Cross-genre Retrieval for Information Integrity: A COVID-19 Case Study

Chaoyuan Zuo[1(✉)][0000−0001−5361−0193], Chenlu Wang[2][0009−0007−6870−5401], and Ritwik Banerjee[2,3][0000−0003−0336−0258]

[1] School of Journalism and Communication
Nankai University, China
`zuocy@nankai.edu.cn`

[2] Department of Computer Science

[3] Institute for AI-Driven Discovery and Innovation
Stony Brook University, USA
{`chenlwang,rbanerjee`}`@cs.stonybrook.edu`

Abstract. Ubiquitous communication on social media has led to a rapid increase in the proliferation of unreliable information. Its ill-effects have perhaps been seen most obviously during the COVID-19 pandemic, and have rightfully raised concerns about the integrity of shared information. This work focuses on *derivative* Twitter posts (tweets), *i.e.*, posts that re-transmit preexisting content. We acknowledge that a considerable number of such tweets do not provide a source of information, which undoubtedly undermines the integrity of the information and poses difficulties in fact-checking. To address this concern, we propose an *ad hoc* information retrieval (IR) task to identify the support for claims made in tweets from reputable news outlets. We demonstrate the feasibility of such cross-genre IR by presenting experiments on a COVID-19 dataset of 11K pairs of tweets and news articles. We describe a two-step methodology: (i) selecting the most relevant candidates from 57K pandemic-related news articles, and (ii) a final re-ranking of this selection. Our method achieves significant improvements over the classical token-based approach using BM25 as well as a state-of-the-art transformer-based language model pretrained on COVID-19 tweets. Our findings demonstrate the viability of cross-genre IR across news and social media in safeguarding the integrity of information disseminated through social media.

Keywords: Dataset · COVID-19 · Twitter · Information retrieval

1 Introduction

The COVID-19 outbreak was a global public health emergency with widespread effects. Major disease outbreaks such as Ebola, Zika, and Yellow Fever, have underscored the importance of combating false information [16,34,40,41]. Misinformation related to the COVID-19 pandemic on social media, particularly on Twitter, has been a pressing social issue. It has, for instance, motivated racial

Table 1. Derivative tweets with (✓) and without (✗) a reference to the claim source.

(✓) JPMorgan Chase is investigating potential misuse of pandemic relief programs by its workers and customers, saying some conduct "may even be illegal".
Cited News: JPMorgan investigates employees over potential misuse of PPP loans [www.cnn.com/2020/09/08/business/jpmorgan-covid-relief-misuse]

(✗) I read a WSJ article that said anxiety is also a symptom of Coronavirus....we're all doomed aren't we?

violence [10], caused destruction of infrastructure due to belief in conspiracy theories [4], and discouraged people from seeking medical assistance and receiving vaccines [37]. Numerous studies have examined the occurrence and spread of such misinformation [6,17]. One study found that around 70% tweets contained medical or public health information, and nearly 25% and 17.5% of the tweets were found to contain misinformation and unverifiable information, respectively [26]. In addition to the direct harm caused by misinformation, these findings are particularly concerning because users are more vulnerable to misinformation after prior exposure, leading to a ripple effect where false information spreads faster than accurate and verifiable news [3,53].

Notable studies on the proliferation of false information have distinguished posts containing original content from those that re-transmit preexisting content (*i.e.*, "derivative" posts) [3,5,42]. This is indeed important, since derivative posts are often shared together with reference to external sources such as renowned news outlets, to establish their own credibility [15]. In Table 1, we see two derivative posts propagating claims related to COVID-19, attributing the respective claims to articles from CNN while providing a reference (✓) and the Wall Street Journal without any reference (✗). The verifiability of information by ordinary users and professional fact-checkers alike is greatly impeded by social media posts of the latter kind, where no source is cited. It is unfortunate that such posts, where information is presented without a clear reference to its source, are quite common. Clearly, this demonstrates the need for an intelligent and automated system for the retrieval of news articles (if any such article exists) that support the information in a tweet.

The unique characteristics of social media posts, including their brevity, informal language, and non-standard grammar and spelling, present significant challenges for natural language processing (NLP) tasks. As a result, Twitter and other short-form social media content have been treated as a linguistic genre distinct from traditional newswire [19,45]. To effectively retrieve information from news articles for social media posts, a retrieval system must be robust and adaptable to the specific nuances of both genres. This necessitates a thorough comprehension of the characteristics that set social media posts apart

[4] UK phone masts attacked amid 5G-coronavirus conspiracy theory. https://www.theguardian.com/uk-news/2020/apr/04/uk-phone-masts-attacked-amid-5g-coronavirus-conspiracy-theory

from conventional news articles and the capacity to modify retrieval techniques accordingly. Developing an approach that can bridge this gap is thus crucial to ensure the integrity of information presented in contemporary social media for consumption by the global populace. To that end, we commence by reviewing relevant literature (Sec. 2) before describing our contributions:

(a) a dataset of $11,444$ tweets and $57K$ news articles related to the COVID-19 pandemic (Sec. 3);
(b) an *ad hoc* cross-genre IR task to retrieve news articles from reputable sources that corroborate the information presented in a tweet related to COVID-19. To this end, we present a two-stage pipeline comprising (i) candidate selection from the large set of news articles, and (ii) re-ranking using a transformer-based cross-encoder (Sec. 4); and
(c) an in-depth analysis of our findings (Sec. 5), demonstrating the viability of cross-genre information retrieval for information integrity.

2 Related Work

Fact-checking and identifying misinformation has a long and rich history in journalism [18]. With the rise of user-generated content and social media platforms, however, the last two decades have seen a massive surge in automated computational (or at least computer-aided and partially automated) means of fact-checking and the detection of various forms of misinformation (see, for example, the multilingual architecture proposed by Martin et al. [33] or the recent survey by Guo et al. [20]). Two threads of research are particularly relevant to our work: (i) identifying misinformation on social networks, and (ii) information retrieval across different genres. The contributions we present in subsequent sections grow both these directions of research, discussed hereunder.

Misinformation detection on social media platforms. News consumption habits have undergone a drastic change since the advent of social media, especially Twitter. There are manifold reasons for it, leading to a large fraction of the populace consuming news and information on social media. For example, nearly half of all Americans use social media as a news source at least sometimes [5]. A manual verification of information being impossible due to the deluge of data and rate at which information spreads, computational means have been increasingly explored. Some techniques involve the retrieval of posts associated with misinformation identified *a priori* by the researchers [24,49]. These are, by design, helpful in the study of the propagation of misinformation in a network, but not in their initial identification. To identify falsehoods, methods can broadly be categorized into three stages: claim detection, evidence discovery, and claim verification. The first of these often relies on a claim being worth checking and it being checkable [23,25]. The discovery of evidence is a more recent

[5] News Consumption Across Social Media in 2021. Pew Research Center, www.pewresearch.org/journalism/2021/09/20/news-consumption-across-social-media-in-2021/

research thrust, in contrast to earlier work on misinformation, which did not use external evidence beyond the immediate text [52,14]. Recently, Dougrez-Lewis et al. [13] introduced a dataset of tweets along with external evidence retrieved from the Web, and showed how the inclusion of external evidence can benefit rumor identification models. Their work, however, was restricted to five pre-identified rumors. On the other hand, Haouari [21] proposed a method to verify tweets immediately upon posting by evidence retrieval from multiple sources, but these sources were limited to other tweets. For a more in-depth survey of misinformation detection in social media, we point the reader to Shu et al. [47].

With regard to COVID-19, misinformation had a particularly deleterious effect in many ways. The extent of misinformation has been studied by many. Kouzy et al. [26], for example, found in a small sample of 673 tweets, that 25% contained misinformation, and 17.4% propagated unverifiable information. We thus use COVID-19 as a timely case study that distinguishes itself from this literature by not only looking for external evidence in a specialized domain, but also doing so across two different genres.

Information Retrieval Methods and Datasets. The majority of modern *ad hoc* IR systems are based on bag-of-words representations with term-weighting approaches such as the BM25 algorithm or its variants [31,44]. BM25 is a ranking function commonly used for text document retrieval tasks, which determines the relevance of a document to a query based on the frequency of query terms in the document and other factors. These methods have traditionally been employed for query-based news retrievals that focused on specific parts of an article [8]. With the success of BERT [12] and its successors [27,29,46] in NLP, however, more recent IR research on evidence retrieval has incorporated these language models [48]. Of particular relevance to this work is that some applications of transformer-based encoders show the value of contextual information in re-ranking candidate documents retrieved by models based on BM25 [39]. With this body of work as our motivation, we design a two-stage pipeline for information retrieval across news and social media wherein classical term-weighting as well as transformer-based cross-encoders are used.

Most IR datasets of sufficient size are able to provide a reasonably accurate representation of a specific domain or topic, making them effective benchmarks for the evaluation and comparison of various IR models. Notable examples include the various TREC collections[6], the NTCIR task datasets[7], and the CLEF initiative task datasets[8]. These are largely confined to single genres and document types. Despite several cross-lingual IR (CLIR) and a few other cross-genre IR datasets (*e.g.*, [50,56]), there is a comparative dearth of this latter category, even more so for retrieval tasks related to COVID-19. In this work, we present a dataset that comprises two distinct linguistic genres (tweets and news articles) as a much needed contribution not just to help mitigate the effects of misin-

[6] trec.nist.gov/data.html

[7] research.nii.ac.jp/ntcir/data/data-en.html

[8] www.clef-initiative.eu

Table 2. News-related keywords: the 22 keywords used to filter Tweets.

bloomberg, cbs, cnbc, cnn, forbes, nypost, reuters, sfgate, theatlantic, wsj, abc news, chicago tribune, fox news , new york post, new york times, ny times, the atlantic, the daily beast , the guardian, us news, usa today, washington post

formation during health crises, but also as a potential benchmark for future IR tasks that query across such distinct linguistic genres.

3 Dataset

Our work starts by analyzing a subset of a large open dataset of COVID-19 related tweets developed and made publicly available by Banda *et al.* [7]. Specifically, our work focuses on 46.86 million tweets collected from March to May. Although this dataset pertains to COVID-19, it required significant pre-processing steps to ensure its suitability for our tasks. This includes the application of rigorous filtering and data cleaning procedures, such as the removal of retweets and non-English posts. Furthermore, as our work is focused on derivative posts that transmit information from reputable news agencies, we implement a more stringent criterion to establish the relevance of each post to the corresponding news article. Specifically, we employ a set of 22 carefully chosen keywords (shown in Table 2) and retain only those tweets containing at least one.

After applying the aforementioned filtering steps, we are left with a total of 876,325 Tweets. Among these derivative posts, we observe that a significant proportion of 23% (197,464) tweets have no hyperlink to any information source. This indicates a potential loss of information integrity and underscores the need to establish an information retrieval system to identify the sources of information for these Tweets. To build such an IR system, we then create a IR dataset for training purposes.

3.1 IR Dataset Creation

Our IR Dataset is based on a large corpus of check-worthy tweets about COVID-19, developed by Zuo *et al.* [57]. It offers about 30K tweets collected from March through May 2020, where each tweet contains a factual claim and is deemed worth checking for factual accuracy[9]. Moreover, each tweet refers to the external source of its information by providing a hyperlink to a news article from a reputable news publisher. This relation between a tweet and a news article provides the ground truth for strong relevance labels for our task of retrieving the article that supports the claim made in a tweet.

We notice that many tweets are mere retweets of a news article, and the linguistic content of the tweet already contains the article headline. The inclusion

[9] The work by Zuo *et al.*[57] follows the notion of *check-worthiness* set forth by a large body of work in NLP research (Arslan *et al.*[4] and Hassan *et al.*[22], among others).

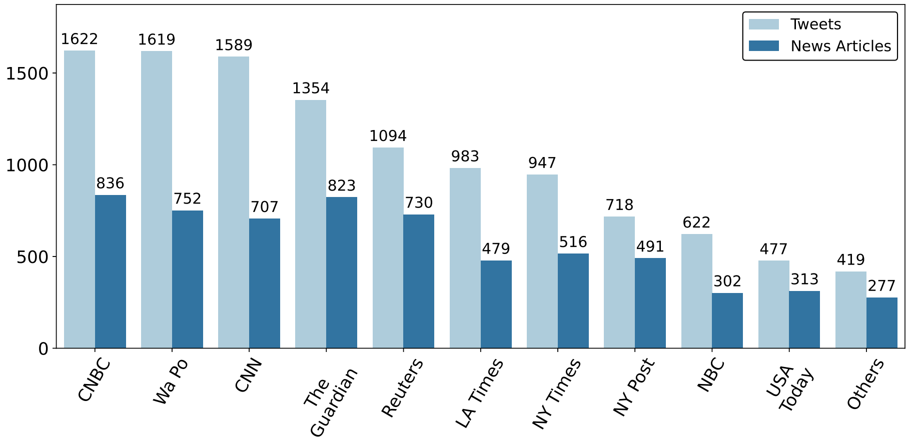

Fig. 1. Distribution of news articles and tweets citing them, across the ten most frequent news agencies (plus "Others", which comprises the combination of all news enterprises that have fewer than 400 tweets).

of such instances will make the IR task of this work overly and impractically simple, since any standard search will immediately find the exact match. Thus, we discard such tweets, retaining a corpus of 11,444 tweets and 6,226 news articles. The number of unique news articles is understandably smaller, since multiple tweets often cite the same article from a well-known news publication. The distribution of tweets and the linked news articles, over the news agencies, is shown in Fig. 1. Our final dataset[10] consists of tuples of the form $(t, \{h, b\})$, where t is the tweet with its hyperlinks removed, and $\{h, b\}$ is the headline-body pair from the news article cited by that tweet. Moreover, we consider a realistic scenario for a lay human reader who wants to verify the information in a tweet. This reader will usually need to search for a news report from a vast collection. To imitate this scenario, we add a collection of 51,003 news articles from the RSS feeds of several popular news websites over a period of three years (2020-2022). We collect these news articles based on a set of six keywords related to COVID-19: *corona, coronavirus, covid, covid-19, pandemic, quarantine*. In our entire dataset, the full text and the headline of all the articles are retained, while images and videos are discarded.

4 Experiments

Given a tweet propagating a claim pertaining to COVID-19, we present the cross-genre IR task of retrieving news articles that support it. Along the lines of other *ad hoc* pipelines [11,32], this comprises two stages: a retrieval system to obtain a

[10] https://github.com/chzuo/adma2023_tweet_IR

Table 3. Results of candidate selection for different models and K values when only the title of news was searched and when both title and body were searched. The best results for each K value are in bold. MiniLM* indicates that the model is pretrained on several corpora, while models with † are fine-tuned on the MS MARCO dataset.

	Title				**Title + Body**			
Model	**R@1**	**R@10**	**R@100**	**R@500**	**R@1**	**R@10**	**R@100**	**R@500**
BM25	63.5	75.0	81.5	86.3	58.3	**79.5**	**89.3**	**93.2**
DistilBERT†	37.5	60.8	76.0	84.1	58.0	69.6	77.7	83.0
MiniLM†	29.0	52.7	71.0	81.1	56.4	68.3	76.4	81.9
MiniLM*	40.4	67.2	82.6	89.3	**59.7**	73.0	81.9	87.3
CT-BERT	**64.1**	**79.9**	**88.5**	**92.2**	1.1	2.6	5.7	12.1

large candidate list, and the re-ranking of this list by a transformer-based cross-encoder. Furthermore, we notice that tweets may often simply modify the news headline. Thus, using only the news headline for retrieval may boost the results as well as improve efficiency. For this reason, we conduct two sets of experiments: by considering (i) only the headline of each news article, and (ii) the headline as well as the body of the article.

4.1 Candidate Selection

In this first stage, we aim to reduce the search space for the final re-ranking task. For this, we consider the classical lexical IR approach of token-based bag-of-words models, *i.e.*, BM25 algorithm, which has remained an extremely viable model for information retrieval [30] for nearly two decades. Additionally, we also use transformed-based Bi-Encoders as semantic search. It encodes the query (*i.e.*, the tweet post) into vector space and retrieves the news embeddings that are nearby in the same space.

For the BM25 approach, proper preprocessing steps are added before retrieval, which comprises converting the words into lowercase, removing function words, and stemming. We also include text cleaning procedures designed for tweet posts, including removing Twitter user handles, emojis, and hash symbols for hashtags (the term is retained, *e.g.*, "#quarantine" to "quarantine").

As part of the Bi-Encoder approach, we use two pretrained models from Sentence-BERT [43]: MiniLM [54] and DistilBERT [46]. Those models encode the tweet t, the news headline h, and the full news (headline and body) h+b respectively. We then obtain the ranked list of news pertinent to the tweet based on cosine similarity. Given that a claim may bear some semantic similarity to the evidence supporting it [35,2], the pretrained models are fine-tuned on several corpus, including the Natural Language Inference (NLI), the Semantic Textual Similarity (STS) benchmark datasets [9] and *etc.*. Considering the newswire article is significantly longer than the given query, we also use the pretrained models tuned on the large-scale information retrieval corpus, *i.e.*, MS MARCO Passage

Ranking Dataset[38]. Moreover, since our task involves COVID-19 tweets, we use the COVID-Twitter-BERT(CT-BERT) [36], which is pretrained on 97M tweets.

For evaluation, since each tweet in our dataset cites only one newswire article, precision is not an important measure for this task. Instead, we measure recall@k ($k = 1, 10, 100, 500$) to report whether the hyperlinked news is in the top-K selection. As it has been argued in other two-stage IR pipelines [48], a high recall is crucial in this case because the correct news report will otherwise be excluded from the final re-ranking. Recall@K would also be useful in a scenario where a reader needs to verify tweets in real-time. Our retrieval system could return a shortlist of up to ten news items that are pertinent to the tweets, and the reader could rapidly scan them for verification.

Table 4.1 shows the results of experiments across the BM25 and Bi-Encoders, which compare the fraction of times the correct document was found in the first K ranked documents. When K is equal to 1, BM25 could locate the correct document more than 63% of the time using only the title. By increasing the value of K to 500, this reaches 86.3%. The search results get worse for smaller K values when the body of the article is also included. However, including the body gives us better results when we have a more considerable K value. One explanation could be some tweets refer to a specific part of the news that is not included in the title, but the algorithm cannot rank them on top of the list; when the K value increases, those documents are retrieved. When the title is used, CT-BERT outperforms all other algorithms in terms of results for all K values, which shows that domain-specific knowledge is crucial in this task. When adding the news body, however, the model's performance drops significantly, as the news body may be too extensive for it to embed the core information. All models, with the exception of CT-BERT, achieve similar performance when the news body is included for lower values of K, while the MiniLM model pretrained on a large corpus performs marginally better. However, the BM25 model outperforms others for $K > 1$, demonstrating this token-based approach is still a hard-to-beat baseline for asymmetric semantic search with long documents and short queries.

4.2 Re-Ranking

We keep $6,000$ tweet-news pairs for training, $2,444$ for development, and $3,000$ for testing. We first use the BM25 approach (on the full news) with highest score in Recall@500 to generate a list of 500 newswire articles for each tweet. It is possible that the correct news was not retrieved during candidate selection. In that case, we add it back to the list. To train the Cross-Encoder, we pair the tweet t with the newswire article with headline h and body b. These concatenated strings serve as training data for our task. The ground-truth label is 1 for an input $t + h$ (or $t + h + b$, for experiments on the full news), where the article is indeed cited by the tweet. For other inputs, which are created by random sampling from the BM25 candidate list, the label is 0. As discussed by Thakur *et al.* [51] and Zuo *et al.*[56], using the right sampling strategy to create negative samples is crucial to achieving performance improvement. We randomly sample 4 news articles from the top 50-100 candidate list retrieved by the BM25 approach.

Table 4. Re-ranking results. The best results for each K value are in bold. BM25 and CT-BERT without training serve as the baseline. MiniLM* indicates that the model is pretrained on several corpus, while models with † are pretrained as Cross-Encoder on the MS MARCO dataset.

	Title				**Title + Body**			
Model	**R@1**	**R@10**	**R@100**	**R@MRR**	**R@1**	**R@10**	**R@100**	**R@MRR**
Baseline								
BM25	63.5	75.7	81.3	17.8	57.8	79.7	89.4	17.2
CT-BERT	64.4	79.7	88.8	18.4	1.3	2.5	5.6	0.5
MiniLM*	68.3	85.3	97.5	74.4	67.7	88.9	99.6	75.4
MiniLM†	70.1	88.3	98.4	76.5	76.9	92.1	99.4	82.5
TinyBERT†	67.9	83.7	97.6	73.5	73.6	88.7	98.1	79.0
CT-BERT	**73.2**	**93.5**	**99.5**	**80.5**	**77.6**	**95.0**	**99.8**	**83.9**

Although random selection typically results in dissimilar pairs, these articles from the candidate list may share similarities with the tweet someway, making them strong negative samples to prevent the classification process from being overly simple. We use this labeled data to tune pretrained transformer-based models. During prediction, we use the softmax probabilities of the classification scores to re-rank the news article for each tweet and calculate recall@k for $k = 1, 3, 5, 20$, as well as the mean reciprocal rank (MRR).

As part of our experiments, we train different models – MiniLM from sentence-BERT, MiniLM (tuned as Cross-Encoder on the MS MARCO dataset), TinyBERT (tuned as Cross-Encoder on MS MARCO dataset), and CT-BERT. All models are trained for 2 and 3 epoch, batch sizes of 16 and 24, and maximum sequence lengths of 256 and 512 tokens. The final hyperparameters are manually chosen based on MRR achieved on the development set.

The results of re-ranking are shown in Table 4. As a benchmark for candidate selection, BM25 is difficult to surpass, but token-based lexical search makes mistakes when words from the newswire article do not match those in the tweet, which frequently happens when the same or similar meanings are expressed across two different genres. A Cross-Encoder based re-ranker can perform attention across the query and the document, improving the final results with higher performance. Following training, CT-BERT performs better than competitors, and it may include the right news story in the Top 10 list more frequently than 93% of the time when only the title is used. Even if a slight improvement could result from including the news body, relying just on the news title would be a more efficient way for this IR task. Also, for MiniLM, fine-tuning on the MS MARCO dataset significantly improves, indicating the importance of task-specific training.

Table 5. The tweet disseminates information pertaining to three newswire articles. In the retrieval results generated by CT-BERT, the cited news ranks third.

Tweet: JC Penney files for bankruptcy during coronavirus pandemic #retailers #retail #retailbankruptcy
Cited News: Long-struggling JC Penney files for bankruptcy as coronavirus crushes hopes for a quick turnaround (Source: CNBC)

Retrieval results
1. *JCPenney files for bankruptcy as the coronavirus hammers retail (Source: NBC)*
2. *JC Penney could join a growing list of bankruptcies during the coronavirus pandemic (Source: CNBC)*

5 Discussion

5.1 News Event Clustering

Our dataset contains a set of 57k covid-19-related news reports, and a subset of 6k newswire articles ($\{n_i\}$) are linked by the tweets. It is also worth pointing out that our evaluation is based on relevance labels from these hyperlinks. However, as noted in the by Liu *et al.* [28] regarding the redundancy of news ("*a news event or story is likely to be reported and discussed by multiple publishers*"), we may encounter the case where several newswire articles regarding the same event are retrieved when we perform information retrieval for the given tweet. It is possible that some documents that are given higher rankings actually support the tweet, but are judged as irrelevant because they are not cited (as shown in Table 4). This will impact our evaluation results of the models. In fact, many IR benchmark datasets – *e.g.*, MS MARCO [38] – do not offer strong non-relevance labels. Instead, one could consider the results as a lower bound in this general evaluation setup (*i.e.*, With exhaustive ground-truth labels of non-relevance, the true performance are better, not worse.)

Additionally, we run same-event news searches to take a second look at the findings. We characterize a group of documents as referring to the same event when they have a simultaneous publication date, have semantic similarity, and share comparable named entities. These criteria are similar to the definition of an event in other previous work [1,55]. We perform a three-step pipeline to obtain such sets. First, for each news n_i linked by tweets, we identify a set of candidate news articles ($\{N_i\}$) from the overall news collection that was published five days before and after it. In the second step, we employ sentence-BERT to measure the semantic similarity between the news n_i and the candidates N_i among the headlines, and only keep the articles N_i above a given threshold. Then, we use the named entity recognition(NER) algorithm to obtain the named entities from the full news (title and body) and compare the intersection between the retained news candidates N_i and news n_i. Only news articles that exceed the predetermined threshold are kept. For articles N_i retained after this step, it is related to the same event with the given news n_i. Two independent readers who each received a random sample of 50 such sets noticed this. They both agreed that the final candidate lists for each set present the same event with the provided news. Finally, for 1,461 news articles linked by tweet (23.5%), we obtained a set of reports on the same news story separately. To adjust the evaluation results, for each tweet, if the linked news n_i has a same-event news set $\{N_i\}$, we additionally assign the relevant label to the news N_i. We find that the MRR for CT-BERT trained on title and body increased slightly, from 83.9 to 85.0, while Precision@1(equal to Recall@1 without same-event news searching) improves from 77.6 to 78.8.

5.2 Error analysis

We conduct a thorough analysis of instances where the CT-BERT model fails to accurately identify the news in the Top 5 list. Our findings reveal that 10% of the errors were due to inaccurate or dirty data, as these tweets do not contain any factual information. Additionally, over 40% of the mistakes are Type I errors, where the tweets are only relevant to the latter half of a lengthy news document, making it challenging for the Cross-Encoder model to fully encode the information due to its token length restriction. The remaining 50% of errors are Type II errors: the tweet contains an additional statement that is not in the news, even though the content is relevant.

This has prompted us to contemplate the utilization of the model. Firstly, for posts on social media platforms that lack reference sources, we can leverage the model to identify relevant authoritative news articles to aid in verifying the authenticity of the original post. Secondly, with regard to tweets that have listed information sources, we can also conduct IR research to uncover additional news reports that pertain to the news event, thus mitigating the potential for any bias inherent in a singular news report. Additionally, by evaluating the similarity score generated by the model between the tweet and the news, we can make an assessment of the extent to which the original tweet is supported by the cited news. If the similarity score is comparatively low, this indicates that the tweet includes information not present in the news, and a more rigorous fact-checking process for the tweet would be necessary.

6 Conclusion

In this work, we provide a novel dataset by linking tweets related to COVID-19 to reliable news articles. We use this linked dataset to then present a pipeline for the retrieval of news related to the tweeted claim(s). The findings of our investigation show that cross-genre information retrieval is practical for confirming the veracity of information about the epidemic, based on the support such information finds in journalistic organizations of repute. Furthermore, our findings emphasize the importance of incorporating domain-specific knowledge in the information retrieval process. This highlights the need for responsible and informed social media usage, particularly in times of crisis, where access to accurate information is of paramount importance. By utilizing reliable sources of information, we can help ensure the integrity of such information across genres, as we demonstrated through our experiments across traditional and social media.

Acknowledgment. This work was supported by the Fundamental Research Funds for the Central Universities (No. 63232115).

References

1. Allan, J., Carbonell, J.G., Doddington, G., Yamron, J., Yang, Y.: Topic Detection and Tracking Pilot Study Final Report. Tech. rep., Carnegie Mellon University (1998). https://doi.org/10.1184/R1/6626252.v1
2. Alonso-Reina, A., Sepúlveda-Torres, R., Saquete, E., Palomar, M.: Team GPLSI. approach for automated fact checking. In: Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER). pp. 110–114 (2019). https://doi.org/10.18653/v1/D19-6617
3. Arif, A., Shanahan, K., Chou, F., Dosouto, Y., Starbird, K., Spiro, E.S.: How information snowballs: Exploring the role of exposure in online rumor propagation. In: Proceedings of the 19th Conference on Computer-Supported Cooperative Work & Social Computing, 2016. pp. 465–476 (2016). https://doi.org/10.1145/2818048.2819964
4. Arslan, F., Hassan, N., Li, C., Tremayne, M.: A Benchmark Dataset of Check-Worthy Factual Claims. Proceedings of the International AAAI Conference on Web and Social Media **14**(1), 821–829 (2020). https://doi.org/10.1609/icwsm.v14i1.7346
5. Badawy, A., Ferrara, E., Lerman, K.: Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. In: ASONAM. pp. 258–265 (2018). https://doi.org/10.1109/asonam.2018.8508646
6. Balakrishnan, V., Ng, W.Z., Soo, M.C., Han, G.J., Lee, C.J.: Infodemic and fake news – A comprehensive overview of its global magnitude during the COVID-19 pandemic in 2021: A scoping review. International Journal of Disaster Risk Reduction **78**, 103144 (2022)
7. Banda, J.M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Artemova, E., Tutubalina, E., Chowell, G.: A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration. Epidemiologia **2**(3), 315–324 (2021). https://doi.org/10.3390/epidemiologia2030024
8. Catena, M., Frieder, O., Muntean, C.I., Nardini, F.M., Perego, R., Tonellotto, N.: Enhanced News Retrieval: Passages Lead the Way! In: SIGIR. p. 1269–1272. SIGIR'19 (2019). https://doi.org/10.1145/3331184.3331373
9. Cer, D.M., Diab, M.T., Agirre, E., Lopez-Gazpio, I., Specia, L.: Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In: SemEval. pp. 1–14 (2017)
10. Chong, M., Froehlich, T.J., Shu, K.: Racial Attacks during the COVID-19 Pandemic: Politicizing an Epidemic Crisis on Longstanding Racism and Misinformation, Disinformation, and Misconception. Proceedings of the Association for Information Science and Technology **58**(1), 573–576 (2021). https://doi.org/10.1002/pra2.501

11. Dai, Z., Callan, J.: Deeper Text Understanding for IR with Contextual Neural Language Modeling. In: SIGIR. pp. 985–988 (2019). https://doi.org/10.1145/3331184.3331303
12. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT. pp. 4171–4186 (2019). https://doi.org/10.18653/v1/n19-1423
13. Dougrez-Lewis, J., Kochkina, E., Arana-Catania, M., Liakata, M., He, Y.: PHEMEPlus: Enriching social media rumour verification with external evidence. In: Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER) (2022)
14. Dungs, S., Aker, A., Fuhr, N., Bontcheva, K.: Can Rumour Stance Alone Predict Veracity? In: COLING. pp. 3360–3370 (2018)
15. Fogg, B.J., Cuellar, G., Danielson, D.: Motivating, influencing, and persuading users: An introduction to captology. In: The Human Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, pp. 159–172. CRC Press (2007). https://doi.org/10.1201/9781410615862
16. Fung, I.C.H., Fu, K.W., Chan, C.H., Chan, B.S.B., Cheung, C.N., Abraham, T., Tse, Z.T.H.: Social Media's Initial Reaction to Information and Misinformation on Ebola, August 2014: Facts and Rumors. Public Health Reports **131**(3), 461–473 (2016). https://doi.org/10.1177/003335491613100312
17. Gabarron, E., Oyeyemi, S.O., Wynn, R.: Covid-19-related misinformation on social media: a systematic review. Bulletin of the World Health Organization **99**, 455 – 463A (2021)
18. Graves, L.: Deciding What's True: The Rise of Political Fact-Checking in American Journalism. Columbia University Press (2016). https://doi.org/10.7312/grav17506
19. Gui, T., Zhang, Q., Gong, J., Peng, M., Liang, D., Ding, K., Huang, X.: Transferring from formal newswire domain with hypernet for twitter POS tagging. In: EMNLP. pp. 2540–2549 (2018). https://doi.org/10.18653/v1/d18-1275
20. Guo, Z., Schlichtkrull, M., Vlachos, A.: A Survey on Automated Fact-Checking. Transactions of the Association for Computational Linguistics **10**, 178–206 (02 2022). https://doi.org/10.1162/tacl_a_00454
21. Haouari, F.: Evidence-based early rumor verification in social media. In: ECIR 2022. p. 496–504 (2022). https://doi.org/10.1007/978-3-030-99739-7_61
22. Hassan, N., Arslan, F., Li, C., Tremayne, M.: Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In: SIGKDD. pp. 1803–1812 (2017). https://doi.org/10.1145/3097983.3098131
23. Hassan, N., Li, C., Tremayne, M.: Detecting Check-Worthy Factual Claims in Presidential Debates. In: CIKM. p. 1835–1838 (2015). https://doi.org/10.1145/2806416.2806652
24. Jin, Z., Cao, J., Guo, H., Zhang, Y., Wang, Y., Luo, J.: Detection and Analysis of 2016 US Presidential Election Related Rumors on Twitter. In: Social, Cultural, and Behavioral Modeling - 10th International Conference. Lecture Notes in Computer Science, vol. 10354, pp. 14–24 (2017). https://doi.org/10.1007/978-3-319-60240-0_2
25. Konstantinovskiy, L., Price, O., Babakar, M., Zubiaga, A.: Toward Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. Digital Threats **2**(2) (2021). https://doi.org/10.1145/3412869
26. Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M.B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E., Baddour, K.: Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. Cureus **12**(3), e7255 (2020)

27. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: A lite BERT for self-supervised learning of language representations. In: ICLR (2020)
28. Liu, J., Liu, T., Yu, C.: Newsembed: Modeling news through pre-trained document representations. In: SIGKDD. pp. 1076–1086 (2021). https://doi.org/10.1145/3447548.3467392
29. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019)
30. Lv, Y., Zhai, C.: Adaptive term frequency normalization for BM25. In: CIKM. pp. 1985–1988 (2011). https://doi.org/10.1145/2063576.2063871
31. Lv, Y., Zhai, C.: When Documents Are Very Long, BM25 Fails! In: SIGIR. p. 1103–1104 (2011). https://doi.org/10.1145/2009916.2010070
32. MacAvaney, S., Yates, A., Cohan, A., Goharian, N.: CEDR: Contextualized Embeddings for Document Ranking. In: SIGIR. p. 1101–1104 (2019). https://doi.org/10.1145/3331184.3331317
33. Martín, A., Huertas-Tato, J., Álvaro Huertas-García, Villar-Rodríguez, G., Camacho, D.: FacTeR-Check: Semi-automated fact-checking through semantic similarity and natural language inference. Knowledge-Based Systems **251**, 109265 (2022). https://doi.org/10.1016/j.knosys.2022.109265
34. Miller, M., Banerjee, T., Muppalla, R., Romine, W., Sheth, A.: What are people tweeting about Zika? An exploratory study concerning its symptoms, treatment, transmission, and prevention. JMIR Public Health and Surveillance **3**(2), e38 (2017)
35. Mohtarami, M., Baly, R., Glass, J.R., Nakov, P., Màrquez, L., Moschitti, A.: Automatic stance detection using end-to-end memory networks. In: NAACL-HLT. pp. 767–776 (2018). https://doi.org/10.18653/v1/n18-1070
36. Müller, M., Salathé, M., Kummervold, P.E.: Covid-twitter-bert: A natural language processing model to analyse COVID-19 content on twitter. CoRR **abs/2005.07503** (2020)
37. Muric, G., Wu, Y., Ferrara, E.: COVID-19 Vaccine Hesitancy on Social Media: Building a Public Twitter Data Set of Antivaccine Content, Vaccine Misinformation, and Conspiracies. JMIR Public Health Surveill **7**(11), e30642 (2021)
38. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: A human generated machine reading comprehension dataset. In: NeurIPS. CEUR Workshop Proceedings, vol. 1773 (2016)
39. Nogueira, R., Cho, K.: Passage Re-ranking with BERT (2020). https://doi.org/10.48550/arXiv.1901.04085
40. Ortiz-Martínez, Y., Jiménez-Arcia, L.F.: Yellow fever outbreaks and Twitter: Rumors and misinformation. American Journal of Infection Control **45**(7), 816–817 (2017)
41. Oyeyemi, S.O., Gabarron, E., Wynn, R.: Ebola, Twitter, and misinformation: a dangerous combination? BMJ **349**, g6178 (2014)
42. Piergiorgio, C., Giulia, A., Riccardo, G., Eugenia, P., Manlio, D.D.: The voice of few, the opinions of many: evidence of social biases in Twitter COVID-19 fake news sharing. R. Soc. Open Sci. **9**(220716) (2022). https://doi.org/10.1098/rsos.220716
43. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: EMNLP-IJCNLP. pp. 3980–3990 (2019). https://doi.org/10.18653/v1/D19-1410
44. Robertson, S.E., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. Found. Trends Inf. Retr. **3**(4), 333–389 (2009). https://doi.org/10.1561/1500000019

45. Rosenthal, S., Ritter, A., Nakov, P., Stoyanov, V.: Semeval-2014 task 9: Sentiment analysis in twitter. In: SemEval@COLING. pp. 73–80 (2014). https://doi.org/10.3115/v1/s14-2009
46. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR **abs/1910.01108** (2019)
47. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake News Detection on Social Media: A Data Mining Perspective. SIGKDD Explor. Newsl. **19**(1), 22—-36 (2017). https://doi.org/10.1145/3137597.3137600
48. Soleimani, A., Monz, C., Worring, M.: BERT for Evidence Retrieval and Claim Verification. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) ECIR. pp. 359–366 (2020). https://doi.org/10.1007/978-3-030-45442-5_45
49. Starbird, K., Maddock, J., Orand, M., Achterman, P., Mason, R.M.: Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing. In: iConference 2014 Proceedings (2014). https://doi.org/10.9776/14308
50. Sun, S., Duh, K.: CLIRMatrix: A massively large collection of bilingual and multilingual datasets for Cross-Lingual Information Retrieval. In: EMNLP. pp. 4160–4170 (2020). https://doi.org/10.18653/v1/2020.emnlp-main.340
51. Thakur, N., Reimers, N., Daxenberger, J., Gurevych, I.: Augmented SBERT: data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In: NAACL-HLT. pp. 296–310 (2021). https://doi.org/10.18653/v1/2021.naacl-main.28
52. Volkova, S., Shaffer, K., Jang, J.Y., Hodas, N.: Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. In: ACL. pp. 647–653 (2017). https://doi.org/10.18653/v1/P17-2102
53. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018)
54. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In: NeurIPS (2020)
55. Yang, Y., Carbonell, J.G., Brown, R.D., Pierce, T., Archibald, B., Liu, X.: Learning approaches for detecting and tracking news events. IEEE Intell. Syst. **14**(4), 32–43 (1999). https://doi.org/10.1109/5254.784083
56. Zuo, C., Acharya, N., Banerjee, R.: Querying across genres for medical claims in news. In: EMNLP. pp. 1783–1789 (2020). https://doi.org/10.18653/v1/2020.emnlp-main.139
57. Zuo, C., Banerjee, R., Chaleshtori, F.H., Shirazi, H., Ray, I.: Seeing should probably not be believing: The role of deceptive support in covid-19 misinformation on twitter. J. Data and Information Quality **15**(1) (2022). https://doi.org/10.1145/3546914