A FACT EXTRACTION AND VERIFICATION FRAMEWORK FOR SOCIAL MEDIA POSTS ON COVID-19

Orkun Temiz[1], Tuğba Taşkaya Temizel[2]

[1]Graduate School of Informatics, Middle East Technical University, Ankara, Turkey
[2]Graduate School of Informatics, Middle East Technical University, Ankara, Turkey

Abstract

Social media has become popular for spreading and consuming information online. On the other hand, the high number of posts has increased the need for fact checking. In the COVID-19 pandemic, the lack of information on the disease paved the way for the spread of false information, negatively affecting public health and society. In this paper, a new zero-shot fact extraction and verification framework for informal user posts on COVID-19 against medical articles is proposed. The framework includes five main steps, which are pre-processing user posts, claim extraction, document & evidence extraction, and verdict assignment. The framework aims to classify user posts while presenting the related evidence set extracted from peer-reviewed medical articles about each claim in user posts, making it interpretable for end users. The proposed framework obtains on-par and stable performance compared with the state-of-the-art supervised techniques for classifying raw user posts (Coaid) and rumors collected from social media (COVID-19 Rumors Dataset). By utilizing the zero-shot capabilities of the present models in the literature, it achieves superior performance detecting newly emerged misinformation posts and topics.

Keywords

LaTeX Fact Checking and Verification System, Fake News, Misinformation Detection with Credible Information Retrieval, COVID-19, Natural Language Processing

1. Introduction

Social media is widely used for creating and sharing information, ideas, interests, and other forms of narration using the World Wide Web. In the meantime, it has also become a facilitator for the spread of misinformation. In the context of the COVID-19 pandemic, misinformation has disseminated faster than the virus [1]; therefore, it has been called an 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

ROMCIR 2023: The 3rd Workshop on Reducing Online Misinformation through Credible Information Retrieval, held as part of ECIR 2023: the 45th European Conference on Information Retrieval, April 2-6, 2023, Dublin, Ireland
orkun.temiz@metu.edu.tr (O. Temiz); ttemizel@metu.edu.tr (T. T. Temizel)
https://github.com/orkuntemiz/FactExtractionAndVerificationCovid19 (O. Temiz)
0000-0002-1430-1686 (O. Temiz); 0000-0001-7387-8621 (T. T. Temizel)

is labor intensive. 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 provided a specific COVID-19 streaming endpoint [5] and an academic version of its API, which allows developers and researchers to study misinformation and hoaxes in user posts. YouTube removes the videos containing COVID-19 misinformation and limits the recommendation of anti-vaccination content to deal with the problem. However, as detecting misinformation takes time, and information spreads rapidly [6, 7], detecting misinformation early, ideally at the time of submission of a post, becomes important. When misinformation goes 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 and verify given information. In this paper, we 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 paper. We used two different datasets, one of which contains misinformative and informative tweets of COVID-19, retrieved from fact-checking sites and reliable organizations [8], and the other comprises claims and raw tweets collected by using the claims as search terms [9]. We built a fact extraction and verification framework, 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 framework regarding the claim. The links to the documents of the retrieved evidence sources are also presented, so that users can inspect them for further information.

1.1. Research Questions

This paper 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? If so, how can we achieve it?

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, the claim is considered as correct or vice versa. To explore this, user posts including medical claims are analyzed.

RQ2: How can we develop an evidence-based fact-checking method without direct supervision? How does it perform compared to the state-of-the-art supervised models on newly emerging medical claims?

The supervised models' performance is expected to be limited to identifying misinformation, particularly on newly emerging topics as there might not be enough data to train them. On the other hand, we need reliable sources for fact-checking a claim. Peer-reviewed published medical articles are a good example of a reliable source. This study hypothesizes that the evidence-based fact-checking framework with a zero-shot based approach might perform better in classifying newly emerging medical claims/topics than the state-of-the-art supervised models.

RQ3: Can medical document retrieval performance be improved using complementary medical terms for query enhancement? It is hypothesized that expanding a search query with the medical terms and synonyms of those medical words (e.g., common flu vs. influenza) might increase document retrieval performance. Search query is important to find relevant documents from a corpus, which is initially constructed using the words in the check-worthy claim. In this study, the search query is expanded by including these words' related terms using MeSH tree to improve document retrieval performance.

The contributions can be summarized as follows:

- The article presents a framework with zero-shot capabilities of the existing models to fact-check user posts, which include medical claims.
- The proposed system requires no supervision for checking newly emerged medical claims.
- Informal user posts are mapped to scientific medical articles, and evidence of the claims is presented from the medical articles in a comprehensible way. To achieve it, text simplification and transformation are used to map formal evidence to informal user claims to improve the textual entailment performance.
- Medical keywords in the claim are expanded using Medical Subject Headings (MeSH) for improving document retrieval and textual entailment performance. The CORD-19 [10] dataset, MedrXiv [11] and PubMed [12] are used to keep up with the latest medical information for medical fact-checking. In addition, claim extraction, document & evidence retrieval, and verdict assignment steps are provided to ensure a comprehensive fact-checking framework.

1.2. Organization of the Paper

Section 2 reviews the related studies and explains the contributions of the study. Section 3 presents the datasets, the method, and presents the results. Section 4 summarizes the results with further discussions. Section 5 concludes the study, discusses certain limitations, and presents future work.

2. Related Work

This section presents the related studies in three categories: claim extraction, evidence retrieval, and fact-checking.

2.1. Claim Extraction

Previous studies have studied extracting claims from various textual sources, known as argumentation mining or claim extraction. They aimed 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 are limited studies on the biomedical domain. The claim retrieval and extraction task were mainly handled by feature or rule-based methods [19, 20] and simple machine-learning methods [21] in the literature. On the other hand, recent argumentation mining techniques used state-of-the-art deep learning architectures or transfer learning

methods (e.g., [22, 23, 24, 25]). Conference and Labs of the Evaluation Forum (CLEF) organized a series of challenges for argumentation mining [32]. They organized challenges related to the detection and fact-checking of claims in political debates in six different languages. They have also added COVID-19 related dataset very recently.

Since identification of misinformation is a complex and domain specific task, an expert contribution is strongly needed to provide necessary annotations. Therefore, there are limited datasets in the domain [19, 21, 28]. For instance, Thorne et al. [26] built the FEVER dataset, which includes factual claims obtained from Wikipedia and labeled by the public. PubMed 200k RCT [27], a dataset based on PubMed, included sentences annotated according to their role in the abstract, such as background, introduction, method, result or conclusion. It was mainly designed for sequential sentence classification. Maria et al. [21] created the CoreSC dataset comprising 270 hundred articles related to physical and biochemistry topics. Dasigi et al. [28] created a dataset, including 75 articles for extracting statements from PubMed. Stateli et al. [19] introduced a dataset with evaluations for claims and additions using 30 articles from the computer science area. Nevertheless, these datasets are still relatively small to train a deep learning model for argument mining on medical texts.

Different techniques have been studied for claim extraction. Yuan et al. [20] used feature-based (rule-based) claim extraction methods, and Yu et al. [29] worked on 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.

As most of the present datasets are limited to predefined domains such as politics, computer linguistics, and chemistry, feature based ClaimBuster API is used in the proposed framework to extract claims, although it is trained on general election debates in user posts from social media.

2.2. Document Retrieval & Evidence Retrieval

Document retrieval aims to identify and retrieve relevant documents from a corpus based on a query. In the earlier studies of the document retrieval task, mention-based methods have been particularly used. Considering most claims include named entities as a subject and object in the sentences, Hanselowski et al. and Tuhin et al. [33, 34] used named entities (NE) to improve document search. Hanselowski et al. [33] studied a method, which carries out mention retrieval, article search, and filtering tasks. Their study used a simple constituency parser in the mention extraction task to help label the noun phrases in a claim as potential entities and the terms on either side of a verb as potential entity mentions [35]. Their article search component employed a search API to retrieve the mentioned potential entities. Candidate filtering component filtered the entity mentions, which were not part of the claim. The method used in [33] was also conducted by [36, 37]. Tuhin et al. [34] used a similar approach, where they used the Google search API along with a dependency parser to expand the scope of the retrieved documents. Their study also resolved the disambiguation problem.

Another line of study for document retrieval tasks is using keyword-based methods. Nie et al. [83] presented a model including three stages and relies on the Neural Semantic Matching Network (NSNM), which is another type of ESIM [39]. In the document retrieval task, the

authors used a key-value matching approach. 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. Standard practices such as BM25 or cosine similarity indexes in the word vector space are commonly used for retrieving related documents. Therefore, BM25 indexing is also utilized in the document retrieval module of our proposed framework for indexing the documents and retrieving the related top related ones only.

Evidence retrieval, an important step in fact checking, is used to select and retrieve the pieces of evidence from a corpus. The evidence sentences explain why a claim has been selected as reliable or not. For this task, different approaches were adopted in the literature. Thorne et al. [26] utilized a TF-IDF (Term Frequency – Inverse Document Frequency) based method. Yoneda et al. [41] applied a logistic regression model. Enhanced Sequential Inference Model (ESIM) [42] was used for evidence retrieval [33, 43]. ESIM utilizes the co-attention mechanism on top of two BiLSTMs to detect related evidence based on a given sentence. Recent works proposed sentence similarity models for this task. Atanasova et al. [38] adopted a BERT-based model for retrieving related evidence collections from articles. In our paper, a zero-shot-based approach is proposed for this task, which uses the BERT-based question-answering model. It is also shown that the use of BERT based question-answering model as evidence retrieval achieves comparable results [44].

2.3. Fact Checking

Machine learning methods are used to detect misinformation, and they generally use (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. The increasing use of deep learning methods has accelerated progress in text classification, particularly fake news classification. Transformers-based models (BERT, XLNET, ROBERTA) and their ensembled versions are widely used. Zhang et al. [46] used a Bayesian-learning based approach, which outputs a distribution that can model both the prediction and its uncertainty simultaneously. Also, most of the past work on the fact-checking topic has been on text classification. However, users tend to trust the posts/ news more when images supplement the text. Hence, Ghai et al. [47] developed deep-learning-based image forgery detection to address the problem of image manipulation.

Few studies used Twitter-specific features, word and lexical features like follower count, verified account information, 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, the 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 a better performance compared to single machine-learning-based models.

Misinformation on social media posts may appear not only within the text but in the images or the hashtags. Wang et al. [52] proposed an ensemble model that works with multi-modal social media posts (images, texts, and hashtags) to detect anti-vaccine content propagating through social media. Similarly, Prinkaya et. al. [54] used a hierarchical attention network, image captioning, headline matching, noise invariance inconsistency, and error level analysis to create an ensemble for multi-modal fake news detection.

Similarly, due to the complexity of medical and public health issues, 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 the COVID-19 pandemic, as researchers learned more about the virus, they observed that statements that initially seemed true became false in the future, and vice versa. Therefore, Sergio et al. [53] proposed a tool for analyzing official medical bulletins with the help of case-based reasoning. Although, the tool proposed by Sergio et. al. shows partial similarity to the proposed framework, it only classifies a claim according to medical bulletins obtained from the medical sites, but it does not give supportive evidence from the corpus, rather it gives the reference sources related to the query itself.

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 framework 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 framework:

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 entail or contradict 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 credibility according to the evidence collection.

Recent related works have advanced the field by partially addressing several aforementioned framework requirements with numerous models and datasets [1, 6, 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 claims retrieved from fact-checking sources, evidence-based documents, and other unstructured information, but it does not give

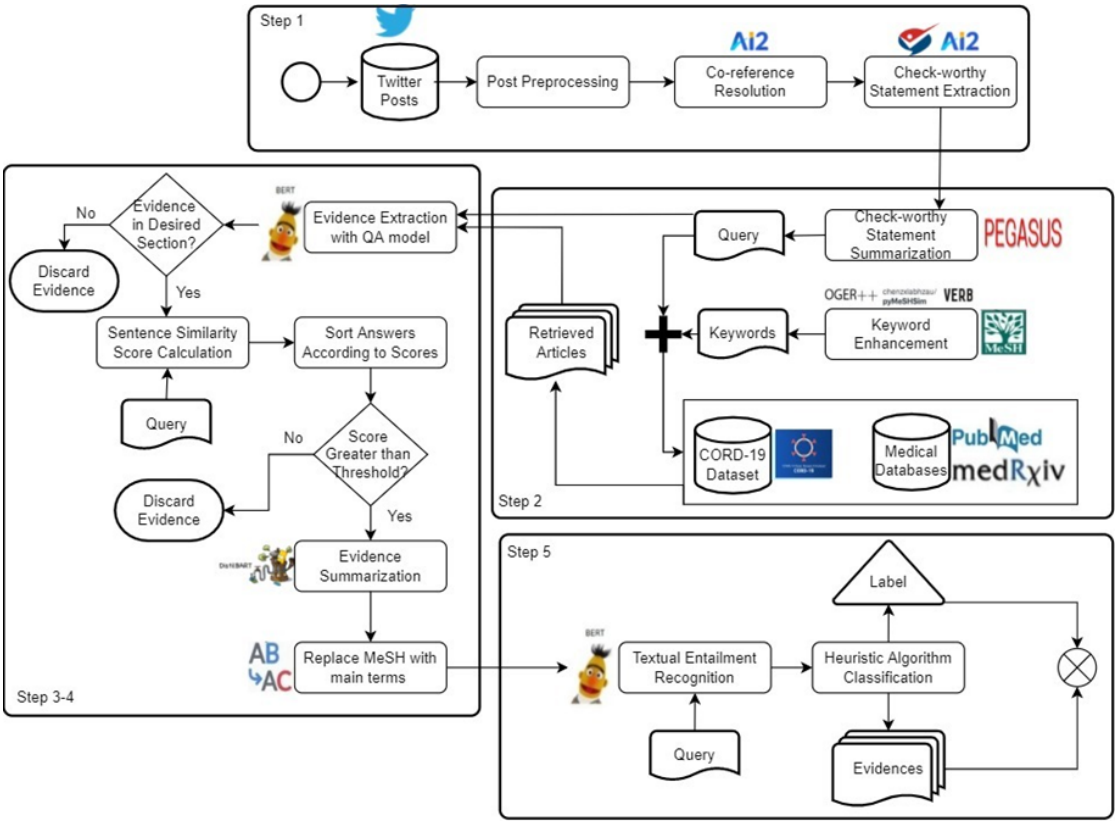

Figure 1: The proposed system framework.

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. On the other hand, our proposed framework aims to deal with all these requirements (See Figure 1).

3. Method

This paper 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 framework with justifications. In addition, we conducted an ablation study to show the impact of each component in Section 4.

3.1. Preprocessing

Firstly, the text is stripped out of 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 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 to select the claimant statement from the text, ClaimBuster API [31] is used with modifications. Given a sentence, ClaimBuster gives it a score between 0.0 and 1.0. The higher score implies that the sentence is more checkworthy. ClaimBuster's score is based on a classification model and scoring scheme. The model was trained using manually labeled past general election debates. To fine-tune the threshold score for our problem, 100 random tweets are selected from the CoAID dataset, then sentences comprising the claim are manually labeled. Then, the best recall threshold is selected, which is 0.35, the default value of which was 0.5. In parallel, each post is parsed with the AllenNLP Constituency Parser [57] to identify sentence and noun phrase structures. If the ClaimBuster API output is neither a noun phrase (i.e., hashtag chunks) nor a question, it is labeled as a claim indicating its eligibility for fact-checking. 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 are 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, consider 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, whereas 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 and Enhancement

In this phase, the search keywords used for the document retrieval model are determined from the claim. 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 a match is found, it is added to the keyword set; otherwise, it is discarded from the keyword set.

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, which uses named entity and concept recognition, which merges a dictionary-based identifier with a corpus-based annotator component for medical subject headings. The pyMeSHSim identifies biological named entities using MetaMap, which links those to medical subject headings to Unified Medical Language System concepts. To successfully identify medical terms, 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 dataset is formed principally by the Allen Institute for AI (AI2), and other contributors such as the White House Office of Science and Technology Policy (OSTP), the National Library of Medicine (NLM). This regularly updated resource includes a large and growing collection of publications and preprints on COVID-19 and previous coronavirus species, such as SARS and MERS. CORD-19 is a resource of over half a million scholarly-reviewed articles, including over 200,000 with full text, about nCov-2019, and related coronaviruses. 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 the CORD-19 dataset), MedRxiv, and Pubmed. The articles published before the post's publication date are considered 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, the retrieved article paragraphs are used to select sentences related to the check-worthy statement as potential evidence. To achieve it, BIOBERT [67] question answering model is employed instead of a sentence similarity model, although the claim is not an interrogative sentence. Although sentence similarity models were frequently adopted in FEVER [26, 68] tasks, QA models may aid to grasp the semantics and the nuance of the query better. 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. Moreover, Google also employs BERT question answering model for its searches [45]. In the proposed framework 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, some words are overlapped with each other to preserve the semantic meaning. Then QA model is applied to all chunks. The one with the highest score is selected as the final answer. After that, to ensure completeness, sentences including the answer and the 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 using a brute-force approach, where all the scores between 0.2-0.8 with 0.1 step size are evaluated and the best performing one is selected as a threshold for the confidence score. The upper bound of the number of returned answers 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 retrieved documents. At the end, the proposed framework returns a query, its answers (related evidences with the query), and the confidence scores between the query and the answers. Since the aim of this framework 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, we need to return 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).

In some cases, the retrieved results are consistent with each other, where all the returned evidence supports the claim stated in the query (check-worthy statement extracted from user posts). But, in some cases, mixed results are encountered, where only some of the articles support the claim in the query while the remaining support the opposite. This phenomenon has been frequently observed during 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 a 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 no need for the members of the public wearing a mask unless they were sick or around people with the coronavirus." Then 8.07.2020 – The World Health Organization changed its stance on wearing facemasks during the COVID-19 pandemic and said, "WHO advises that governments should encourage the public to wear masks where there is widespread transmission in crowded environments and public transportation." [70, 71].

3.6. Textual Entailment

Initially, a preprocessing step is conducted to improve the performance of entailment scoring. Since entailment scores between 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, summarization models are utilized. A claim in a user post may include irrelevant text and numerous noisy characters. Therefore, to emphasize the claim rather than the opinion, text simplification is made on the check-worthy claim with PEGASUS trained on CNN dataset, a state-of-the-art abstractive summarization model trained on C4 and HugeNews [72]. This model ensured that the sentences in the claim were more formal 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, the texts are rewritten to bring both informal query

and formal answers to a similar level, 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). Because standardizing nouns both in claims and evidence may improve the entailment model's performance. The impact of this improvement is experimented with in the ablation study in Section 4.3.2.

Finally, for the entailment model, the BERT model trained on the biomedical 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 a general-purpose corpus, as it can identify medical words to better understand relations between two given texts (see step 5 in Figure 1).

3.7. Heuristic Verdict Assignment

In this phase, the final verdict is determined according to a heuristic. If no related evidence is found, Neutral label is assigned. If the "neutral" score between the query and answer is higher than 0.5, it is added as a Neutral vote. For the remaining cases, the entailment scores between the query and the answers are considered. If the entailment score is higher than the contradiction score, then one vote is added for the Support label or vice versa. In the end, a majority vote is taken 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 tiebreaker, 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 user post including "Shoes can spread COVID-19" (see step 5 in Figure 1). The given threshold scores for both similarity and entailment are determined with a brute force approach by taking independent runs. Values from 0.1 to 0.9 with a 0.1 increment were experimented separately, and the best performing threshold for those was determined as a threshold score.

4. Experiments

4.1. Datasets

The experiments were conducted using two different datasets of user posts. (1) CoAID (COVID-19 heAlthcare mIsinformation Dataset) includes misleading and fake news retrieved from various websites and highly used social platforms (e.g., Twitter), along with users' interactions with this news and posts related to COVID-19. CoAID dataset comprises around 4,250 news, 300,000 related comments, and 900 social media posts. Here, manually labeled ground truth claims were used as search queries to automatically retrieve related tweets and label them. The retrieved social media posts and news include Twitter related attributes. In total, around 10,500 tweets about fake news articles, 140,000 tweets about verified news, 480 posts about false claims, and 8,000 posts about true claims were obtained. In the experiments, the tweets including solely user posts were also used (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".

(2) COVID-19 Rumour Dataset includes manually labeled 4,129 rumors from news and 2,705 rumors from user posts in Twitter with sentiment and stance labels. The true status of the rumors was manually retrieved from fact-checking websites. 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 our 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. This task requires to extracting textual evidence from Wikipedia pages that supports or refutes the claim including one or more entities. 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 system comprising document retrieval, sentence-level evidence selection, and textual entailment steps. Although the tasks are similar, we cannot evaluate the framework using this task's dataset, as the proposed framework 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, we 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 framework does not only label claims, but it retrieves the evidence from the medical articles and then assigns the verdict. 2) Since the proposed framework 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 framework has three distinct labels (Supports, NotEnoughInfo, Refuses).

Consequently, to ensure a fair comparison with the supervised models, we 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, we 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, we have utilized the pre-trained word embeddings BERT Tokenizer and Glove embeddings of dimension 100, respectively. We have also applied the same tweet preprocessing steps for both the proposed framework and the supervised baseline methods. We 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 with the binary cross-entropy loss.

The proposed system generates three distinct labels (Supports, NotEnoughInfo, Refuses), although the dataset includes two labels (True, False). We kept the NotEnoughInfo category because it is needed to show the following outcomes: (1) the framework 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 framework due to the evidence-based classification approach of claims and zero-shot-based learning of the framework. It is also difficult to manually select and label user posts that cannot be checked against scientific articles. In the end, we have mapped two ground-truth class labels as follows: Supports stands for True, Refuses stands for False. 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 framework and dataset labels and for a fair comparison with the supervised models, claims labeled as NotEnoughInfo by the proposed framework are fed to the Baseline 1 model, and metrics are calculated accordingly. Two evaluation schemes were considered.

4.2.1. First Evaluation Scheme

We aim to assess whether the framework 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, we 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.

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 1 show 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 framework's aim, which is to detect newly emerging claims. Therefore, we 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

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*-

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

	Training %	A	F1	P	MCC
Baseline 1	10%	0.65	0.66	0.95	0.16
	20%	0.83	0.84	0.97	0.35
	30%	0.9	0.91	0.97	0.49
	40%	0.93	0.94	0.98	0.58
	50%	0.94	0.95	0.98	0.59
	60%	0.94	0.96	0.97	0.61
	70%	0.94	0.96	0.97	0.61
	80%	0.95	0.97	0.97	0.62
	90%	0.96	0.97	0.98	0.63
Baseline 2	10%	0.63	0.68	0.64	0.26
	20%	0.71	0.74	0.76	0.41
	30%	0.75	0.8	0.78	0.48
	40%	0.76	0.81	0.82	0.49
	50%	0.78	0.84	0.85	0.5
	60%	0.8	0.87	0.86	0.44
	70%	0.83	0.89	0.9	0.5
	80%	0.85	0.91	0.92	0.41
	90%	0.87	0.93	0.98	0.5

1 number of training dataset samples are created, where k represents the number of clusters. To be more specific, the training sample *k1* 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 *k2* 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 framework. Since the proposed framework does not require training, there is no significant difference expected in the framework'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.

The number of clusters, k, determined by the elbow method, was chosen as 91 for the CoAID dataset and 51 for the COVID-19 Rumors dataset respectively. To handle the changing training dataset sizes, the epoch number is kept high and early stopping criteria is used.

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

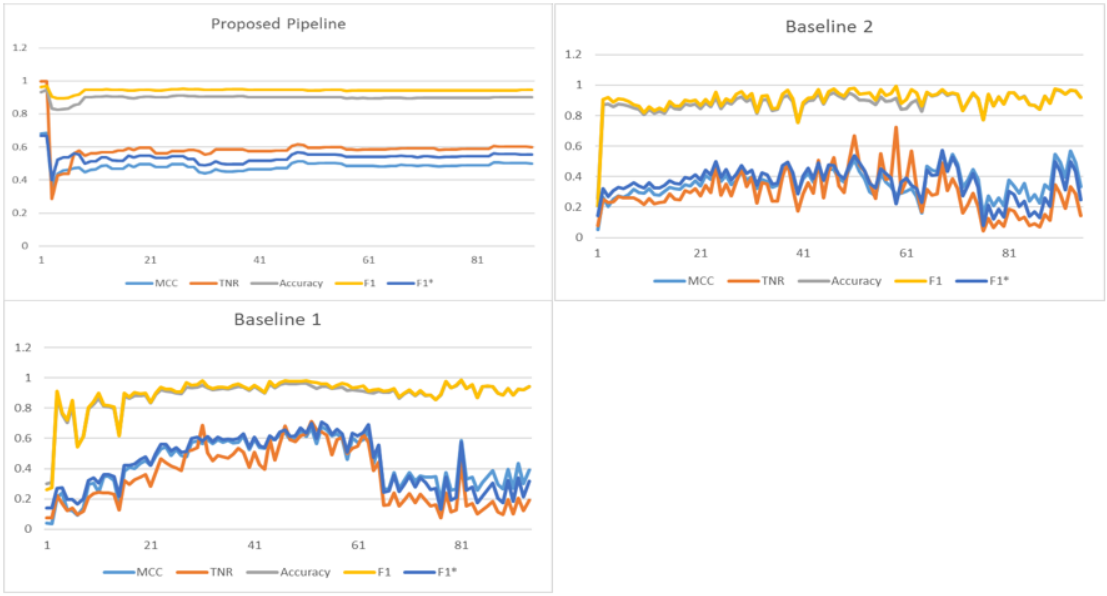

Figure 2: The testing results of the proposed framework and the baseline models. The vertical axis represents the value of metrics, and the horizontal axis represents the cluster count.

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. The proposed framework 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 (see Table 2). More importantly, the framework gives a steady performance across all runs unlike supervised models as can be seen in Figure 4.

The analysis shows that the framework 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 framework since the framework checks the statements against the medical articles. To further analyze this situation, we 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

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.

Dateset	Metric	PP	B1	B2
Baseline 1	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
Baseline 2	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

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 compared to the whole dataset. Table 3 shows the results of this experiment. As expected, the baseline models perform better than the proposed framework 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 framework labeled 64% of the user posts as NotEnoughInfo, compared to 28% in the CoAID dataset also supports our reasoning.

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 framework 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 framework does not include any supervised training, the ablation studies are constructed using the whole dataset. Since the different framework configurations are compared in ablation studies, the results are also reported for the classes Supports/NotEnoughInfo/Refutes.

4.3.1. Study 1

We conducted an ablation study to measure the impact of the Natural Language Inference Model chosen for the proposed framework. 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 by a slight margin. Also, 36% NotEnoughInfo labelled class posts reduced to 28% with M2, as the labelling 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
Accuracy	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

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 framework 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 improve the performance. When the textual entailment score is computed after summarization is applied on both query and answer, the model outperforms, as shown in Table 4. Furthermore, if we do not use summarization, the proposed framework labels 48% of the user

posts as NotEnoughInfo in CoAid Dataset. On the other hand, when we use summarization, the framework's labelling performance increases, and the NotEnoughInfo label percentage drops to 28%. As a result, it is observed that the labelling 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
Accuracy	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

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 framework (Abstract, Introduction, Conclusion, Discussion and Result). As stated, since the framework 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

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 framework 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, we have considered the "Abstract", "Introduction", "Results", "Discussion" and "Conclusion" sections to increase the possibility of the QA model in finding relevant answers to the query. In addition, the framework using these selected paragraphs from the articles labelled 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 labelling performance of the framework.

5. Discussion and Future Work

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

All the codes and tests are carried out 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. 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, we utilized the GPU cores and parallelization; however, this cannot be done for the API calls and database connections. Nevertheless, the framework can be scalable for larger applications, especially if the non-parallelizable operations can be optimized in the framework.

CoAID dataset includes automatically labeled user posts, which were determined based on the labels of the search queries used to retrieve them. 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, we 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, we will implement a semantic classification model in the framework for sarcasm detection and adjust the final verdict about the statement accordingly.

After manually inspecting the results, we 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, we aim to improve the textual entailment model's performance by training or implementing a model dedicated to this task as future work. Also, the tokenization algorithm's performance is important for the model's performance.

In the framework, we have used ClaimBuster API [31] to identify the claims in the text; however, since ClaimBuster was trained on the political claims, we 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 framework. 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 by the LIWC tool, TF-IDF text

features, and a TF-IDF representation of domain names retrieved from a Google search. With this framework, they achieved the first place in the English version of the check-worthiness task. To improve the proposed framework's performance, the models that achieve a high score in CheckThat can be fine-tuned for the proposed framework.

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. We will implement a better ranking/sentence similarity algorithm for answers as a future work. We will also consider different features other than sentence similarity to achieve a more precise ranking.

We 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, we investigated the performances of other framework-based approaches proposed for different domains. For instance, in FEVER 2018 Shared task, the winning framework [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 frameworks might achieve. In this domain, the framework 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, we needed to manually annotate the evidence related to the claims and evaluate the document retrieval and evidence selection performance. For this purpose, we aim to implement a solution similar to FEVER SCORER [86] in the future. Saakyan et al mentioned a FEVER like dataset [87], however, the evidence was collected from Google Search results instead of COVID-19 related medical articles. Therefore, we cannot utilize this dataset either for evaluating document retrieval or evidence selection. Since the datasets we used contained only verdict information and no evidence related to that verdict, we were unable to consider the metrics, including evidence recall and evidence precision, in this study. Therefore, we left this step as future work.

The proposed model has the potential to be applied to different medical domains. The corpus to be searched can be replaced with any corpus using Anserini. There are three main hyperparameters that need to be changed for a different domain: (1) the threshold used for claim extraction, (2) the confidence score used to retrieve the relevant evidence, (3) the threshold used for heuristic verdict assignment. For the first one, the default values can work reasonably well in a general corpus. The latter two thresholds are needed as part of a question-answering system. We note that these parameters remain valid even in the case of posts for newly emerged topics

since they are related to parameters for domain adaptation. In future work, the generalizability of these thresholds across different domains is planned to be investigated

6. Conclusion

In this paper, we proposed a new zero-shot fact extraction and verification framework for user posts related to COVID-19 against the medical articles that have the potential to be applied to other health domains. The framework includes the preprocessing, claim extraction, keyword extraction & query enhancement, document retrieval, evidence selection, textual entailment, and verdict assignment steps. The framework 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.). This study aimed to develop a fact-checking framework that does its fact-checking based on evidence retrieved from 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 formal medical articles in social media? If so, how can we achieve it? "

In the proposed framework, 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. To achieve this, a question-answering model is applied for evidence selection; summarization models are applied for simplifying the retrieved evidence 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 are 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 framework outputs. The results show that the proposed framework gives on-par/better results than the baseline models.

Second, "How can we develop an evidence-based fact-checking method without direct supervision? How does it perform compared to the state-of-the-art supervised models on newly emerging medical claims?"

The proposed framework 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 models 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. The results of the experiments in Section 4.2.2 revealed that the proposed framework achieves better and steady performance in newly emerging claims than the supervised baseline models. However, it should be noted that, if there is no research conducted yet related to a statement in a claim, the proposed framework may not find useful evidence related with the claim. In these situations, the proposed framework returns no related evidence or evidence, which does not include enough info to contradict or entail with the claim.

Third, “Can medical document retrieval performance be improved using complementary medical terms for query enhancement?”

There are numerous 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 were retrieved by querying the MeSH tree. Then they were added to the search query as keywords. Also, using synonyms in the MeSH tree, all the medical terms, which are related to the same term, were replaced with one of the synonyms. In this way, the natural language inference model’s performance was 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 framework 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 framework 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 framework 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/https://doi.org/10.17780/ksujes.435734

[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] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

[7] Wattal, Schuff, Mandviwalla, & Williams. (2010). Web 2.0 and Politics: The 2008 U.S. Presidential Election and an E-Politics Research Agenda. MIS Quarterly, 34(4), 669. https://doi.org/10.2307/25750700

[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] Lu Wang, L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Wilhelm, C., Kohlmeier, S. (2020). CORD-19: The Covid-19 Open Research Dataset. ArXiv, arXiv:2004.10706v2.

[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] Dusmanu, M., Cabrio, E., & Villata, S. (2017). Argument Mining on Twitter: Arguments, Facts and Sources. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2317–2322. https://doi.org/10.18653/v1/D17-1245

[14] Habernal, I., Eckle-Kohler, J., & Gurevych, I. (2014). Argumentation mining on the web from information seeking perspective. CEUR Workshop Proceedings, 1341.

[15] Sardianos, C., Katakis, I. M., Petasis, G., & Karkaletsis, V. (2015). Argument Extraction from News. Proceedings of the 2nd Workshop on Argumentation Mining, 56–66. https://doi.org/10.3115/v1/W15-0508

[16] Fréard, D., Denis, A., Détienne, F., Baker, M., Quignard, M., & Barcellini, F. (2010). The role of argumentation in online epistemic communities: The anatomy of a conflict in Wikipedia. Proceedings of the 28th Annual European Conference on Cognitive Ergonomics - ECCE '10, 91. https://doi.org/10.1145/1962300.1962320

[17] Frej, J., Schwab, D., & Chevallet, J.-P. (2020). WIKIR: A Python Toolkit for Building a Large-scale Wikipedia-based English Information Retrieval Dataset. Proceedings of the 12th Language Resources and Evaluation Conference, 1926–1933. https://aclanthology.org/2020.lrec-1.237

[18] Stab, C., Kirschner, C., Eckle-Kohler, J., & Gurevych, I. (2014). Argumentation mining in persuasive essays and scientific articles from the discourse structure perspective. CEUR Workshop Proceedings, 1341.

[19] Sateli, B., & Witte, R. (2015). Semantic representation of scientific literature: Bringing claims, contributions and named entities onto the Linked Open Data cloud. PeerJ Computer Science, 1, e37. https://doi.org/10.7717/peerj-cs.37

[20] Yuan, S., & Yu, B. (2019). HClaimE: A tool for identifying health claims in health news headlines. Information Processing and Management: An International Journal, 56(4), 1220–1233. https://doi.org/10.1016/j.ipm.2019.03.001

[21] Liakata, M., Teufel, S., Siddharthan, A., & Batchelor, C. (2010, May). Corpora for the Conceptualisation and Zoning of Scientific Papers. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). LREC 2010, Valletta, Malta. http://www.lrec-conf.org/proceedings/lrec2010/pdf/644_Paper.pdf

[22] Achakulvisut, T., Bhagavatula, C., Acuna, D., & Kording, K. (2019). Claim Extraction in Biomedical Publications using Deep Discourse Model and Transfer Learning. https://doi.org/10.48550/ARXIV.1907.00962

[23] Dehghani, M., Severyn, A., Rothe, S., & Kamps, J. (2017). Learning to Learn from Weak Supervision by Full Supervision. NIPS Workshop on Meta-Learning (MetaLearn 2017).

https://arxiv.org/abs/1711.11383

[24] Ratner, A., Hancock, B., Dunnmon, J., Sala, F., Pandey, S., & Ré, C. (2019). Training Complex Models with Multi-Task Weak Supervision. Proceedings of the AAAI Conference on Artificial Intelligence, 33, 4763–4771. https://doi.org/10.1609/aaai.v33i01.33014763

[25] Augenstein, I., Ruder, S., & Søgaard, A. (2018). Multi-Task Learning of Pairwise Sequence Classification Tasks over Disparate Label Spaces. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 1896–1906. https://doi.org/10.18653/v1/N18-1172

[26] Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A Large-scale Dataset for Fact Extraction and VERification. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 809–819. https://doi.org/10.18653/v1/N18-1074

[27] Dernoncourt, Franck & Young Lee, Ji. (2017). PubMed 200k RCT: a Dataset for Sequential Sentence 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] Dasigi, P., Burns, G., & Waard, A. (2017). Experiment Segmentation in Scientific Discourse as Clause-level Structured Prediction using Recurrent Neural Networks.

[29] Yu, B., Li, Y., & Wang, J. (2019). Detecting Causal Language Use in Science Findings. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 4663–4673. https://doi.org/10.18653/v1/D19-1473

[30] Li, Z., Li, Q., Zou, X., & Ren, J. (2019). Causality Extraction based on Self-Attentive BiLSTM-CRF with Transferred Embeddings. https://doi.org/10.48550/ARXIV.1904.07629

[31] Arslan, F., Naeemul Hassan, Li, C., & Tremayne, M. (2020). ClaimBuster: A Benchmark Dataset of Check-worthy Factual Claims [Data set]. Zenodo. https://doi.org/10.5281/ZENODO.3609356

[32] Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum (n.d.). Retrieved May 9, 2022, from https://sites.google.com/view/clef2020-checkthat/

[33] Hanselowski, A., Zhang, H., Li, Z., Sorokin, D., Schiller, B., Schulz, C., & Gurevych, I. (2018). UKP-Athene: Multi-Sentence Textual Entailment for Claim Verification. Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), 103–108. https://doi.org/10.18653/v1/W18-5516

[34] Chakrabarty, T., Alhindi, T., & Muresan, S. (2018). Robust Document Retrieval and Individual Evidence Modeling for Fact Extraction and Verification. Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), 127–131. https://doi.org/10.18653/v1/W18-5521

[35] Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., & Zettlemoyer, L. (2018). AllenNLP: A Deep Semantic Natural Language Processing Platform. Proceedings of Workshop for NLP Open Source Software (NLP-OSS), 1–6. https://doi.org/10.18653/v1/W18-2501

[36] Chernyavskiy, A., & Ilvovsky, D. (2019). Extract and Aggregate: A Novel Domain-

Independent Approach to Factual Data Verification. Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), 69–78. https://doi.org/10.18653/v1/D19-6612

[37] Liu, Z., Xiong, C., Sun, M., & Liu, Z. (2020). Fine-grained Fact Verification with Kernel Graph Attention Network. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 7342–7351. https://doi.org/10.18653/v1/2020.acl-main.655

[38] Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2020). Generating Fact Checking Explanations. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 7352–7364. https://doi.org/10.18653/v1/2020.acl-main.656

[39] Chen, Q., Zhu, X., Ling, Z.-H., Wei, S., Jiang, H., & Inkpen, D. (2017). Enhanced LSTM for Natural Language Inference. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1657–1668. https://doi.org/10.18653/v1/P17-1152

[40] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 55–60. https://doi.org/10.3115/v1/P14-5010

[41] Yoneda, T., Mitchell, J., Welbl, J., Stenetorp, P., & Riedel, S. (2018). UCL Machine Reading Group: Four Factor Framework For Fact Finding (HexaF). Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), 97–102. https://doi.org/10.18653/v1/W18-5515

[42] Chen, Q., Zhu, X., Ling, Z.-H., Wei, S., & Jiang, H. (2016). Enhancing and Combining Sequential and Tree LSTM for Natural Language Inference.

[43] Malon, C. (2018). Team Papelo: Transformer Networks at FEVER. Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), 109–113. https://doi.org/10.18653/v1/W18-5517

[44] Jobanputra, M. (2019). Question Answering for Fact-Checking. Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), 52–56. https://doi.org/10.18653/v1/D19-6609

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

[46] Zhang, Q., Lipani, A., Liang, S., & Yilmaz, E. (2019). Reply-Aided Detection of Misinformation via Bayesian Deep Learning. The World Wide Web Conference on - WWW '19, 2333–2343. https://doi.org/10.1145/3308558.3313718

[47] Ghai, A., Kumar, P., & Gupta, S. (2021). A deep-learning-based image forgery detection framework for controlling the spread of misinformation. Information Technology & People. https://doi.org/10.1108/ITP-10-2020-0699

[48] Enayet, O., & El-Beltagy, S. R. (2017). NileTMRG at SemEval-2017 Task 8: Determining Rumour and Veracity Support for Rumours on Twitter. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 470–474. https://doi.org/10.18653/v1/S17-208

[49] Dadgar, S., & Ghatee, M. (2021). Checkovid: A COVID-19 misinformation detection system on Twitter using network and content mining perspectives. https://doi.org/10.48550/ARXIV.2107.09768

[50] Elhadad, M. K., Li, K. F., & Gebali, F. (2020). Detecting Misleading Information on COVID-19.

IEEE Access, 8, 165201–165215. https://doi.org/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. https://doi.org/10.1109/ACCESS.2020.3019600
[52] Wang, Z., Yin, Z., & Argyris, Y. A. (2020). Detecting Medical Misinformation on Social Media Using Multimodal Deep Learning. https://doi.org/10.48550/ARXIV.2012.13968
[53] Martinez Monterrubio, S. M., Noain-Sánchez, A., Verdú Pérez, E., & González Crespo, R. (2021). Coronavirus fake news detection via MedOSINT check in health care official bulletins with CBR explanation: The way to find the real information source through OSINT, the verifier tool for official journals. Information Sciences, 574, 210–237. https://doi.org/10.1016/j.ins.2021.05.074
[54] Meel, P., & Vishwakarma, D. K. (2021). HAN, image captioning, and forensics ensemble multimodal fake news detection. Information Sciences, 567, 23–41. https://doi.org/10.1016/j.ins.2021.03.037.
[55] Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., & Hajishirzi, H. (2020). Fact or Fiction: Verifying Scientific Claims. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 7534–7550. https://doi.org/10.18653/v1/2020.emnlp-main.609
[56] Augenstein, I., Lioma, C., Wang, D., Chaves Lima, L., Hansen, C., Hansen, C., & Simonsen, J. G. (2019). MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 4684–4696. https://doi.org/10.18653/v1/D19-1475
[57] Joshi, V., Peters, M., & Hopkins, M. (2018). Extending a Parser to Distant Domains Using a Few Dozen Partially Annotated Examples. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1190–1199. https://doi.org/10.18653/v1/P18-1110
[58] Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). SpanBERT: Improving Pre-training by Representing and Predicting Spans. Transactions of the Association for Computational Linguistics, 8, 64–77. https://doi.org/10.1162/tacl_a_00300
[59] Lee, K., He, L., & Zettlemoyer, L. (2018). Higher-Order Coreference Resolution with Coarse-to-Fine Inference. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 687–692. https://doi.org/10.18653/v1/N18-2108
[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. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3613–3618. https://doi.org/10.18653/v1/D19-1371
[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] Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., Kinney, R., Kohlmeier, S., Lo, K., Murray, T., Ooi, H.-H., Peters, M., Power, J., Skjonsberg, S., Wang, oren. (2018). Construction of the Literature Graph in Semantic Scholar. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers), 84–91. https://doi.org/10.18653/v1/N18-3011.

[64] Luo, Z.-H., Shi, M.-W., Yang, Z., Zhang, H.-Y., & Chen, Z.-X. (2020). pyMeSHSim: An integrative python package for biomedical named entity recognition, normalization, and comparison of MeSH terms. BMC Bioinformatics, 21(1), 252. https://doi.org/10.1186/s12859-020-03583-6

[65] Dozat, T., & Manning, C. D. (2016). Deep Biaffine Attention for Neural Dependency Parsing. https://doi.org/10.48550/ARXIV.1611.01734

[66] Yang, P., Fang, H., & Lin, J. (2017). Anserini: Enabling the Use of Lucene for Information Retrieval Research. Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1253–1256. 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] Bulian, J., Boyd-Graber, J., Leippold, M., Ciaramita, M., & Diggelmann, T. (2020). CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims. NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning.

[69] Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., & Kurzweil, R. (2018). Universal Sentence Encoder for English. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 169–174. https://doi.org/10.18653/v1/D18-2029

[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. J. (2020). PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. Proceedings of the 37th International Conference on Machine Learning, 11328–11339.

[73] Shleifer, S., & Rush, A. M. (2020). Pre-trained Summarization Distillation. https://doi.org/10.48550/ARXIV.2010.13002

[74] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Transactions on Computing for Healthcare, 3(1), 2:1-2:23. https://doi.org/10.1145/3458754

[75] Williams, A., Thrush, T., & Kiela, D. (2022). ANLIzing the Adversarial Natural Language Inference Dataset. Proceedings of the Society for Computation in Linguistics 2022, 23–54. https://aclanthology.org/2022.scil-1.3

[76] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. https://doi.org/10.18653/v1/N19-1423.

[77] Maiya, A. S. (2022). ktrain: A Low-Code Library for Augmented Machine Learning. Journal of Machine Learning Research, 23(158), 1–6.

[78] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1), 6. https://doi.org/10.1186/s12864-019-6413-7

[79] Festus Ayetiran, E., & Barnabas Adeyemo, A. (2012). A Data Mining-Based Response Model for Target Selection in Direct Marketing. International Journal of Information Technology and Computer Science, 4(1), 9–18. https://doi.org/10.5815/ijitcs.2012.01.02.

[80] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. In Proceedings of the 33rd International Conference on Neural Information Processing Systems (pp. 5753–5763). Curran Associates Inc.

[81] Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 632–642. https://doi.org/10.18653/v1/D15-1075

[82] Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. https://doi.org/10.48550/ARXIV.1508.05326

[83] Nie, Y., Chen, H., & Bansal, M. (2019). Combining fact extraction and verification with neural semantic matching networks. Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, 6859–6866. https://doi.org/10.1609/aaai.v33i01.33016859

[84] Ölçer, D., & Taşkaya Temizel, T. (2022). Quality assessment of web-based information on type 2 diabetes. Online Information Review, 46(4), 715–732. https://doi.org/10.1108/OIR-02-2021-0089

[85] Martinez-Rico, J., Martinez-Romo, J., & Araujo, L. (2021). NLP&IR@UNED at CheckThat! 2021: Check-worthiness estimation and fake news detection using transformer models.

[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. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International

Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2116–2129. https://doi.org/10.18653/v1/2021.acl-long.165