (if there is no sufficient evidence to either support or refute it). Here, the baseline model is a simple pipelined system comprising document retrieval, sentence-level evidence selection, and textual entailment. Although the tasks are similar, I cannot evaluate the pipeline using this task's dataset as the proposed pipeline is specialized for the medical domain, whereas the Wikipedia dataset in the FEVER task is generic. In addition, the models developed for the FEVER task are fine-tuned considering the Wikipedia corpus. As a result, I use the supervised models proposed in the literature as baseline models for comparison. However, direct comparison with the supervised models is not also possible for three main reasons, 1) The proposed pipeline does not only label claims, but it retrieves the evidence from the medical articles and then assigns the verdict. 2) Since the proposed pipeline does not learn from the data, the decision of training and testing dataset splits is important for comparison. 3) Supervised model assigns one of the labels from the dataset given (True or False), whereas the proposed pipeline has three distinct labels (Supports, NotEnoughInfo, Refuses).

Consequently, to ensure a fair comparison with the supervised models, I use different data sampling schemes/splits to create training and testing datasets. For instance, the datasets are split to reflect newly emerging topics in user posts. As a baseline supervised model, I used the BERT-Base-Uncased model for sequence classification [76] (henceforward called "Baseline 1"), and a simple CNN with one convolution and one fully connected layer ("Baseline 2"). For BERT and CNN models, I have utilized the pre-trained word embeddings BERT Tokenizer and Glove embeddings of dimension 100, respectively. I have also applied the same tweet preprocessing steps for both the proposed pipeline and the supervised baseline methods. I kept the epoch size high for the supervised model and employed early stopping criteria to ensure the model's convergence. Both models were trained with the generic Adam optimizer and used the binary cross-entropy loss.

The proposed system generates three distinct labels (Supports, NotEnoughInfo, Refuses), although the dataset includes two labels (True, False). I kept the NotEnoughInfo category because it is needed to show the following outcomes: (1) the pipeline could not retrieve the related article sections concerning the claim successfully. (2) Given the related article sections, there is no information supporting

or refuting the claim in the scientific articles. (3) The claim is not subject to a medical article (e.g., **The social media post is**; *Show me your papers!? What is a coronavirus immunity certificate? US may start issuing them G8M personal sovereignty under God.* **The extracted claim is**; "*US may start issuing coronavirus immunity certificate G8M personal sovereignty under God*"). In other words, the claims that are not related to any topic in scientific articles may not be classified correctly by the pipeline due to the evidence-based classification approach of claims and zero-shot-based learning of the pipeline. It is also difficult to manually select and label user posts that cannot be checked against scientific articles. In the end, I have mapped two ground-truth class labels as follows: Supports stands for True, Refuses stands for False, and if there is no evidence retrieved to refute or support the claim or NLI model gives Neutral output, which is labeled as NotEnoughInfo. The further classification of the NotEnoughInfo label is left as future work. However, to mitigate the problem between the pipeline and dataset labels and for a fair comparison with the supervised models, claims labeled as NotEnoughInfo by the proposed pipeline are fed to the Baseline 1 model, and metrics are calculated accordingly. Two evaluation schemes were considered.

4.2.1 First Evaluation Scheme

I aim to assess whether the pipeline can correctly classify claims in user posts, especially on newly emerging COVID-19 topics. Supervised methods are limited to classifying such posts successfully as it is unlikely to observe similar posts in the training dataset. In this evaluation scheme, which uses an out-of-time sampling approach, first, the tweets are sorted according to the dates of posts. Then, I incrementally split the dataset timewise. In the first split, the first 10% of the dataset is utilized as training, and the rest is separated for testing.

Table 1: The test results of the CoAid (Tweets + News Titles), which were split according to user postdates. T, A, F1, P and MCC stand for Training percentage, Accuracy, F1 score, Precision, and Matthews Correlation Coefficient respectively.

Model	T	A	F1	P	Recall	MCC
Baseline 1	10%	0.65	0.66	0.95	0.51	0.16
	20%	0.83	0.84	0.97	0.74	0.35
	30%	0.90	0.91	0.97	0.86	0.49
	40%	0.93	0.94	0.98	0.90	0.58
	50%	0.94	0.95	0.98	0.92	0.59
	60%	0.94	0.96	0.97	0.95	0.61
	70%	0.94	0.96	0.97	0.95	0.61
	80%	0.95	0.97	0.97	0.97	0.62
	90%	0.96	0.97	0.98	0.96	0.63
Baseline 2	10%	0.63	0.68	0.64	0.73	0.26
	20%	0.71	0.74	0.76	0.72	0.41
	30%	0.75	0.80	0.78	0.82	0.48
	40%	0.76	0.81	0.82	0.80	0.49
	50%	0.78	0.84	0.85	0.83	0.50
	60%	0.80	0.87	0.86	0.88	0.44
	70%	0.83	0.89	0.90	0.88	0.50
	80%	0.85	0.91	0.92	0.90	0.41
	90%	0.87	0.93	0.98	0.88	0.50

In the second split, the first 20% of the dataset is used as training, and testing is done on the rest of the dataset. This process is repeated until 90% of the dataset is reserved for training. The test results of the baseline models in Table 1show that as the training percentage increases, the supervised models' performances increase. Even with small training dataset percentages comparing the test percentage, the models perform quite well. Further analysis of the dataset shows that similar tweets may have been posted at different times, hence appearing in both training and testing datasets. In other words, there is a data-leakage problem present in this scheme. Due to the data collection methodology chosen for constructing these datasets, the topics of the user posts vary almost uniformly over time. This means the same user post can be encountered at different periods, and this situation contradicts the pipeline's aim, which is to detect newly emerging claims. Therefore, I conclude that this scheme is biased and prone to data leakage, and the models' capability to responding newly emerging claims cannot be tested properly.

4.2.2 Second Evaluation Scheme

In order to mitigate the problem in the first evaluation scheme, it is preferred to cluster the tweets by using ktrain's zero-shot topic classification model [77] in both datasets separately to simulate newly emerging topics and to prevent the possibility of any data leakage into the testing phase. *k-1* number of training dataset samples are created, where *k* represents the number of clusters. To be more specific, the training sample k_1 included the data points in cluster 1, and the remaining ones were reserved for the

testing dataset. Training dataset samples are gradually formed by including new cluster data instances into the training datasets. For instance, while the training sample k_2 had all the data points in both clusters 1 and 2, the other cluster instances were reserved for the testing dataset. Finally, separate models are fitted for each dataset configuration, making a total of *k-1* different models. In this way, it is ensured to reduce the likelihood of seeing similar posts both in the training and testing datasets. This validation approach will also help assess the generalizability of the pipeline. Since the proposed pipeline does not require training, there is no significant difference expected in the pipeline's performance while increasing the dataset size based on the clusters. The same dataset splits were also used for the baseline models *Baseline 1* and *Baseline 2.* The results were reported by computing the average of the metrics over *k*-1 iterations. For clustering, the k-means algorithm is used and KL Divergence is used as performance metric to determine the number of clusters.

Figure 4 The testing results of the proposed pipeline and the baseline models. The vertical axis represents the value of metrics, and the horizontal axis represents the cluster count

The number of clusters, *k*, is determined by the elbow method heuristic which can be seen in Figure 5. Cluster numbers were chosen as 91 for the CoAID dataset and 51 for the COVID-19 Rumors dataset which corresponds to elbow of the curve in Figure 5. To handle the changing training dataset sizes, the epoch number is kept high and early stopping criteria is used.

Figure 5 The KL divergence results for different total cluster numbers for given dataset

In addition, there is a significant class imbalance problem in favor of True posts in the CoAID dataset (7% False, 93% True posts). As a remedy, under-sampling is carried out on the training dataset by selecting True posts from each cluster in equal numbers. The "COVID-19 Rumors" dataset is a balanced dataset (42% True, 58% False posts); therefore, no under-sampling is done for this dataset in the training phase. As for performance metrics, Matthews Correlation Coefficient (MCC) [78], TNR [79] and F1 scores are used. As can be seen in Table 2, the proposed pipeline surpasses the baseline models in classifying False posts (TNR) and F1*, which gives more emphasis to predicting "False" posts, a requirement preferred for fake posts/news detection. More importantly, the pipeline gives a steady performance across all runs unlike supervised models as can be seen in Figure 4.

Table 2: The results of the CoAid (Tweets + News Titles), COVID-19 Rumors. MCC refers to Matthews Correlation Coefficient. F1* refers to when the desired class (TP) is the true identification of "False" posts. B1 and B2 refer to Baseline 1 and Baseline 2 models respectively.

Dataset	Metric	PP	B1	B2
CoAID	Acc.	0.90	0.90	0.89
	F1	0.94	0.91	0.91
	F1*	0.54	0.46	0.32
	TNR	0.58	0.38	0.31
	MCC	0.50	0.46	0.36
	Recall	0.92	0.98	0.96
COVID-19 Rumors	Acc.	0.84	0.91	0.73
	F1	0.79	0.87	0.57
	F1*	0.87	0.93	0.61
	TNR	0.83	0.92	0.60
	MCC	0.66	0.77	0.37
	Recall	0.82	0.97	0.96

The analysis shows that the pipeline underperforms, particularly in social media messages expressing an opinion or including popular news rather than medical facts or claims, and cannot be verified from the medical articles. i.e., "*I told you guys that someone or perhaps many would die from listening to Trump and Trump's admin.*

Health officials warn against self-medicating with chloroquine for coronavirus after man dies from taking fish tank cleaner". Such opinion or daily news-related posts cannot be validated using the proposed pipeline since the pipeline checks the statements against the medical articles. To further analyze this situation, I have conducted an experiment on user posts labeled as "Claim" only, considering those posts comprise significantly more non-medically verifiable statements than the user posts labeled as "News". i.e., news title: "*Antiviral used to treat cat coronavirus also works against SARS-CoV-2", claim: "I spent several minutes this morning chatting with the first volunteer in the Oxford COVID-19 vaccine trial via Skype.*" The motivation of this experiment is to demonstrate that the baseline models might be learning features that are specific to post characteristics or words rather than learning the semantic meaning of the actual content. When the "Claims" are taken from the dataset, the dataset size has significantly reduced (6275 against 54221 observations in the whole dataset). The baseline models are expected to have similar or slightly degraded performance on the dataset composed of only "Claims" compared to the whole dataset, and since there are fewer medical claims to check in this subset, the proposed approach is expected to underperform further comparing to the whole dataset. Table 3 shows the results of this experiment. As expected, the baseline models perform better than the proposed pipeline when the claim is opinion or popular news-related rather than medically related.

Table 3: The results of the models on the CoAID "Claim" Posts Only

Metric	PP	B1	B2
Accuracy	0.80	0.83	0.81
F1	0.89	0.83	0.85
TPR	0.96	0.97	0.95
TNR	0.20	0.33	0.25
MCC	0.27	0.42	0.26

The speculation on the "CoAid" dataset also holds for the "COVID-19 Rumors" dataset. "COVID-19 Rumors" dataset includes a significant number of claims which express an opinion or are about popular news. For instance, "*China regulator says epidemics impact on industry major in February*". Another example is, "*If you do not have insurance and can't afford to take the $3,200 test for the virus ($1,000 with insurance), DONATE BLOOD. They HAVE to test you for the virus in order to donate blood.*" The fact that the pipeline labeled 64% of the user posts as NotEnoughInfo, comparing to 28% in the CoAID dataset also supports the speculation.

4.3. Ablation Studies

CoAid dataset is employed to conduct the ablation studies because it comprises a significant amount of medically verifiable claims. As the pipeline outputs three

distinct labels (Supports, NotEnoughInfo, Refuses), the metrics are used for reporting multi-class classification performance specifically; accuracy, F1, Matthews Correlation Coefficient, and precision. Since the pipeline does not include any supervised training, the ablation studies are constructed using the whole dataset. Since the different pipeline configurations are compared in ablation studies, the results are also reported for the classes Supports/NotEnoughInfo/Refutes.

4.3.1 Study 1

I conducted an ablation study to measure the impact of the Natural Language Inference Model chosen for the proposed pipeline. Two models are compared for this purpose: (1) XLNET [80] trained on the composition of SNLI [81], MLI, FEVER [26], ANLI [82], and NLI [83] datasets (hereafter called ***M1***) and (2) BERT trained on bio-medical PubMed corpus [74] then fine-tuned on MNLI [75] (hereafter called ***M2***). The bio-medical PubMed corpus has medical articles on various medical topics. The results show that *M2* outperforms *M1* with a slight margin. Also, 36% *NotEnoughInfo* labeled class posts reduced to 28% with *M2*, as the labeling performance improves specifically on the user posts, including medically verifiable claims ("News" as discussed in section 4.1.2).

Table 4: The results of the ablation study for Natural Language Inference Model

Metric	M1	M2
Acc.	0.67/0.73 /0.67	0.66/0.73/0.87
F1	0.72/NA/0.37	0.78/NA/0.38
MCC	0.05/NA/0.23	0.06/NA/0.23
Precision	0.77/NA/0.59	0.94/NA/0.34
Recall	0.67/NA/0.27	0.66/NA/0.43

The superior performance of *M2* can be explained by better identification of medical terms, which results in more accurate matching between the claims and the evidence. For instance, while the XLNET model tokenizes "*naloxone*" word as "*na-lo-xon-e*", BERT trained on PubMed corpus tokenizes it as "*naloxone*" thus preserving the meaning of the noun and improving the results. Moreover, when the XLNET model cannot relate the words between the evidence and claim, especially in the cases where both include medical words, the pipeline tends to give "Neutral" (Not Enough Info") as a result.

4.3.2 Study 2

This ablation study investigates whether the MeSH term addition and Query and Answer summarization significantly improve the performance. When the textual entailment score is computed after summarization is applied on both query and answer, the model overperforms, as shown in Table 4. Furthermore, if I do not use summarization, the proposed pipeline labels 48% of the user posts as *NotEnoughInfo* in CoAid Dataset. On the other hand, when I use summarization, the pipeline's

labeling performance increases, and the *NotEnoughInfo* label percentage drops to 28%. As a result, it is observed that the labeling performance increased (52% to 72%).

Table 5: Ablation study results showing the effect of summarization and MeSH terms CoAID dataset

Metric	PP	w/o Summarization & MeSH	w/o Summarization
Acc.	0.66/0.73/0.87	0.52/0.56/0.56	0.64/0.70/0.69
F1	0.78/NA/0.38	0.50/NA/0.24	0.71/NA/0.33
MCC	0.06/NA/0.23	-0.12/NA/0.0	0.01/NA/0.20
Precision	0.94/NA/0.34	0.55/NA/0.85	0.85/NA/0.56
Recall	0.66/NA/0.43	0.45/NA/0.13	0.61/NA/0.23

4.3.3 Study 3

In this study, we investigate the effect of using only the abstract, all paragraphs, and the selected paragraphs on the performance of the pipeline (Abstract, Introduction, Conclusion, Discussion and Result). As stated, since the pipeline aims to map informal user posts to formal medical articles, instead of highly scientific evidence, direct and simple evidence is needed. For better matching, the scientific method names, statistical analysis results, and formal domain-specific explanations should be avoided. Also, providing such simple explanations from the articles will make the evidence set more comprehendible for end-users.

Table 6: The results of the third ablation study for evidence retrieval.

Metric	w. Selected Paragraphs	w. Abstracts Only	w. All Paragraphs
Accuracy	0.66/0.73 /0.87	0.57/0.58 /0.58	0.53/0.58/0.58
F1	0.78/NA/0.38	0.54/NA/0.34	0.51/NA/0.32
MCC	0.06/NA/0.23	0.02/NA/0.21	0.01/NA/0.19
Precision	0.94/NA/0.34	0.56/NA/0.87	0.54/NA/0.83
Recall	0.66/NA/0.43	0.52/NA/0.21	0.49/NA/0.20

Table 6 shows that including the highly scientific sections of the articles, such as the method and the results sections, causes a decrease in the pipeline performance. The model using only the abstract section of an article for evidence retrieval performs slightly better than the model, which considers the evidence retrieved from the whole article. It is speculated that due to the complexity of the answers found in the scientific sections of an article (i.e., the method, the experiment section, etc.), the NLI model could not find any relation between the evidence and the claim since the expressions, language and writing style used is considerably different. Therefore, I have considered the "*Abstract*", "*Introduction*", "*Results*", "*Discussion*" and "*Conclusion*" sections to increase the probability of the QA model in finding relevant answers to the query. In addition, the pipeline using these selected paragraphs from the articles labeled 28% of the predictions as *NEI*, compared to 42% in the abstract only and 43% in all paragraph

configurations. Therefore, it can be concluded that selecting related paragraphs improves the labeling performance of the pipeline.

CHAPTER 5

DISCUSSION & FUTURE WORK

This study proposes a pipeline that performs fact-checking and verification of informal user claims using scientific medical articles specifically on the COVID-19 domain. The pipeline gives NotEnoughInfo for the claims, which cannot be verified from the medical articles. In future work, I plan to implement a supplementary pipeline, which will be able to classify such claims.

All the codes and tests are done in the Google Colab VM's. The system settings of the machine as follows: *GPU: Tesla P100-PCIE-16GB (UUID: GPU-29178be5-63d1-377b-3122-61552b5e3030), NVIDIA-SMI 460.32.03, Driver Version: 460.32.03, CUDA Version: 11.2, Intel(R) Xeon(R) CPU @ 2.20GHz, L3 cache: 56320K, 127G Available Memory.*

The time for fact-checking a user post from Twitter takes 3.4 seconds per tweet on average. The verdict (True, False, NEI) and the related evidence are returned by the pipeline as model output. It has been observed that a significant amount of this time passes on REST API calls for the retrieving MeSH Terms, Check-worthy statement extraction, and database connections for retrieving the medical articles. For the deep-learning models, I utilized the GPU cores and parallelization; however, this cannot be done for the API calls and database connections. Nevertheless, the pipeline can be scalable for larger applications, especially if it can be optimized to reduce the time consumption of non-parallelizable operations in the pipeline. As a future work, to decrease the time it takes in the REST API calls for MeSH terms, a dictionary might be created and kept in cache beforehand for the common words, therefore the pipeline does not need to make an REST API request every time, instead it searches the dictionary in the cache and if the specific word does not present in there it may make the API call. In this way, I anticipate the time it takes might go down to 2.5 seconds on average.

In the proposed pipeline it is observed that, pipeline is mostly underperforming on the claims which includes non-medically verifiable statement as check-worthy statement. This necessitates the requirement of separating non-medically verifiable claims from medically verifiable claims. Then there should be another fact-checking path for non-

medically veriable claims. As a future work, it is planned to add another component which seperates non-medically verifiable claims from the medically veriable claims. After that, again as a future work it is planned to build a separate path for fact-checking non-medically verifiable claims. With this way, it is aimed to increase the performance and applicability of the model in general.

CoAID dataset includes automatically labeled user posts, which were determined based on the labels of the search queries given to each claim. However, this automatic labeling approach resulted in some data quality issues [84] and caused some user posts to be mislabeled. For instance, in the cases of the presence of sarcasm or contradiction in the tweet, I observed incorrect labels; for example, **User Post:** "*To suggest COVID-19 is just like the flu is to buy in to disinformation fellas." is labeled as "False*". As seen from the user post, there is sarcasm in the sentence, which makes a claim "True" inherently. These user posts affect the evaluation performance negatively. In future work, I will implement a semantic classification model in the pipeline for detection of that sarcasm and adjust the final verdict about the statement accordingly.

After manually inspecting the results, I observe that the majority of mislabeling comes from the textual entailment step, i.e., sentences including the list items or partially sharing the same sub words (e.g., **Claim**: *Coronavirus Covid-19 is not transmitted through the houseflies.* **Evidence**: *The present pediatric case of COVID-19 was acquired through household transmission*). Therefore, I aim to improve the textual entailment model's performance by training or implementing a model dedicated to this task in future work. Also, the tokenization algorithm's performance is important for the model's performance.

In the pipeline, I have used ClaimBuster API [31] to identify the claims in the text; however, since ClaimBuster was trained on the political claims, I speculate that a model trained on health-related documents might improve the "check-worthy sentence" detection performance. For this purpose, papers in the CheckThat [32] challenge can be analyzed and fine-tuned in the pipeline. For example, in Martinez Rico et al. [85], transformer models are used to extract linguistic features that identify fraudulent articles. On the other hand, a simple Gradient Boosting classifier uses linguistic features extracted through the LIWC tool, TF-IDF text features, and a TF-IDF representation of domain names retrieved from a Google search. With this pipeline, they achieved 1st place in the English version of the check-worthiness task. In order to improve the proposed pipeline's performance, models that achieve a high score in CheckThat can be fine-tuned for the proposed pipeline.

Other than ClaimBuster API, solely considering the similarity score between the answers and the query to rank the answers may be misleading for selecting the most relevant answers. For instance, in the cases like when the query and the answer are very similar sentences, except their subject, e.g., **Claim**: *The new coronavirus cannot be transmitted through mosquito bites*, **Evidence**: *Sindbis virus (sinv) a positive-*

stranded RNA virus that causes mild symptoms in humans is transmitted by mosquito bites. As in the given example, the subjects of the claim and evidence are different, whereas the sentences are alike. Although both sentences mention the transmission with mosquito bites, the diseases mentioned are different. Consequently, although the evidence is unrelated to the claim itself, a high similarity score obtained between evidence and claim causes incorrect ranking result in terms of evidence-claim relation. I will implement a better ranking/sentence similarity algorithm for answers in future work. I will also consider different features other than sentence similarity to achieve a more precise ranking.

I did not find any baseline models in the literature for evaluating the document retrieval, evidence selection, or labeling performances on medical articles about COVID-19 to the given claims. Nevertheless, I investigated the performances of other pipeline-based approaches proposed for different domains. For instance, in FEVER 2018 Shared task, which is similar to the given problem, the winning pipeline [83] has an accuracy score of 0.69 and F1 score of 75.7/69.4/63.3 *Supports/ NotEnoughInfo /Refutes*, respectively. Although a direct comparison might be misleading due to the differences in the problem domains and datasets, the performances reported for FEVER tasks might give valuable insights in terms of how much performance similar pipelines might achieve. In this domain, the pipeline achieved 0.64 accuracy on average and 0.78/NA/0.37 Supports/ NotEnoughInfo/ Refutes F1 score, indicating that the performance achieved is reasonable and scalable.

To evaluate the performance of document retrieval and evidence selection, I needed to manually annotate the evidence related to the claims and evaluate the document retrieval and evidence selection performance. For this purpose, I aim to implement a solution similar to FEVER SCORER [86] in the future. In Saakyan et al. [87], the authors mentioned FEVER like dataset; however, the evidence was collected from Google Search results instead of COVID-19 related medical articles. Therefore, I can not utilize this dataset either for evaluating document retrieval or evidence selection. Since the datasets I used contained only verdict information and no evidence related to that verdict, I was unable to consider the metrics, including evidence recall and evidence precision, in this study. Therefore, I left this step as future work.

CHAPTER 6

CONCLUSION

In this thesis, I proposed a new zero-shot fact extraction and verification pipeline for user posts related to COVID-19 against the medical articles that have the potential to be applied to other health domains. The pipeline includes the preprocessing, claim extraction, keyword extraction & query enhancement, document retrieval, evidence selection, textual entailment, and verdict assignment steps. The pipeline provides final verdicts including "True", "False" and "Not Enough Evidence" as well as evidence corresponding to that verdict assignment. In this way, the fact-checking process ensures to give comprehendible results and can be entangled with the related evidence. The model is applied to the COVID-19 dataset collected from Twitter (CoAID) and COVID-19 Rumors Dataset compiled from various independent fact-checking platforms (e.g., Snopes, Politifact, etc.).

Back to the main research question, it is aimed to develop a fact-checking pipeline that does its fact-checking on evidence found in medically verified and peer review articles. In this research question, three sub research questions have arisen;

First, "Is it possible to map an informal medical claim to the formal medical articles on social media?"

In the proposed pipeline, informal medical claims in the user posts retrieved from Twitter are fact-checked against the medical articles retrieved from PubMed, MedrXiv, and the CORD-19 article repository. In order to achieve this, a question-answering model is applied for evidence selection; summarization models are applied for simplifying the retrieved evidences and check-worthy statements which are extracted from user posts, and a natural language inference is applied for calculating entailment score between simplified evidence and check-worthy statements in user posts. As a result, informal user posts mapped to the formal medical articles and related evidence regarding the claim, as well as the final verdict determining the correctness of the user post, are obtained as pipeline outputs. In addition, when the pipeline results are analyzed on the experimental datasets, it can be seen that the pipeline gives on-par/better results than the baseline models.

Second, "Does evidence-based fact-checking without explicit supervision for newly emerging medical claims perform on-par/better results than state-of-art supervised models?"

The proposed pipeline uses the zero-shot capabilities of existing supervised models. Therefore, there is no need for explicit supervision as stated before. It is known that one of the major weaknesses of supervised architecture is finding data to be trained on. In addition, since the misinformation spreads very fast, supervised models often become inadequate for determining misinformation in newly emerging claims. When the results of the experiments in section 4.2.2 are analyzed, it can be seen that the proposed pipeline achieves better and steady performance in newly emerging claims than the supervised baseline models. However, it should be noted that, if a research paper on the claim has not been previously published, proposed pipeline may not find usefull evidences related with the claim, in the situations like these the proposed pipeline return no related evidence or evidences which does not include enough info to contradict or entail with the claim.

Third, "Is it possible to improve medical document retrieval performance by using MeSH (Medical Subject Headings) tree structure for query enhancement?"

As stated before, there are many synonym words present in the medical domain. Different words in the medical domain and common usage may correspond to mainly the same (e.g., influenza, flu, common cold, etc.). It is observed that different medical articles have different medical term usages (nCov-2019, COVID-19, etc.) to cover all terms in the medical domain. In addition to medical terms found in a claim by BioBert, synonyms of them found by querying the MeSH tree are added to the search query as keywords. Also, using synonyms in the MeSH tree, all the medical terms, which correspond to the same thing, are replaced with one of the synonyms. In this way, the natural language inference model's performance is increased. The improvement of MeSH term usage is given in section 4.3.2, and as can be seen from Table 5, the improvement of MeSH term usage is significant.

The steps of the proposed pipeline are as follows:

First, the main check-worthy statements are extracted from raw user posts to eliminate noise in the claims. Then, irrelevant data (e.g., hashtag chunks, questions, etc.) is discarded using ClaimBuster API and the constituency parser. In order to preserve the semantic meaning and subject-object relationship in the main check-worthy statement, coreference resolution is used. Second, gold evidence is needed to compare them with the claim to determine the final verdict about a user post, whether it includes false information or not. For this purpose, document retrieval and evidence selection steps are employed. After retrieving the related documents, the claim is summarized to generate a simplified query with fewer noisy words in it while enhancing it with the MeSH words. Verbs directly depending on the medical terms found in the MeSH tree are included as well to focus more on the subject and verb relationship of the claim. After modifying the query (summarized claim + MeSH keywords), medical databases (Pubmed, Medrixv) and the dataset comprising medical articles (CORD-19) is queried to retrieve the related documents. In order to overcome the evidence selection problem from the paragraphs, question answering models are utilized. Evidence that corresponds to "Method" section of the articles is discarded to eliminate noise. Then the list of retrieved evidence is ordered according to their semantic similarity scores with the

query. The evidence with higher scores than the threshold is determined as gold evidence. Third, the final verdict is determined by evaluating the entailment score between the gold evidence and the claim. For this purpose, I initially summarized the evidence to decrease noise in the evidence and used Natural Language Entailment model to determine the entailment score. In the end, the final verdict is determined by counting the "entailment" and "refuses" labels in the gold evidence with a majority vote.

The pipeline is applied to the COVID-19 dataset collected from Twitter (CoAID) and COVID-19 Rumors Dataset. Compared with the supervised state of art models, the pipeline shows superior performance on these datasets in detecting fake information, including new emerging topics. The system can successfully use user posts as search queries and find relevant sentences from scientific health-related articles. In the end, it returns results/evidence, which is comprehendible for end users. It is expected that using this pipeline alongside the existing methods might help to deal with the spread of misinformation on social media.

REFERENCES

[1] The MSC 2020 at a Glance. (n.d.). Security conference. Retrieved October 17, 2021, from https://securityconference.org/en/msc-2020/overview/

[2] Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation. (2020, September 23). WHO. Retrieved October 17, 2021, from https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation

[3] Snopes. (2021, May 14). Mask Off: Researching Face Mask Rumors. Snopes.Com. Retrieved October 17, 2021, from https://www.snopes.com/collections/covid-face-mask-rumors/

[4] PolitiFact | Coronavirus. (n.d.). Politifact.Com. Retrieved October 17, 2021, from https://www.politifact.com/coronavirus/

[5] Twitter Developer. COVID-19 Stream. Developer.twitter.Com. Retrieved October 17, 2021, from https://developer.twitter.com/en/docs/twitter-api/tweets/covid-19-stream/overview

[6] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. Science 359, 6380 (2018), 1146–1151 Transactions of the Association for Computational Linguistics, 8, 64-77.

[7] Sunil Wattal, David Schuff, Munir Mandviwalla, and Christine B Williams. 2010. Web 2.0 and politics: The 2008 U.S. presidential election and an e-politics research agenda. MIS Quarterly 34, 4 (2010), 669–688.

[8] Cheng, M., Wang, S., Yan, X., Yang, T., Wang, W., Huang, Z., Xiao, X., Nazarian, S., & Bogdan, P. (2021). A COVID-19 Rumor Dataset. Frontiers in psychology, 12, 644801. https://doi.org/10.3389/fpsyg.2021.644801.

[9] Cui, L., & Lee, D. (2020). Coaid: Covid-19 healthcare misinformation dataset. arXiv preprint arXiv:2006.00885.

[10] Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R.M., Liu, Z., Merrill, W.C., Mooney, P., Murdick, D.A., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A.D., Wang, K., Wilhelm, C., Xie, B., Raymond, D.A., Weld, D.S., Etzioni, O., & Kohlmeier, S. (2020). CORD-19: The COVID-19 Open Research Dataset. *ArXiv*

[11] MedRxiv. (n.d.). Www.Medrxiv.Org. Retrieved October 17, 2021, from https://www.medrxiv.org/

[12] Pubmed. (n.d.). PubMed. Retrieved October 17, 2021, from https://pubmed.ncbi.nlm.nih.gov/

[13] Mihai Dusmanu, Elena Cabrio, and Serena Villata. Argument mining on twitter: Arguments, facts and sources. In EMNLP, 2017.

[14] Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. Argumentation mining on the web from information seeking perspective. In ArgNLP, 2014.

[15] Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. Argument extraction from news. In ArgMining@HLT-NAACL, 2015.

[16] Dominique Fréard, Alexandre Denis, Françoise Détienne, Michael Baker, Matthieu Quignard, and Flore Barcellini The role of argumentation in online epistemic communities: the anatomy of a conflict in wikipedia. In Proceedings of the 28th Annual European Conference on Cognitive Ergonomics, pages 91–98. ACM, 2010.

[17] Frej, J., Schwab, D., & Chevallet, J. (2020). WIKIR: A Python Toolkit for Building a Large-scale Wikipedia-based English Information Retrieval Dataset. *ArXiv, abs/1912.01901*.

[18] Christian Stab, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. Argumentation mining in persuasive essays and scientific articles from the discourse structure perspective. In ArgNLP, pages 21–25, 2014.

[19] Bahar Sateli and René Witte. Semantic representation of scientific literature: bringing claims, contributions and named entities onto the linked open data cloud. PeerJ Computer Science, 1:e37, 2015.

[20] Shi Yuan and Bei Yu. Hclaime: A tool for identifying health claims in health news headlines. Information Processing & Management, 56(4):1220–1233, 2019.

[21] Maria Liakata, Simone Teufel, Advaith Siddharthan, Colin R Batchelor, et al. Corpora for the conceptualisation and zoning of scientific papers. In LREC. Citeseer, 2010.

[22] Achakulvisut, T., Bhagavatula, C., Acuna, D., & Kording, K. (2019). Claim extraction in biomedical publications using deep discourse model and transfer learning. arXiv preprint arXiv:1907.00962.

[23] Mostafa Dehghani, Aliaksei Severyn, Sascha Rothe, and Jaap Kamps. Learning to learn from weak supervision by full supervision. arXiv preprint arXiv:1711.11383, 2017.

[24] Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. Training complex models with multi-task weak supervision. CoRR, abs/1810.02840, 2018.

[25] Isabelle Augenstein, Sebastian Ruder, and Anders Søgaard. Multi-task learning of pairwise sequence classification tasks over disparate label spaces. arXiv preprint arXiv:1802.09913, 2018.

[26] Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). Fever: a large-scale dataset for fact extraction and verification. arXiv preprint arXiv:1803.05355. Alexander V. Mamishev and Murray Sargent. 2013. Creating Research and Scientific Documents Using Microsoft Word. Microsoft Press, Redmond, WA.

[27] Classification in Medical Abstracts. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.

[28] Pradeep Dasigi, Gully APC Burns, Eduard Hovy, and Anita de Waard. Experiment segmentation in scientific discourse as clause-level structured prediction using recurrent neural networks. arXiv preprint arXiv:1702.05398, 2017

[29] Bei Yu, Yingya Li, and Jun Wang. Detecting causal language use in science findings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4656–4666, 2019

[30] Zhaoning Li, Qi Li, Xiaotian Zou, and Jiangtao Ren. Causality extraction based on self-attentive bilstm-crf with transferred embeddings. arXiv preprint arXiv:1904.07629, 2019.

[31] Arslan, F., Hassan, N., Li, C., & Tremayne, M. (2020, May). A benchmark dataset of check-worthy factual claims. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 14, pp. 821-829).

[32] CLEF2020-checkthat! CLEF2020-CheckThat! (n.d.). Retrieved May 9, 2022, from https://sites.google.com/view/clef2020-checkthat/

[33] Hanselowski, A., et al.: UKP-Athene: multi-sentence textual entailment for claim verification. In: Proceedings of the First Workshop on Fact Extraction

and VERification (FEVER), pp. 103–108. Association for Computational Linguistics, Brussels, November 2018. https://doi.org/10.18653/v1/W18-5516

[34] Tuhin Chakrabarty, Tariq Alhindi, and Smaranda Muresan. 2018. Robust Document Retrieval and Individual Evidence Modeling for Fact Extraction and Verification.. In Proceedings of the First Workshop on Fact Extraction and VERification (FEVER). Association for Computational Linguistics, Brussels, Belgium, 127–131.

[35] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. In Proceedings of Workshop for NLP Open Source Software (NLP-OSS).

[36] Anton Chernyavskiy and Dmitry Ilvovsky. 2019. Extract and Aggregate: A Novel Domain-Independent Approach to Factual Data Verification. In Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)

[37] Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained Fact Verification with Kernel Graph Attention Network. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics

[38] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).

[39] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for Natural Language Inference. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Vancouver, Canada, 1657–1668

[40] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics, Baltimore, Maryland.

[41] Yoneda, T., Mitchell, J., Welbl, J., Stenetorp, P., Riedel, S.: UCL machine reading group: four factor framework for fact finding (HexaF). In: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pp. 97–102. Association for Computational Linguistics, Brussels, November 2018. https://doi.org/10.18653/v1/W18-5515

[42] Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H.: Enhancing and combining sequential and tree LSTM for natural language inference. arXiv preprint arXiv:1609.06038 (2016)

[43] Malon, C.: Team Papelo: transformer networks at FEVER. In: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pp. 109–113. Association for Computational Linguistics, Brussels, November 2018. https://doi.org/10.18653/v1/W18-5517

[44] Soleimani, A., Monz, C., Worring, M. (2020). BERT for Evidence Retrieval and Claim Verification. In: , et al. Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science(), vol 12036. Springer, Cham. https://doi.org/10.1007/978-3-030-45442-5_45

[45] Pandu Nayak. (2019, October 25). Understanding searches better than ever before. https://blog.google/products/search/search-language-understanding-bert/

[46] Qiang Zhang, Aldo Lipani, Shangsong Liang, and Emine Yilmaz. 2019. Reply-Aided Detection of Misinformation via Bayesian Deep Learning. In The World Wide Web Conference (WWW '19). Association for Computing Machinery, New York, NY, USA, 2333–2343. https://doi.org/10.1145/3308558.3313718

[47] Ghai, A., Kumar, P. and Gupta, S. (2021), "A deep-learning-based image forgery detection framework for controlling the spread of misinformation", Information Technology & People, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/ITP-10-2020-0699

[48] Omar Enayet and Samhaa R. El-Beltagy. 2017. NileTMRG at SemEval-2017 Task 8: Determining Rumour and Veracity Support for Rumours on Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 470–474, Vancouver, Canada. Association for Computational Linguistics.

[49] Dadgar, S., & Ghatee, M. (2021). Checkovid: A COVID-19 misinformation detection system on Twitter using network and content mining perspectives. arXiv preprint arXiv:2107.09768.

[50] Elhadad, M. K., Li, K. F., & Gebali, F. (2020). Detecting Misleading Information on COVID-19. IEEE Access, 8, 165201-165215. doi:10.1109/ACCESS.2020.3022867

[51] Al-Rakhami, M. S., & Al-Amri, A. M. (2020). Lies Kill, Facts Save: Detecting COVID-19 Misinformation in Twitter. IEEE Access, 8, 155961-155970. doi:10.1109/ACCESS.2020.3019600

[52] Z. Wang, Z. Yin and Y. A. Argyris, "Detecting Medical Misinformation on Social Media Using Multimodal Deep Learning," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 6, pp. 2193-2203, June 2021, doi: 10.1109/JBHI.2020.3037027.

[53] Varma, R., Verma, Y., Vijayvargiya, P. and Churi, P.P. (2021), "A systematic survey on deep learning and machine learning approaches of fake news

detection in the pre- and post-COVID-19 pandemic", International Journal of Intelligent Computing and Cybernetics, Vol. 14 No. 4, pp. 617-646. https://doi.org/10.1108/IJICC-04-2021-0069

[54] Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact checking. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pages 493–503, Hong Kong, China. Association for Computational Linguistics.

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

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

[57] Joshi, V., Peters, M., & Hopkins, M. (2018). Extending a parser to distant domains using a few dozen partially annotated examples. arXiv preprint arXiv:1805.06556.

[58] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. Transactions of the Association for Computational Linguistics, 8:64–77.

[59] Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-Order Coreference Resolution with Coarse-to-Fine Inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

[60] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., & Gatford, M. (1995). Okapi at TREC-3. Nist Special Publication Sp, 109, 109.

[61] Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. EMNLP.

[62] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler

Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Wang, et al.. 2018. Construction of the Literature Graph in Semantic Scholar. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers), pages 84–91, New Orleans - Louisiana. Association for Computational Linguistics.

[63] Basaldella, M., Furrer, L., Tasso, C. et al. Entity recognition in the biomedical domain using a hybrid approach. J Biomed Semant 8, 51 (2017). https://doi.org/10.1186/s13326-017-0157-6.

[64] Luo, Z., Shi, M., Yang, Z. et al. pyMeSHSim: an integrative python package for biomedical named entity recognition, normalization, and comparison of MeSH terms. BMC Bioinformatics 21, 252 (2020).

[65] Dozat, Timothy & Manning, Christopher. (2016). Deep Biaffine Attention for Neural Dependency Parsing.

[66] Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the Use of Lucene for Information Retrieval Research. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17). Association for Computing Machinery, New York, NY, USA, 1253–1256. DOI:https://doi.org/10.1145/3077136.3080721

[67] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240. Dan Gusfield. 1997. Algorithms on Strings, Trees and Sequences. Cambridge University Press, Cambridge, UK.

[68] Diggelmann, Thomas & Boyd-Graber, Jordan & Bulian, Jannis & Ciaramita, Massimiliano & Leippold, Markus. (2020). CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims.

[69] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder for English. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

[70] CNN. (2020, June 5) "WHO calls on nations to encourage the public to wear fabric face masks where coronavirus is spreading" Cnn.com. Retrieved March 8, 2022, from https://edition.cnn.com/2020/06/05/health/face-mask-coronavirus-who-recommendations-bn/index.html

[71] World Health Organization. (2020). Advice on the use of masks in the context of COVID-19: interim guidance, 6 April 2020. World Health Organization.

https://apps.who.int/iris/handle/10665/331693. License: CC BY-NC-SA 3.0 IGO

[72] Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020, November). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In International Conference on Machine Learning (pp. 11328-11339). Proceedings of the 37th International Conference on Machine Learning, PMLR 119:11328-11339, 2020

[73] Shleifer, S., & Rush, A. M. (2020). Pre-trained summarization distillation. arXiv preprint arXiv:2010.13002.

[74] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 3(1), 1-23.

[75] Williams, Adina; Thrush, Tristan; and Kiela, Douwe (2022) "ANLIzing the Adversarial Natural Language Inference Dataset," Proceedings of the Society for Computation in Linguistics: Vol. 5 , Article 4. DOI: https://doi.org/10.7275/gatd-1283.

[76] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

[77] Maiya, A. S. (2020). ktrain: A low-code library for augmented machine learning. arXiv preprint arXiv:2004.10703.

[78] Chicco, D., Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21, 6 (2020). https://doi.org/10.1186/s12864-019-6413-7

[79] Sadaf Hossein Javaheri, Mohammad Mehdi Sepehri, Babak Teimourpour, Chapter 6 - Response Modeling in Direct Marketing: A Data Mining-Based Approach for Target Selection, Editor(s): Yanchang Zhao, Yonghua Cen, Data Mining Applications with R, Academic Press, 2014, Pages 153-180, ISBN 9780124115118, https://doi.org/10.1016/B978-0-12-411511-8.00006-2.

[80] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: generalized autoregressive pretraining for language understanding. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, Article 517, 5753–5763.

[81] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

[82] Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2019). Adversarial NLI: A new benchmark for natural language understanding. arXiv preprint arXiv:1910.14599.

[83] Nie, Y., Chen, H., & Bansal, M. (2019, July). Combining fact extraction and verification with neural semantic matching networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 6859-6866).

[84] Ölçer, D. and Taşkaya Temizel, T. (2021), "Quality assessment of web-based information on type 2 diabetes", Online Information Review, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/OIR-02-2021-0089

[85] Martinez-Rico, J. R., Martinez-Romo, J., & Araujo, L. (2021). NLP&IR@UNED at CheckThat! 2021: Check-worthiness estimation and fake news detection using transformer models. CEUR Workshop Proceedings, 2936, 545–557. http://ceur-ws.org

[86] Fact Extraction and VERification. (n.d.). Fever.Ai. Retrieved October 17, 2021, from https://fever.ai/2018/task.html

[87] Saakyan, A., Chakrabarty, T., & Muresan, S. (2021). COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic. arXiv preprint arXiv:2106.03794.

TEZ İZİN FORMU / THESIS PERMISSION FORM

ENSTİTÜ / INSTITUTE

3 Fen Bilimleri Enstitüsü / Graduate School of Natural and Applied Sciences	
Sosyal Bilimler Enstitüsü / Graduate School of Social Sciences	
Uygulamalı Matematik Enstitüsü / Graduate School of Applied Mathematics	
Enformatik Enstitüsü / Graduate School of Informatics	X
Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences	

YAZARIN / AUTHOR

Soyadı / Surname : Temiz
Adı / Name : Orkun
Bölümü / Department : Bilişim Sistemleri

TEZİN ADI / TITLE OF THE THESIS (**İngilizce** / English) : Fact Extraction And Verification Pipeline For Covid-19 Related User Posts In Social Media

TEZİN TÜRÜ / DEGREE: Yüksek Lisans / Master [X] **Doktora** / PhD [ ]

1. **Tezin tamamı dünya çapında erişime açılacaktır. /** Release the entire work immediately for access worldwide. [X]
2. **Tez iki yıl süreyle erişime kapalı olacaktır.** / Secure the entire work for patent and/or proprietary purposes for a period of **two year. *** [ ]
3. **Tez altı ay süreyle erişime kapalı olacaktır.** / Secure the entire work for period of **six months**. ***** [ ]

***** *Enstitü Yönetim Kurulu Kararının basılı kopyası tezle birlikte kütüphaneye teslim edilecektir.*
A copy of the Decision of the Institute Administrative Committee will be delivered to the library together with the printed thesis.

Yazarın imzası / Signature ..... **Tarih** / Date 06.06.2022