Automated Evidence Collection for Fake News Detection

Mrinal Rawat
UpGrad Education Pvt. Ltd., Mumbai, India.
Liverpool John Moores University,
Liverpool, United Kingdom.
rawatmrinal06@gmail.com

Diptesh Kanojia
Centre for Translation Studies,
University of Surrey,
Surrey, United Kingdom.
d.kanojia@surrey.ac.uk

Abstract

Fake news, misinformation, and unverifiable facts on social media platforms propagate disharmony and affect society, especially when dealing with an epidemic like COVID-19. The task of Fake News Detection aims to tackle the effects of such misinformation by classifying news items as fake or real. In this paper, we propose a novel approach that improves over the current automatic fake news detection approaches by automatically gathering evidence for each claim. Our approach extracts supporting evidence from the web articles and then selects appropriate text to be treated as evidence sets. We use a pre-trained summarizer on these evidence sets and then use the extracted summary as supporting evidence to aid the classification task. Our experiments, using both machine learning and deep learning-based methods, help perform an extensive evaluation of our approach. The results show that our approach outperforms the state-of-the-art methods in fake news detection to achieve an F1-score of 99.25 over the dataset provided for the CONSTRAINT-2021 Shared Task. We also release the augmented dataset, our code and models [1] for any further research.

1 Introduction

The ability to consume readily available information from the internet is alarming for both individuals and organizations. The quality of content on social media platforms has been significantly affected due to the spread of fake news, misinformation and unverifiable facts. The current tally of internet users stands at 4.66 billion[2] (Kemp, 2015); and many of these users generate, post and consume content without any regulation, in a large number of countries[3]. Due to the unrestricted nature of online platforms, there is a significant increase in the amount of misinformation on social media (Allen et al., 2020), especially in developing nations (Badrinathan, 2020; Wasserman and Madrid-Morales, 2019). Studies show that events such as the presidential election of the United States in 2016 were affected due to *moderated fake news* campaigns (Tavernise, 2016). Shu et al.(2017) (Shu et al., 2017) propose that *fake news is intentionally written, verifiably false, and is created in a way that makes it look authentic*. Manual efforts by other online platforms such as Poynter[4], FactCheck[5], AltNews[6] *etc.* to detect fake news, requires a lot of human effort and can prove to be cumbersome. Such manual efforts can be time-consuming, challenging, and at times, can also be ineffective as fake news can spread faster than verified claims over social media platforms.

Automatic Fake News Detection is a task that aims to mitigate the problem of misinformation with the help of evidence supported by various sources. Most of the approaches in this recently devised task aim to use the classical machine learning-based methods or the recent deep learning-based methods to help classify news items as fake or as real. Initially proposed methods for the task applied machine learning-based techniques but cited insufficient data as a major concern (Vlachos and Riedel, 2014). Recent deep learning and ensemble approaches (Malon, 2018; Roy et al., 2018) were proposed on the FEVER (Thorne et al., 2018a) and LIAR (Wang, 2017) datasets, and have been shown to perform very well. Studies have proposed a combination of evidence detection with textual entailment

---

[1] https://github.com/rawat-mrinal06/fake_news
[2] Internet Live Stats (as on 15-07-2021)

[3] Internet Censorship in Countries
[4] Poynter: Online
[5] FactCheck: Online
[6] AltNews: Online

concerning the claim (Vijjali et al., 2020). FEVER Shared Tasks (Thorne et al., 2018b, 2019) have helped the automatic fact verification task gather attention towards the problem and helped generate approaches to mitigate the issues with previously proposed solutions. This study shows that our novel approach improvises over state-of-the-art approaches and helps detect fake news related to COVID-19. Our approach performs web-search for evidence collection and uses BERT-score similarity to match the unverified claim with the top-k searches. Further, we propose the use of summarization to mitigate problems with the evidence collection. We summarize the top-n selected lines from these articles and use them as evidence to support or reject the news item claim. Our experiments perform an extensive evaluation of the approach over the datasets released as a part of the CONSTRAINT-2021 Shared Task (Patwa et al., 2020) shared task.

Our summarized contributions with this paper are:

- We propose a novel approach to help automate the evidence collection for any fake news detection dataset.

- Additionally, we incorporate a summarization component that helps outperform the state-of-the-art approaches for automatic fact verification on the CONSTRAINT-2021 dataset.

2 Related Work

Automatic detection and classification of fake news, especially in epidemic situations like COVID-19, is a significant issue for society. Most of the recent works have identified that fake news is written intentionally and factually false (Shu et al., 2017). Several datasets have been released for the AI community in the field of fake news detection, such as LIAR (Wang, 2017), Fake News Challenge-1, and FEVER (Thorne et al., 2018a). Some recent techniques extract the evidence from Wikipedia to classify a claim as SUPPORTED, REFUTED or NOTENOUGHINFO (Thorne et al., 2018a). They formulate the problem as a three-step process (i) first the top-k documents are identified based on the TF-IDF based approaches (ii) then top-k sentences are identified from the documents, and (iii) finally the textual entailment based approaches (Parikh et al., 2016) are used to classify the claim. Team

Papelo (Malon, 2018) used the Transformer-based approach for the textual entailment and selected the evidence-based on tf-idf and entities present in the title. Hanselowski et al. (2018) selects the documents and sentence using the entity mentions and recognizes textual entailment using Enhanced Sequential Inference Model (ESIM) (Chen et al., 2017). Despite the several attempts, fake news detection is a challenging problem and countering fake news is a typical issue that requires continuous studies. Recently, some researchers released the datasets related to COVID-19 fake news detection. Shahi and Nandini (2020) (Shahi and Nandini, 2020) proposed the first multi-lingual cross-domain dataset for COVID-19 that consists of 5182 fact-checked news articles from Jan-2020 to May-2020. They collected data from 92 different websites and manually classified them into 23 classes. Kar et al. (Kar et al., 2020) also released a multi-indic-lingual dataset besides English to detect the fake news in social media tweets. They obtained 480 tweets in Bengali and 460 tweets in Hindi. In addition to tweets, they also included several features related to tweets such as retweet count, favourite count, total URL in description, URL, friend list, followers, *etc.* Recently, a very relevant dataset was released by Patwa et al. (Patwa et al., 2020) which consists of 10,700 tweets or claim collected from various sources such as Twitter, PolitiFact, Snopes, Boomlive. They experimented with various machine learning techniques like Decision Trees, Logistic Regression, SVM, Gradient Boosting DT and achieved the F1-score of 93.32. Most of the previous work on COVID-19 dataset proposed an ensemble approach of various models such as BERT, RoBERTa, XLNet, *etc.* (Shifath et al., 2021; Raha et al., 2021; Shushkevich and Cardiff, 2021). Chen et al. (Chen et al., 2021) trained the model with additional words such as covid-19, coronavirus, pandemic, indiafightscorona since the BERT tokenizer will split these words into separate tokens. Some works leveraged the fine-tuned models like COVID-Twitter-BERT (CT-BERT) (Müller et al., 2020) and demonstrated a boost in performance (Li et al., 2021; Glazkova et al., 2020; Wani et al., 2021). The fake news detection methods described above mainly uses the claim for the classification. Our method focuses on extracting and summarizing the evidence from the external source and uses it to classify the claim on the COVID-19 fake news detection dataset (Li et al., 2021; Patwa et al., 2020).

Split	Real	Fake	Total
Training	3360	3060	6420
Validation	1120	1020	2140
Test	1120	1020	2140
Total	5600	5100	10700

Table 1: Dataset Statistics

3 Dataset

For our work, we use the pre-released COVID-19 fake news dataset as a part of the CONSTRAINT-2021 shared task (Patwa et al., 2020). This gold-standard manually annotated dataset comprises social media posts and articles which are related to COVID-19. Each post or tweet contains content in the English language and is classified in either of the two categories- **(1) Real:** where tweets or articles which are factually correct and verified from authentic sources, for example, "Wearing mask can protect you from the virus. (Twitter)"; or **(2) Fake:** where tweets or posts related to COVID-19 which are factually incorrect and verified as false, for example, "If you take Crocin thrice a day you are safe. (Facebook)".

The authors collect fake news from two different sources- social media platforms and public fact-checking platforms. The social media posts include text from Facebook posts, Instagram posts, and Twitter posts, whereas the fact-checking websites such as PolitiFact, Snopes, and Boomlive are used to collect fact-checked news items. To further collect real news, they sample tweets from official government channels, news channels, and medical institutes. Overall, a total of 14 such sources were used to prepare this dataset.

The dataset comprises 10700 manually annotated samples and is split into (60%) train, (20%) validation and (20%) test sets. We provide the exact numbers for each split/class in Table 1 for clarity. The dataset is class-balanced as it contains 52.3% samples of real posts and 47.7% samples of fake posts. As an analysis on it, we obtained a word-cloud illustration for both real and fake samples and observed a high lexical overlap between both the classes, where words like '*coronavirus*', '*covid19*', '*people*', '*cases*', '*number*', '*test*', *etc.* are repeatedly used in both the sets. We do not show the word cloud due to space constraints. We create and present a wordcloud for the dataset in Figure 1.

(a) Worcloud of Real Posts   (b) Worcloud of Fake Posts

Figure 1: Wordcloud of real/fake posts in our dataset.

4 Our Approach

In this section, we provide details of our novel approach to augment the dataset with evidence from web search and the use of this evidence to complement the task of fact verification. The algorithm for our approach can be seen in Algorithm 1.

Algorithm 1: Algorithm to collect the evidence from the input claim

---

1   Claim c, Blocked URLs u Evidence e
2   **Function** ArticleRet(c, $k = 3$)**:**
3   results $\leftarrow GoogleSearch(\text{c})$;
4   filtered_results $\leftarrow \emptyset$;
5   **foreach** $r_i \in results$ **do**
6   **if** $r_i \notin u$ **then**
7         $s_i \leftarrow Similarity(\text{c},\text{r}_i)$;
          // Document Similarity using spacy library
8   **if** $s_i > 0.7$ **then**
9   filtered_results $\leftarrow (\text{r}_i, s_i)$
10   filtered_results $\leftarrow Sort(filtered\_results)[:\text{k}]$;
11   **return** filtered_results;
12   $articles \leftarrow$ ArticleRet(c);
13   $e \leftarrow \emptyset$;   // Evidences
14   **foreach** $a_i \in articles$ **do**
15     $d \leftarrow WebsiteData(a_i.url)$;
16     $sents \leftarrow d['<h>'] + d['<p>']$;
      // Extract <p> and <h> tags from html
17   **foreach** $s_i \in sents$ **do**
18       $sim \leftarrow Similarity(\text{c},\text{s}_i)$;
        // Document Similarity using spacy library
19   **if** $sim > 0.5$ **then**
20         $e \leftarrow (\text{s}_i, sim)$;
21     $e \leftarrow Sort(\text{e})[:3]$;
22   **return** e;

As discussed above, we collect this evidence and prune to top-k related news items based on semantic similarity via BERTScore (Devlin et al., 2019). We also select top-n lines from each article for further building an evidence repository, as detailed below in further subsections.

4.1 Evidence Collection

In the original dataset of COVID-19, evidence is not released along with the claim. We hypothesize that evidence is equally relevant to classify the claim as proposed by Thorne et al. (2018a). As per our approach, given a claim or post text, we first select K relevant articles using a BERT-based sentence similarity score as detailed here; and can be seen as an architecture component in Figure 2.

4.1.1 Article Retrieval

For each claim c, we search the claim as a query using a publicly available search API. The response returned by this API consists of (heading, text) pairs. We use the spacy (Honnibal et al., 2020) library to get the similarity score of response text with respect to the input claim. Based on this similarity score, we select top K results that have the similarity score greater than 0.7[7]. While selecting documents, we prune for webpages in other languages and pages which are direct links to PDF or other such non-text files. As an immediate next step, we scrape the selected web pages to obtain the matching N sentences concerning the claim as detailed here.

4.1.2 Sentence(s) Retrieval

In the previous step, we extract the relevant article URLs $U = (u_1, u_2, u_3)$. We employ a similar method to find the sentences within each article. For every url u, we first scrape the webpage and extract the text from $< h >$ and $< p >$ tags. Further, we use the same similarity score to select the top N sentences with respect to the claim. We obtain a similarity threshold of 0.5 after performing a similar empirical evaluation as mentioned in the footnote. Eventually, we concatenate the selected sentences from these articles, which act as our evidence for the claim. We would like to note that increasing the threshold significantly higher returned an empty set in some case and hence we choose a relatively lower threshold (0.5).

An example of evidence collected via our approach is shown in Table 2 where the column titled "Evidence" shows the output after these steps.

4.2 Dataset Preprocessing

To map claims with evidence, we pre-process both the dataset and the evidence collected from external sources. Following are the details of the pre-processing steps: **(1) URL Mapping:** We observe that some posts contain URLs in a masked form, *e.g.,* https://t.co/z5kk XpqkYb. Our approach extracts these URLs using a regular expression-based match and maps them to the original URL using the python '*requests*' library. Any additional information from the URL is removed, and only appropriate URLs remain in the text. For example, https://t.co/z5kkXpqkYb → https://www.cdc.gov/; **(2) Special symbols:** We removed extra whitespaces, special symbols and brackets like ;, (,), {,}; **(3) Hashtags, Emojis and Mentions:** Additionally, we remove hashtags and replace it with the token "HASHTAG:".

For example, #COVID-19 becomes HASHTAG:COVID-19. Similarly, we also replace mentions "@" with "MENTION:" token. At the end, we convert emojis to their text form using the 'demoji' library[8]; **(4) Lowercasing:** Eventually, we lowercase the claim and the evidence text to obtain the input data used for the next Summarization step.

4.3 Summarization of Evidence

Our pre-processed evidence for claims in many cases were multiple paragraphs resulting in performance degradation. Therefore, we propose the addition of a summarization component to our pipeline which utilizes state-of-the-art Text-to-Text Transfer Transformer (T5) language model (Raffel et al., 2019) for the inherent summarization task[9]. Due to the nature of the usual summarization task input, a large body of text (full documents), we believe that our comparatively short paragraphs would be better summarized. This language model is fine-tuned for the task of summarization helps us obtain a summarized text for each piece of evidence resulting in what we call Summarization-1 or S1. An output obtained is shown in Table 2.

---

[7]Selected with empirical evaluation and manual analysis after trying 0.5, 0.6, 0.7, 0.8 using the 'en_nli_roberta_base' model for document similarity

[8]GitHub: Demoji

[9]This language model can perform the summarization task with the help of a prefix "summarize" to the input text provided.
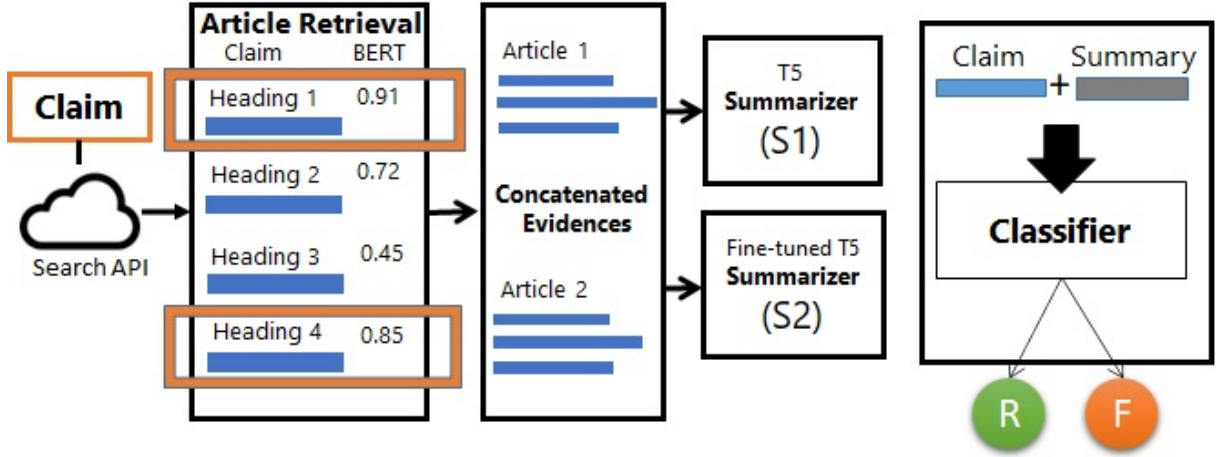
Figure 2: Full architecture of our proposed approach.

Claim	Evidence	Summarization-1 (S1)	Summarization-2 (S2)
There is no evidence that children have died because of a COVID-19 vaccine. No vaccine currently in development has been approved for widespread public use. https://t.co/9ecvMR8SAf	Currently there is no coronavirus vaccine that has been approved for the American public. And there is no evidence that children have died because they received one of the COVID-19 vaccines being developed. PolitiFact found no evidence that anyone has died from complications related to a trial COVID-19 vaccination. There is no evidence that children have died because of a COVID-19 vaccine.	There is no evidence that children have died because they received a COVID-19 vaccine. No evidence that anyone has died from complications related to a trial COVID-19.	There is no evidence that children have died because they received one of the COVID-19 vaccines being developed. PolitiFact found no evidence that anyone has died from complications related to a trial COVID-19 vaccination.

Table 2: Illustrative example of our approach pipeline shown as Claim → Evidence → Summarization-1 → Summarization-2, where Summarization-2 is obtained after fine-tuning T5 language model, and used as **evidence input** for classification

4.3.1 Fine-tuning T5 on FEVER Dataset

As an additional experimental step, we further fine-tune the Text-to-Text Transfer Transformer (T5) model using the original FEVER dataset (Thorne et al., 2018a). The original T5 summarization model is trained on the CNN/Daily Mail (Hermann et al., 2015) data where the input is the news article text, and the objective is to highlight summarized text as the output. The T5 is an encoder-decoder model pre-trained on a multi-task mixture of unsupervised and supervised tasks and for which each task is converted into a text-to-text format. This allows for the use of the same model, loss function, hyperparameters, *etc.* across our diverse set of tasks. T5 works well on a variety of tasks out-of-the-box by prepending a different prefix to the input corresponding to each task, e.g., for the task of translation→ *translate English to German: <English Sentence>*, for the task of summarization→ *summarize: <English Text>*.

For our experiments, the aim is to summarize the pre-processed evidence while including the claim. Thus, we hypothesize that fine-tuning on an auxiliary dataset will improve the quality of the generated summary. For fine-tuning, we use the same hyperparameters as described in their paper to generate another model. We perform another iteration of the summarization step using this fine-tuned model to generate a parallel set of evidence and label the output as Summarization-2 or S2 as shown in Table 2. Further, we provide the details of **as the classification task, which uses either S1 or S2 as evidence input to classify** the claims as real or fake (Figure 2).

5 Experiment Setup

In this section, we discuss the experiment setup in detail. We perform the task of fake news detection as a binary classification task in a supervised setting. We choose to perform our experiments with both conventional machine learning- and deep learning- based classifiers. From the machine learning-based approaches, we choose Logistic Regression (LR) and Support Vector Machines (SVM) with the GridSearch implementation for best results over multiple hyperparameters (values of c, different kernels, *etc.*) We also utilize LSTMs with various contextual language models from the deep learning methods. From the deep learning-based approaches, we use a simple LSTM

	Previous Approaches		Our Approach w/ various Classification methods								
	Chen et al. (2021)	Li et al. (2021)	Logistic Regression			SVM			LSTM		
			-	S1	S2	-	S1	S2	-	S1	S2
P	0.9902	0.986	0.9531	0.9565	0.9701	0.9641	0.9671	0.9764	0.9589	0.9598	0.9612
R	0.9901	0.985	0.9531	0.9564	0.9700	0.9639	0.9668	0.9761	0.9584	0.9596	0.9612
F	0.9901	0.985	0.9531	0.9565	<u>0.9700</u>	0.9639	0.9668	<u>0.9761</u>	0.9584	0.9596	<u>0.9612</u>

Table 3: Results obtained after the fake news classification task where the values for previous approaches are from the latest shared task results and the results for each iteration of our approach are shown [P (Precision), R (Recall), and F (F-Score)]. (-) → No Evidence, S1 → Summarization-1 as Evidence, S2 → Summarization-2 as Evidence.

	Our Approach w/ various Deep Learning Classification methods								
	$BERT_{base}$			$RoBERTa_{base}$			$XLNet_{base}$		
	-	S1	S2	-	S1	S2	-	S1	S2
P	0.9612	0.9916	0.9917	0.9918	0.9929	0.9922	0.9920	0.9934	0.9947
R	0.9864	0.9888	0.9897	0.9897	0.9911	0.9916	0.9892	0.9911	0.9925
F	0.9858	0.9888	<u>0.9893</u>	0.9893	0.9908	**0.9908**	0.9892	0.9910	**0.9925**

Table 4: Results obtained after the fake news classification task where the results for each iteration of our approach with various deep learning classification methods are shown [P (Precision), R (Recall), and F (F-Score)]. (-) → No Evidence, S1 → Summarization-1 as Evidence, S2 → Summarization-2 as Evidence.

implementation with pre-trained GloVE[10] vectors, $BERT_{base}$, $RoBERTa_{base}$, and $XLNET_{base}$ -based classifiers. Our LSTM implementation uses *Adam* optimizer with a learning rate of 0.001, and 256 as the batch size. For classifiers based on $BERT_{base}$, $RoBERTa_{base}$, and $XLNET_{base}$, we use the HuggingFace implementations with a batch size of 32, L2 regularization and cross-entropy loss. The regularization parameter λ was set to 0.1. Each classification method is iterated **(1)** without evidence (-), **(2)** with augmented summarized evidences from S1, **(3)** and then with S2, thus giving us three sets of results for each method; as shown in Table 4.

As an input to the classifier, we use the claim as-is from the dataset as described above. We have a dataset $D = (x_n, y_n)_{n=1}^N$ comprising of N training samples. Here $x_n = (c_n, e_n)$, where c_n represents the claim, and e_n represents evidence gathered using our approach. $X \in \mathcal{X}$ is defined on input space, and $Y \in \mathcal{Y} = \{0, 1\}$ are the corresponding labels. Thus, given a claim c and evidence e, the aim of this task is to train a classifier such that the claim c is predicted as fake news or not, i.e $F_\theta : \mathcal{X} \to \mathcal{Y} \in \{0, 1\}$.

$$F(c, e; \theta) = \begin{cases} 1, & \text{if } c \text{ is the fake news} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $F(c, e)$ is the function our model aims to learn over each iteration or epoch.

---

[10]https://nlp.stanford.edu/projects/glove/

6 Results and Discussion

The results for our classification task are shown in Table 4. Using our approach, we are able to marginally outperform (+0.24, F-Score) the previous state-of-the-art (SoTA) approaches for the task of fake news detection as shown in the last column ($XLNet_{base}$, S2). Even the $RoBERT_{base}$ model is able to outperform the SoTA approaches by a small margin. We present the values of our top two best models in boldface in Table 4. Although the improvement margin is small, we would like to note that the previous SoTA approaches are already performing at almost a 0.99 F-score. *We executed our model run multiple times to ensure that our improvement margin is indeed truly obtained.* We also observe that $RoBERTA_{base}$, and $XLNet_{base}$ outperform the SoTA approaches (Chen et. al. / Li et. al.) even with S1 summarization component. Classical machine learning-based approaches are also shown to perform very well for this task as the scores of 0.96 can be considered to be a good performance for any classification method.

However, this is not the only key takeaway from these results. *We observe that by using our novel approach, a consistent improvement is seen in the task results*. The efficacy of our approach can be seen from Table 4, as either S1 or S2 consistently outperforms all the base models (-) [no evidence] in the table. Moreover, using our approach, we are able to gather key evidence for such a dataset

Text	XLNet	LR	SVM
We always appreciate questions about the quality of our data. If you see a number that doesn't look right please file an issue at and we will investigate. [SEP] SOURCES: github.com	✗	✓	✓
The number of daily tests has been increasing in a steep climb. Average daily tests during the past three weeks also strongly depict the progress made in enhancement of #COVID19 tests across the country. [SEP] SOURCES: twitter.com/MoHFW_INDI	✗	✓	✓
Bill Gates said thousands of people will die with the COVID-19 vaccine [SEP] SOURCES:	✗	✗	✗

Table 5: Qualitative error analysis of some output cases both in terms of successes and failures of our approach.

where, to begin with, only claims were present with manually annotated labels. Table 5 illustrates the success and failure cases from XLNet, Logistic Regression and Support Vector Machines. We observe that first two cases were incorrectly predicted by the XLNet but were predicted correctly by LR and SVM. Last case was predicted incorrectly by all of the models. We believe that the absence of source in the text could be a potential reason for this failure. Our approach can gather the evidence using a fully automated method with summarization component(s) in the pipeline. The importance of this component can be gathered from manual observations of examples in the augmented dataset. We observe that summarized evidences shorten the length of the evidences, which helps the Transformer architecture-based classifiers like BERT$_{base}$, RoBERTa$_{base}$, XLNet$_{base}$ perform better. These pre-trained models have a token length limitation of 512 tokens which is easily able to capture our summarized evidence. We also manually observe that the summarization component helps reduce redundancy in the generated sentences and removes duplicates. Hence, improving the quality of evidence used as additional input helps reduce the training time. The performance of our models with the fine-tuned summarization component (S2) seems to perform better than S1, and the model without any evidence, as can be seen in Table 4.

We acknowledge that the CONSTRAINT dataset is saturated in terms of possible improvements. However, with this paper, *our aim is to show the efficacy of our summarization technique which can help the evidence detection for news*. We chose this dataset at an early stage of our work, and our experiments do show that improvements can, in fact, still be shown on this dataset. Our best-performing system surpasses the state-of-the-art by 0.23% points.

7 Conclusion and Future Work

In this paper, we present an automated method to collect evidence for the fake news detection task. We use our novel approach to augment the dataset, released in the CONSTRAINT-2021 Shared Task, with evidence sets collected from the web. Our method helps process these evidence sets, clean them and use them to generate summarized evidence based on two different methodologies. We use either of the summarized evidence as an additional input to the fake news classification task and perform an evaluation of our approach. We discuss the results of the classification task and conclude that our approach helps outperform the previous SoTA approaches by a small margin, however, helping generate evidence for a crucial dataset. We show that a summarization module can help collect evidence more effectively. We augment this dataset with the summarized evidence and release it along with the code and generated models for further research. We would also like to conclude that our method is generalizable; since it uses pre-trained metrics (BERTScore) and models (T5), it can be used to gather evidence for other datasets. The overall pipeline is also not very time-consuming (2 seconds per sample) once fine-tuned models are included in it. We hope our method and the resources are helpful to the NLP community.

In future, we would like to use our method to gather evidence for other fact detection/verification datasets as well. Our initial aim is to reproduce this study with other datasets and ensure that our method performs well in a real-world scenario. We would also like to apply this method and gather further evidence for existing fake news datasets, and perform our experiments to evaluate this approach over multiple exisiting datasets, including existing multilingual datasets.

References

Jennifer Allen, Baird Howland, Markus Mobius, David Rothschild, and Duncan J Watts. 2020. Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, 6(14):eaay3539.

Sumitra Badrinathan. 2020. Educative interventions to combat misinformation: Evidence from a field experiment in india. *American Political Science Review*, pages 1–17.

Ben Chen, Bin Chen, Dehong Gao, Qijin Chen, Chengfu Huo, Xiaonan Meng, Weijun Ren, and Yang Zhou. 2021. Transformer-based language model fine-tuning methods for covid-19 fake news detection. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 83–92, Cham. Springer International Publishing.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Anna Glazkova, Maksim Glazkov, and Timofey Trifonov. 2020. g2tmn at constraint@aaai2021: Exploiting CT-BERT and ensembling learning for COVID-19 fake news detection. *CoRR*, abs/2012.11967.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Debanjana Kar, Mohit Bhardwaj, Suranjana Samanta, and Amar Prakash Azad. 2020. No rumours please! A multi-indic-lingual approach for COVID fake-tweet detection. *CoRR*, abs/2010.06906.

Simon Kemp. 2015. Global digital & social media stats: 2015. *Social Media Today*.

Xiangyang Li, Yu Xia, Xiang Long, Zheng Li, and Sujian Li. 2021. Exploring text-transformers in aaai 2021 shared task: Covid-19 fake news detection in english. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 106–115, Cham. Springer International Publishing.

Christopher Malon. 2018. Team papelo: Transformer networks at FEVER. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 109–113, Brussels, Belgium. Association for Computational Linguistics.

Martin Müller, Marcel Salathé, and Per Egil Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse COVID-19 content on twitter. *CoRR*, abs/2005.07503.

Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *CoRR*, abs/1606.01933.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Tathagata Raha, Vijayasaradhi Indurthi, Aayush Upadhyaya, Jeevesh Kataria, Pramud Bommakanti, Vikram Keswani, and Vasudeva Varma. 2021. Identifying COVID-19 fake news in social media. *CoRR*, abs/2101.11954.

Arjun Roy, Kingshuk Basak, Asif Ekbal, and Pushpak Bhattacharyya. 2018. A deep ensemble framework for fake news detection and classification. *CoRR*, abs/1811.04670.

Gautam Kishore Shahi and Durgesh Nandini. 2020. Fakecovid - A multilingual cross-domain fact check news dataset for COVID-19. *CoRR*, abs/2006.11343.

S. M. Sadiq-Ur-Rahman Shifath, Mohammad Faiyaz Khan, and Md. Saiful Islam. 2021. A transformer based approach for fighting COVID-19 fake news. *CoRR*, abs/2101.12027.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on socialwang media: A data mining perspective.

Elena Shushkevich and John Cardiff. 2021. TUDublin team at Constraint@AAAI2021 – COVID19 Fake News Detection. *arXiv e-prints*, page arXiv:2101.05701.

Sabrina Tavernise. 2016. As fake news spreads lies, more readers shrug at the truth. *The New York Times*, 6.

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

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and verification (fever) shared task.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019. The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 1–6, Hong Kong, China. Association for Computational Linguistics.

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

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Apurva Wani, Isha Joshi, Snehal Khandve, Vedangi Wagh, and Raviraj Joshi. 2021. Evaluating deep learning approaches for covid19 fake news detection. *CoRR*, abs/2101.04012.

Herman Wasserman and Dani Madrid-Morales. 2019. An exploratory study of "fake news" and media trust in kenya, nigeria and south africa. *African Journalism Studies*, 40(1):107–123.