Multiple Evidence Combination for Fact-Checking of Health-Related Information

Pritam Deka, Anna Jurek-Loughrey, Deepak P
Queen's University Belfast, UK
{pdeka01, a.jurek}@qub.ac.uk, deepaksp@acm.org

Abstract

Fact-checking of health-related claims has become necessary in this digital age, where any information posted online is easily available to everyone. The most effective way to verify such claims is by using evidences obtained from reliable sources of medical knowledge, such as PubMed. Recent advances in the field of NLP have helped automate such fact-checking tasks. In this work, we propose a domain-specific BERT-based model using a transfer learning approach for the task of predicting the veracity of claim-evidence pairs for the verification of health-related facts. We also improvise on a method to combine multiple evidences retrieved for a single claim, taking into consideration conflicting evidences as well. We also show how our model can be exploited when labelled data is available and how back-translation can be used to augment data when there is data scarcity.

1 Introduction

In today's age of easy access to the internet, information exchange among people has increased rapidly, which has also resulted in the spread of misinformation (Vosoughi et al., 2018) within the society. Misinformation has been found to spread faster than real news, and the rise of social media popularity has aided the spread of misinformation (Vosoughi et al., 2018). Research on health misinformation is still an ongoing area, as it is different from political misinformation on the basis of the complexity level of fact-checking (Deka et al., 2022b). Manual fact-checking of health information requires domain-specific experts, which increases both time taken and cost incurred. Automated fact-checking of health information found online has been aided by the release of datasets such as SCIFACT (Wadden et al., 2020), HEALTHVER (Sarrouti et al., 2021), COVID-FACT (Saakyan et al., 2021). Fact-checking of health information comprises of retrieving evidences from reliable resources which either supports or refutes the key claim (Zeng et al., 2021; Guo et al., 2022). Recent works have focused on building end-to-end fact-checking models evaluating on the aforementioned datasets (Pradeep et al., 2020; Zhang et al., 2021; Li et al., 2021; Wadden et al., 2022). However, they do not take into account conflicting evidences retrieved for a single claim. Any claim can have more than one evidence, and these evidences can be conflicting in real-world scenarios wherein one evidence would be supporting the claim and another evidence may be refuting the claim.

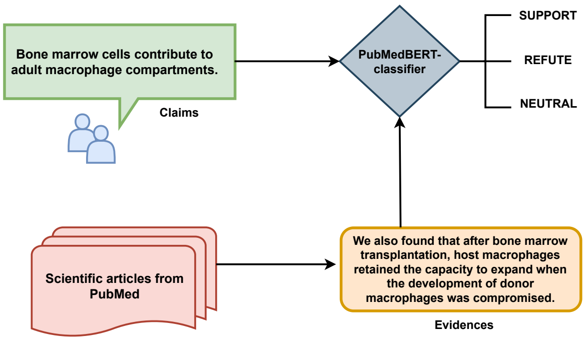
Bone marrow cells contribute to adult macrophage compartments.
Claims
PubMedBERT-classifier
SUPPORT
REFUTE
NEUTRAL
Scientific articles from PubMed
We also found that after bone marrow transplantation, host macrophages retained the capacity to expand when the development of donor macrophages was compromised.
Evidences

Figure 1: An example of claim and evidence from SCIFACT (Wadden et al., 2020) dev set

In this work, we have focused on the subtask of classifying a claim-evidence pair as either supporting, refuting, or neutral as shown in Figure 1. We assume in this work that evidences for claims are already retrieved. We proposed a domain specific BERT-based model using a transfer learning approach where the model is trained over textual entailment data which can then be applied directly over fact-checking data. We have also used the Dempster-Shafer theory (Dempster et al., 2008; Shafer, 1976) of evidence combination for mitigating the conflicting evidences issue to provide an end result. We then extend our work by showing

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 237–247
July 13, 2023 ©2023 Association for Computational Linguistics

how data augmentation techniques can help in a more robust training for smaller datasets with the help of neural machine translation language models. We also analyse the performance of our model when it is trained over other similar datasets. We further share our trained model publicly for further research[1].

2 Related work

In this section, we will discuss the research work that has been done for fact-checking scientific claims using evidences retrieved from existing medical article repositories. With the release of the SCIFACT dataset, various transformer-based methods of predicting the veracity labels using evidences for scientific claims have been proposed and evaluated using the dataset. (Wadden et al., 2020) established a pipeline model using a RoBERTa-large (Liu et al., 2019) model to retrieve evidences from PubMed abstracts. The retrieved evidence sentences are then passed along with the claims to predict whether the evidences SUPPORT or REFUTE the claims using a RoBERTa-large model fine-tuned over the training set of SCIFACT.

VerT5erini (Pradeep et al., 2020) uses a T5 (Raffel et al., 2020) model-based pipeline for their work. For the evidence sentence selection task for the claims, as well as for label prediction from PubMed abstracts, they used two different T5 models. For the sentence selection task, the T5 model used is fine-tuned over the MS-MARCO (Bajaj et al., 2016) dataset and then further trained on SCIFACT. For the label prediction task, the T5 model is trained on the SCIFACT dataset.

PARAGRAPH-JOINT (Li et al., 2021) uses a RoBERTa-large model similar to (Wadden et al., 2020) for both the evidence sentence selection as well as the label prediction task which is fine-tuned over SCIFACT. However, the training approach is different, as (Li et al., 2021) uses a multitask learning approach for model training. Both the tasks of sentence selection and label prediction are done using a joint cross-entropy loss as the training objective. For the label prediction task, the authors have also used two different approaches which includes a simple sentence-level attention and KGAT which is a Kernel Graph Attention Network (Wang et al., 2019a).

Similarly, ARSJOINT (Zhang et al., 2021) also uses a joint approach where their proposed method jointly learns the three tasks of abstract retrieval, sentence selection, and label prediction. Similar to (Wadden et al., 2020), they have also used RoBERTa-large for their work together with BioBERT-large (Lee et al., 2020).

All the above works focus on the three tasks of abstract retrieval, evidence sentence selection and label prediction as a pipeline approach. However, there is a difference in the sentence selection task as well as the label prediction. VerT5erini selects sentences independently, whereas PARAGRAPH-JOINT and ARSJOINT use the abstracts to select the sentences. The label prediction also differs as both PARAGRAPH-JOINT and ARSJOINT use a joint approach unlike VerT5erini. The models used in the tasks also differ as PARAGRAPH-JOINT and ARSJOINT use BERT-based models whereas VerT5erini uses a much larger T5 model having superior performance. However, the current state-of-the-art method, MULTIVERS (Wadden et al., 2022) differs from these works in the approach and the transformer model used. MULTIVERS uses a Longformer (Beltagy et al., 2020) architecture to encode both claims and abstracts together so that there is a minimum loss of information. The authors have used a weak supervision approach, in which the Longformer model is trained on available scientific data before fine-tuning on SCIFACT. However, the overall pipeline training method is a multi-task approach similar to (Li et al., 2021). It outperforms the other approaches in the label prediction task of SCIFACT.

Contrary to the above works, our approach is different in the way that for a given pair of claim and evidence, our model can predict the labels in a zero-shot approach, surpassing the state-of-the-art results without the need for any supervision. The above mentioned works have the end goal of predicting the labels of the claim-evidence pairs. However, to have a final prediction for the claims whether it is a “True” claim or “False” claim, we need to have a combined judgement of all the evidence sentences for that claim which is not addressed by the above works. We have extended our approach to include the final prediction for the claims taking into consideration conflicting evidences as well.

[1] https://huggingface.co/pritamdeka/PubMedBERT-MNLI-MedNLI

3 Task Formulation

In this section, we will first discuss the problem statement and then proceed with the formulation of the tasks. The problem statement is "Given a claim and a number of evidence sentences, determine whether the claim is **True** or **False** or **Neutral**". We can formulate the problem as two tasks:

- **Classification of claim-evidence pair** Given a claim c and an evidence sentence s for that claim, classify the claim-evidence sentence pairs as supporting, refuting or neutral.

$$[c, s] \xrightarrow{\text{classify}} \text{(support, refute, neutral)}$$

- **Prediction of the claims** Given a list of supporting or refuting evidence sentences S where $S = [s_1, s_2 \ldots s_n]$ for a claim c, the task is to predict whether the claim is **True** or **False** or **Neutral** by combining all the evidences.

$$[c, S] \xrightarrow{\text{predict}} \text{c(true, false, neutral)}$$

4 Methodology

In this section we will describe in detail the proposed methods we have adopted for the formulated tasks.

4.1 Classification of claim-evidence pair

Previous studies have focused on using fact-checking datasets such as FEVER (Thorne et al., 2018) for training models for the task of fact-checking. However, we have modelled the classification task as a natural language inference (NLI) problem, since (Pradeep et al., 2020) found in their study that models learn better from NLI data than datasets such as FEVER. Textual entailment or NLI is defined as the task of determining if, given a "premise", a "hypothesis" is true (entailment) or false (contradiction) or not determined (neutral) (Williams et al., 2017). Fact-checking has similarities with the NLI task, in which premises can be modelled as evidences and the hypothesis as claims (Thorne et al., 2018). The idea is to train domain specific BERT (Devlin et al., 2018) model using NLI data to see if the model can learn knowledge that can be transferred to fact-checking task in biomedical domain. In order to achieve this, we have trained PubMedBERT[2] (Gu et al., 2021) which is a domain specific BERT model on the multi-NLI (MNLI) dataset (Williams et al., 2017) by minimizing the cross-entropy loss. We also experimented with other models such as BioBERT (Lee et al., 2020) and SciBERT (Beltagy et al., 2019), however, we achieved the best performance with the PubMedBERT model which is why we chose this model.

In order to fine-tune PubMedBERT on the MNLI data, the sentence pair of hypothesis-premise is used as the initial input sequence. Once the model is trained over the MNLI dataset, we can directly transfer the model to the claim-evidence sentence pairs for the fact-checking task. We then trained it on the MedNLI (Romanov and Shivade, 2018) dataset, which is a domain-specific NLI dataset, as previous research (Phang et al., 2018; Wang et al., 2018) has shown that further training over a similar domain-specific dataset increases model performance. The learned model can then make predictions whether an evidence supports or refutes a claim or if it is undetermined. The model can also be adapted in a supervised way when fact-checking datasets are available. We have done extensive experiments to show how the model can also be adapted for unseen data.

4.2 Prediction of the claims

The classifier can assign a claim-evidence sentence pair as support or refute. In order to further predict the claims as true or false, we need to combine all the evidences for each of the claims. In more complex scenarios, when some evidences support a claim whereas others refute the same claim and yet others may be undetermined, this task is not trivial. To resolve such conflicting cases, we have used an improvised Dempster-Shafer (D-S) theory of evidence combination. Research has mainly focused on using the D-S theory for multi-sensor domain where evidences from multiple sources are combined to achieve a final decision (Xiao, 2019; Khan and Anwar, 2019; Smets, 2000; Jiang et al., 2016). This is similar to our work and we have used the D-S theory for combining multiple conflicting evidences for claims in order to achieve a final decision for the claims. The D-S theory is mathematically defined as follows (Dempster et al., 2008):

Definition 1. The set of all the possible sets of the hypotheses or class categories is known as the frame of discernment (FOD). A frame of discernment consisting of N elements where each element

[2] https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext

a_i is mutually exclusive to each other can be defined as:

$$\theta = \{a_1, a_2, a_3 \ldots, a_N\} \tag{1}$$

Definition 2. If A is a subset of $P(\theta)$ where $P(\theta) = 2^\theta$, the basic probability assignment (BPA) or mass function, $m(A)$ is a function that maps $A \rightarrow [0, 1]$ and satisfies the following conditions:

$$m(\phi) = 0, \sum_{A \subseteq \theta} m(A) = 1 \tag{2}$$

Definition 3. If $m_1(A_i)$ and $m_2(A_j)$ are the BPA of two bodies of evidence (BOE), then according to D-S combination rule, they can be combined as follows:

$$m(A) = \frac{1}{1-K} \sum_{A_i \cap A_j = A} m_1(A_i) m_2(A_j), \quad A \neq 0 \tag{3}$$

where K is a normalization factor defined as follows:

$$K = \sum_{A_i \cap A_j = \phi} m_1(A_i) m_2(A_j) \tag{4}$$

Definition 4. The combination formula can be extended for n terms as well which is defined as:

$$m(A) = \frac{1}{1-K} \sum_{A_{i_1} \ldots \cap A_{i_n} = A} m_1(A_{i_1}) \ldots m_n(A_{i_n}), \quad A \neq 0 \tag{5}$$

and K is defined as follows:

$$K = \sum_{A_{i_1} \ldots \cap A_{i_n} = \phi} m_1(A_{i_1}) \ldots m_n(A_{i_n}) \tag{6}$$

4.3 Illustrative example

In order to understand the working of the D-S theory, let us take a few examples. According to our work, let us take three classes for the FOD, $\theta = \{a, b, c\}$ where a, b, c are “support”, “refute” and “neutral” respectively.

Example 1. Let us take two conflicting evidences with respective probabilities for a, b and c.

$E1 : m_1(a) = 0.062$	$m_1(b) = 0.937$	$m_1(c) = 0.001$
$E2 : m_2(a) = 0.952$	$m_2(b) = 0.048$	$m_2(c) = 0$

We can see that for both E1 and E2, equation 2 is fulfilled. According to equation 4, we get

$$K = m_1(a)m_2(b)m_2(c) + m_1(b)m_2(a)m_2(c) + m_1(c)m_2(a)m_2(b)$$

Putting the respective values, we get $K = 0.896$. Using equation 3, we get the following

$$m(a) = \frac{m_1(a)m_2(a)}{(1-K)}, \; m(b) = \frac{m_1(b)m_2(b)}{(1-K)} \; \text{and} \; m(c) = \frac{m_1(c)m_2(c)}{(1-K)}$$

After calculation, we get $m(a) = 0.436$, $m(b) = 0.563$ and $m(c) = 0$. We can see that $m(b)$ has the highest probability value using the D-S combination theorem.

In certain situations, the D-S theorem fails. Let us look at one such example.

Example 2. Let us take four conflicting evidences with respective probabilities for a, b and c.

$E1 : m_1(a) = 0.889$	$m_1(b) = 0.106$	$m_1(c) = 0.005$
$E2 : m_2(a) = 0.0$	$m_2(b) = 0.999$	$m_2(c) = 0.0$
$E3 : m_3(a) = 1.0$	$m_3(b) = 0.0$	$m_3(c) = 0.0$
$E4 : m_4(a) = 0.481$	$m_4(b) = 0.515$	$m_4(c) = 0.004$

We can see that for both E1 and E2, equation 2 is fulfilled. However, here, we find that $K = 1$ which means that the denominator is $1 - K = 0$. In such situations, the D-S combination rule will fail as division by zero is mathematically undefined.

Definition 5. In order to overcome such situations, we adapted the base belief function from (Wang et al., 2019b) which is defined as follows: Let δ be a set of N possible values that are mutually exclusive. The power set of δ is 2^δ, in which the number of elements is 2^N. According to (Wang et al., 2019b), the base belief function m_{base} is then defined as:

$$m_{base}(A_i) = \frac{1}{2^N - 1} \tag{7}$$

where A_i is the subset of δ except for the empty set ϕ. The modified BPA then becomes

$$m'(A_i) = \frac{m_1(A_i) + m_{base}(A_i)}{2} \tag{8}$$

where $m_1(A_i)$ is the original BPA. This modified BPA allows us to mitigate situations when BPA values are 0. However, this leads to the violation of the condition $\sum_{A \subseteq \theta} m(A) = 1$. In order to preserve the condition, we normalize the value of $m'(A_i)$ and therefore the final BPA is:

$$m'_{norm}(A_i) = \frac{m'(A_i)}{\sum m'(A_i)} \tag{9}$$

Example 3. Using the modifications, from Example 2, the modified BPAs are as follows

$E1 : m_1(a) = 0.723$	$m_1(b) = 0.174$	$m_1(c) = 0.104$
$E2 : m_2(a) = 0.100$	$m_2(b) = 0.801$	$m_2(c) = 0.100$
$E3 : m_3(a) = 0.801$	$m_3(b) = 0.100$	$m_3(c) = 0.100$
$E4 : m_4(a) = 0.437$	$m_4(b) = 0.461$	$m_4(c) = 0.102$

We can see that for both E1 and E2, equation 2 is fulfilled. From equation 6, we can calculate $K = 0.0318$ and $1 - K = 0.968$. After that we can use equation 5 and get the values of $m(a) = 0.795$, $m(b) = 0.201$ and $m(c) = 0.0033$. Using the modified BPAs have helped overcome situations where the D-S combination rules fail.

From the illustrative examples, we have shown how the modified D-S method can be used for the task of combining evidences for a claim taking into account conflicting evidences as well.

5 Experimental Details

In this section, we will describe the experimental details for our tasks.

5.1 Dataset used

For evaluation purposes, we have used the SCIFACT dataset. The SCIFACT dataset has a train file, a dev file and a test file. However, the test file is part of a shared task and as such the labels are not available. This is why we are evaluating directly on the dev file. However, some claims in the dev set do not have the evidence sentences and as such we cannot evaluate on those claims which is why we have dropped those claims.

5.2 Classification of claim-evidence pair

For this task, we have experimented on different scenarios by doing an evaluation study over different classification settings. First we experimented directly on the dev set where we use our PubMedBERT fine-tuned model directly on the dev set examples by passing the claim-evidence pair and predicting the labels. The results are shown in Table 1 showing improvements over other models where P, R and F-1 are the precision, recall and f score respectively.

Model	**P**	**R**	**F-1**
PubMedBERTmnli	0.666	0.599	0.631
DeBERTa-v3-base-mnli	0.426	0.390	0.408
DeBERTa-v3-base-mnli-fever-anli	0.428	0.380	0.403

Table 1: Comparison of models directly on SCIFACT dev set examples

For the second experiment, we fine-tuned the MNLI fine-tuned model over MedNLI to see if there is any performance difference. Experiments using this model yielded a very good result which shows that in order to achieve an increased performance while fine-tuning over a smaller dataset, it is better to first fine-tune over a larger dataset and then use that model to further fine-tune over the smaller dataset. To confirm the results, we also compared the performance of a few more models from Table 1 by further fine-tuning these models over MedNLI. We can see from Table 2 that there is a performance increase in all models. This is in line with the findings by (Phang et al., 2018; Wang et al., 2018; Clark et al., 2019; Sap et al., 2019).

Model	**P**	**R**	**F-1**
PubMedBERT-mnli	0.666	0.599	0.631
PubMedBERT-mednli	0.543	0.465	0.501
PubMedBERT-mnli-mednli	0.847	0.753	0.797
DeBERTa-v3-base-mnli-fever-anli	0.428	0.380	0.403
DeBERTa-v3-base-mnli-fever-anli-mednli	0.748	0.666	0.705
DeBERTa-v3-base-mnli	0.426	0.390	0.408
DeBERTa-v3-base-mnli-mednli	0.781	0.705	0.741

Table 2: Comparison of MedNLI fine-tuned models on SCIFACT dev set

We also experimented in a zero-shot setting for the SCIFACT pipeline where we first retrieve relevant PubMed abstracts for the claims in the dev set using the corpus provided in the dataset (Deka et al., 2022b). After that, the top n evidence sentences are extracted from the abstracts (Deka et al., 2022a) and then we use the claim-evidence pairs to predict whether the evidence supports or refutes the claim. We compared our model with state-of-the-art zero-shot as well as few-shot baselines evaluated on the SCIFACT dataset. However, it should be noted that the baselines have different trade-off points in calculating the results due to our method being different from theirs. The results for the experiment are shown below in Table 3 where top n sentences are the evidence sentences from the relevant abstracts For each setting, we retrieve the top 2, 3, 5 and 10 evidence sentences and then the label for claim-evidence pair is predicted.

The baselines use a supervised approach where they use the train set of the SCIFACT dataset. In our method, however, we are using a transfer learning approach where we directly use our method over the dev set without using the train set. From Table 3, we can see that we have outperformed the baseline models in the zero-shot setting. We can also see that our best-performing model setting outperforms even the few-shot baselines as well as the fully fine-tuned VERISCI (Wadden et al., 2020) baseline.

5.3 Prediction of the claims

In order to use the D-S theory for our work in resolving conflicting evidences, we first need to calculate the probabilities of the classes. As we have approached our task as an NLI problem, we have three different classes: SUPPORT, REFUTE and NEUTRAL. For each claim-evidence pair, our

Model	**Top n sentences**	**P (sentence + label)**	**R (sentence + label)**	**F-1 (sentence + label)**
Our approach (top 2 abstracts)	2	0.515	0.393	0.446
	3	0.509	0.409	0.454
	5	0.503	0.420	0.458
	10	0.513	0.444	0.476
Our approach (top 3 abstracts)	2	0.505	0.390	0.440
	3	0.492	0.401	0.441
	5	0.505	0.428	0.463
	10	**0.534**	**0.465**	**0.497**
Our approach (top 5 abstracts)	2	0.510	0.389	0.441
	3	0.500	0.403	0.446
	5	0.510	0.422	0.462
	10	0.528	0.459	0.491
Our approach (top 10 abstracts)	2	0.500	0.375	0.428
	3	0.472	0.384	0.424
	5	0.512	0.419	0.461
	10	0.526	0.449	0.484
VERISCI (zero-shot)		0.248	0.334	0.284
VERISCI		0.469	0.392	0.426
MULTIVERS (zero-shot)		0.390	0.216	0.278
MULTIVERS (few-shot)		0.517	0.403	0.453
ParagraphJoint (Zero-Shot)		0.364	0.149	0.211
ParagraphJoint (Few-Shot)		0.330	0.351	0340

Table 3: Zero-shot evaluation on SCIFACT dev set for the whole pipeline process

classifier calculates the probabilities of the three classes. We are using these probabilities as the BPAs from Equation 2. Once we have the BPAs, we then use equation 9 to calculate the final modified BPAs to mitigate the denominator error. Once we have calculated the modified BPAs, we use equations 5 and 6 to combine the BPAs according to the D-S combination rules.

The SCIFACT dataset does not have labels that can be used for the evaluation of the combination method. In order to infer these labels for the evaluation, we give the final class label for a claim as either “Fake”, “Truth” or “Neutral”. This label is based on the gold standard label of the evidences. We have seen that in the SCIFACT dataset, all evidences for a claim can either be “SUPPORT” or “REFUTE”. Based on this, we label claims as “True” which has evidences labelled as “SUPPORT” and “False” for claims that have evidence labels as “REFUTE”. Some of the evidences do not enough information to either “SUPPORT” or “REFUTE” claims. These are labelled as “Neutral”. This will be our gold standard and the results from the D-S combination theory will be evaluated against this gold standard. As evaluation metrics, we have used macro precision, recall and f-1 score. We experimented in two different scenarios. Initially, we experimented directly on the dev set using our model from Table 2. We got the results as follows: Precision = 0.898, Recall = 0.893, F-1 score = 0.894.

For the next experiment, we have used the whole pipeline process where we first retrieve top n abstracts and then from these abstracts we retrieve the top n evidences. Once we have the evidence sentences, we then use the classifier to classify them accordingly. The results are shown in Table 4.

Top n abstracts	**Top n sentences**	**P**	**R**	**F-1**
2	**2**	0.867	0.595	0.705
	3	0.887	0.663	0.738
	5	0.884	0.649	0.747
	10	0.894	0.680	0.772
3	**2**	0.865	0.585	0.697
	3	0.885	0.617	0.726
	5	0.891	0.654	0.753
	10	0.897	0.702	0.786
5	**2**	0.866	0.590	0.702
	3	0.874	0.627	0.729
	5	0.885	0.659	0.755
	10	0.892	0.707	0.788
10	**2**	0.864	0.579	0.693
	3	0.869	0.601	0.709
	5	0.886	0.665	0.758
	10	0.891	0.702	0.784

Table 4: D-S method evaluation on SCIFACT dev set for the pipeline process

6 Supervised approach using augmented data

In situations where we have labelled data, our model can be used to train over such data in a supervised way which means that the knowledge from our model can be transferred over such data. However, a problem with the available datasets such as SCIFACT, is the fact that it has very less labelled data for training which may not lead to improved performance of the model. To improve it, there should be more data for training, and data augmentation is one way of increasing the number of training examples (Shorten et al., 2021; Feng et al., 2021). There are various ways of augmenting data such as rule-based, interpolation-based and model-based (Shi et al., 2022). In rule-based methods, words and phrases are manipulated in order to generate augmented text. But a problem with such methods is that changing words or phrases may lead to change in the meaning of the sentences (Niu and Bansal, 2018). In the context of biomedical text, if the meaning of the sentence changes then the sampled augmented data may lead to negative performance in model training. By performing interpolation operations directly on the source text (Chawla et al., 2002; He et al., 2008) or latent space representations (Chen et al., 2020), interpolation-based approaches produce new instances. However, such methods can be error-prone due to noisy generated data (Chawla et al., 2002). Model-based methods use language models such as BERT to generate new training examples. One popular way of using these models to generate new training examples is back translation (Edunov et al., 2018). Recent research works have explored these language models for data augmentation via back translation (Melton et al., 2022).

For our work, we explored the following research questions:

RQ 1. Can we use back-translation method for data augmentation on domain specific fact checking task without loss of context?

RQ 2. How well does the model fare without using augmented examples vs the model which uses augmented examples?

In order to answer the research questions, we explored two different ways: using Google Translate and transformer-based language models. Using Google Translate for back translation has been studied in previous research (Pappas et al., 2022). We have used Google Translate to convert the claim-evidence pairs to different languages such as German, French, Russian, Chinese and Spanish. Each language has a different language structure and since biomedical text is different than general text, a comparison of all the different languages would show which languages can be better suited for such tasks in the medical domain. We have used the deep translator python API [3] for the Google Translate method.

Transformer-based language models have been proven to be very good in neural machine translation tasks (Przystupa and Abdul-Mageed, 2019; Uhrig et al., 2021). For the study, we have used the OpusMT (Tiedemann and Thottingal, 2020) models which are pretrained transformer models for the neural machine translation task based on the Marian MT framework (Junczys-Dowmunt et al., 2018). We have used the models from the HuggingFace repository for the OpusMT [4] models.

For the experiment using the Google translator API, we translate all the claims as a batch to different languages and then back to English and the same approach is taken for the evidences as well. We then merge the synthetic data with the original data by removing any duplicates. However, for NMT models, all claims and evidence are back-translated one at a time using the HuggingFace pipeline[5]. Once we get the back-translated examples, we then merge them with the original data. For evaluation, we use our model from Table 3 with the best results and train it over the augmented data. The results are shown below in Table 5.

Methods/Models	**P**	**R**	**F1**
OpusMT-German-English	0.5992	0.5587	0.5783
OpusMT-Spanish-English	0.5638	0.5219	0.5420
OpusMT-Chinese-English	0.5899	0.5467	0.5675
OpusMT-Russian-English	0.5753	0.5347	0.5543
OpusMT-French-English	0.5859	0.5454	0.5649
Googletranslate(German)	0.5780	0.5348	0.5555
Googletranslate(Spanish)	0.5620	0.5215	0.5410
Googletranslate(Chinese)	0.5630	0.5224	0.5419
Googletranslate(French)	0.5762	0.5370	0.5559
Googletranslate(Russian)	0.5576	0.5184	0.5373
Without augmentation(fine tuned)	0.5443	0.5051	0.5239
Without finetuning on train set	0.5340	0.4651	0.4970

Table 5: Data augmentation results

As seen from the table above the model without fine-tuning on train set performs poor which is expected. Training the model with the train set but without augmentation results in slight improvement on the results. However, we can see there is a significant improvement in the results once we augment the train file using the back translation approaches. Out of the two different approaches that we have experimented with, the transformer-based NMT models perform better than Google translate API. However, these models are also time consuming while performing the back-translation task unlike the Google translate approach. The results show that data augmentation using back translation gives us better results for such domain-specific fact checking tasks which answer both **RQ1** and **RQ2**.

7 Transferring over other datasets

In order to know how well our model generalizes over other similar data, we experimented with two similar fact-checking datasets on biomedical data, HEALTHVER (Sarrouti et al., 2021) and COVID-FACT (Saakyan et al., 2021). Both datasets focus on Covid-19 data, however, the way claims and evidences are collected in both these datasets differ. HEALTHVER claims are collected from CORD-19 (Wang et al., 2020) corpus article snip-

[3] https://deep-translator.readthedocs.io/en/latest/

[4] https://huggingface.co/Helsinki-NLP

[5] https://huggingface.co/docs/transformers/main_classes/pipelines

pets which were retrieved to answer questions for TREC-COVID (Voorhees et al., 2021). The claims in HEALTHVER are complex and evidences are provided for each claim. The dataset has three labels, SUPPORT, REFUTE and NEUTRAL based on the evidences which are collected from the article snippets itself. COVIDFACT, on the other hand, has claims collected from Covid-19 subreddit and evidences are collected from linked scientific papers and documents collected from Google search. The claims in COVIDFACT are also complex and has two labels for the evidences collected, SUPPORT and REFUTE. Both HEALTHVER and COVIDFACT have one annotated evidence per claim, whereas in SCIFACT, there may be more than one evidence for one claim. Also, in SCIFACT, relevant abstracts are needed to be retrieved first from the corpus provided and then evidence sentences are needed to be retrieved from those abstracts.

Although our model has not been trained over Covid-19 specific text, we wanted to experiment how well it generalizes over such data by performing two different experiments. In the first experiment, we applied our model to the test set of both HEALTHVER and COVIDFACT without using the training set in a zero-shot approach. For SCIFACT, we have used the dev set.

	PubMedBERT-mnli-mednli		
	P	**R**	**F1**
SCIFACT	0.847	0.753	0.797
HEALTHVER	0.429	0.431	0.354
COVIDFACT	0.425	0.401	0.338

Table 6: Zero-shot comparison of our model with different datasets

We can see from Table 6 that in a zero-shot setting, our model performs better with the SCIFACT dataset. This can be attributed to the fact that SCIFACT data contain PubMed abstracts and PubMedBERT (Gu et al., 2021) has been trained over PubMed text which is why it performs better. HEALTHVER and COVIDFACT, on the other hand do not contain PubMed data and as such the model does not generalise well over the other datasets.

For the second experiment, we have transferred a trained model over one dataset to the other two to see how well models trained on one dataset generalize to other datasets. We use the train sets of the datasets to train the model and then use the test sets of the other datasets to evaluate the model performance. Since the HEALTHVER dataset has the NEUTRAL label, we have dropped instances from its test set having that label in order to maintain consistency over all the datasets when the model was trained over SCIFACT and COVIDFACT train sets since these datasets only have SUPPORT and REFUTE labels. The results of the experiment are shown in Table 7.

PubMedBERT-mnli-mednli (trained on HEALTHVER)	**HEALTHVER**	**SCIFACT**	**COVIDFACT**
P	0.6287	0.5197	0.4283
R	0.5780	0.5242	0.3918
F1	0.5040	0.5215	0.3349
PubMedBERT-mnli-mednli (trained on SCIFACT)	**(NEUTRAL instances are dropped)**		
P	0.6827	0.8730	0.6347
R	0.6272	0.8497	0.6143
F1	0.6251	0.8591	0.5082
Pubmedbert-mnli-mednli (trained on COVIDFACT)	**(NEUTRAL instances are dropped)**		
P	0.6352	0.7133	0.6851
R	0.6417	0.7313	0.6933
F1	0.6245	0.7009	0.6884
PubMedBERT-mnli-mednli (trained on HEALTHVER without NEUTRAL instances)			
P	0.8080	0.8636	0.6739
R	0.7347	0.7603	0.6194
F1	0.7462	0.7811	0.4836

Table 7: Transfer learning comparison of our approach on different datasets

From Table 7, it can be seen that when the model is trained on SCIFACT and HEALTHVER, transferring to the COVIDFACT test set does not give very good results. This is due to the fact that COVIDFACT contains both scientific as well as non-scientific claim-evidence pairs and therefore a model trained on either SCIFACT or HEALTHVER does not generalize well as they are based on scientific data. We can also see that the model trained on HEALTHVER generalizes better on the SCIFACT data and vice-versa as both these datasets are based on scientific claims and evidences, they learn better and generalize well on each other. However, we can also see that the model trained on COVIDFACT generalizes well on the other datasets since it contains both scientific and non-scientific data. These results confirm the findings by (Saakyan et al., 2021) that models trained on scientific data do not generalize well on data that contain non-scientific data as well. This is important as real-world health misinformation data may contain both scientific and non-scientific claims. In such situations, we need to have both scientific as well as non-scientific data so that models can learn to generalize on such data.

8 Conclusion and future work

We have explored the prediction of veracity for health-related fact-checking tasks that can be learned from NLI data. By doing experiments, we showed that training domain-specific BERT-based models on domain-specific NLI data improves the model performance for fact-checking task. We also explored a method that can be used to combine different evidences for a claim, even for situations that have conflicting evidences for the same claim. We have also shown by experiments that augmenting data using back-translation helps in situations where there is a lack of training data. Although fact-checking of scientific claims is still a new task, there is a potential for improvement of the current methods being used for the task. With the advent of more capable large language models, new research direction such as prompt based methods can also be explored. As future work, we are interested in exploring such prompt-based approaches along with multimodal data in this space.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. *arXiv preprint arXiv:2004.12239*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Pritam Deka, Anna Jurek-Loughrey, and Deepak P. 2022a. Evidence extraction to validate medical claims in fake news detection. In *Health Information Science: 11th International Conference, HIS 2022, Virtual Event, October 28–30, 2022, Proceedings*, pages 3–15. Springer.

Pritam Deka, Anna Jurek-Loughrey, and Deepak P. 2022b. Improved methods to aid unsupervised evidence-based fact checking for online health news. *Journal of Data Intelligence*, 3(4):474–504.

Arthur P Dempster et al. 2008. Upper and lower probabilities induced by a multivalued mapping. *Classic works of the Dempster-Shafer theory of belief functions*, 219(2):57–72.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE.

Wen Jiang, Miaoyan Zhuang, Xiyun Qin, and Yongchuan Tang. 2016. Conflicting evidence combination based on uncertainty measure and distance of evidence. *SpringerPlus*, 5(1):1–11.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*.

Md Nazmuzzaman Khan and Sohel Anwar. 2019. Time-domain data fusion using weighted evidence and dempster–shafer combination rule: Application in object classification. *Sensors*, 19(23):5187.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Xiangci Li, Gully A Burns, and Nanyun Peng. 2021. A paragraph-level multi-task learning model for scientific fact-verification. In *SDU@ AAAI*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Chad A Melton, Brianna M White, Robert L Davis, Robert A Bednarczyk, and Arash Shaban-Nejad. 2022. Fine-tuned sentiment analysis of covid-19 vaccine–related social media data: Comparative study. *Journal of Medical Internet Research*, 24(10):e40408.

Tong Niu and Mohit Bansal. 2018. Adversarial oversensitivity and over-stability strategies for dialogue models. *arXiv preprint arXiv:1809.02079*.

Dimitris Pappas, Prodromos Malakasiotis, and Ion Androutsopoulos. 2022. Data augmentation for biomedical factoid question answering. *arXiv preprint arXiv:2204.04711*.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2020. Scientific claim verification with vert5erini. *arXiv preprint arXiv:2010.11930*.

Michael Przystupa and Muhammad Abdul-Mageed. 2019. Neural machine translation of low-resource and similar languages with backtranslation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 224–235.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic. *arXiv preprint arXiv:2106.03794*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Mourad Sarrouti, Asma Ben Abacha, Yassine M'rabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512.

Glenn Shafer. 1976. *A mathematical theory of evidence*, volume 42. Princeton university press.

Yiwen Shi, Taha ValizadehAslani, Jing Wang, Ping Ren, Yi Zhang, Meng Hu, Liang Zhao, and Hualou Liang. 2022. Improving imbalanced learning by pre-finetuning with data augmentation. In *Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications*, pages 68–82. PMLR.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8:1–34.

Philippe Smets. 2000. Data fusion in the transferable belief model. In *Proceedings of the third international conference on information fusion*, volume 1, pages PS21–PS33. IEEE.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Jörg Tiedemann and Santhosh Thottingal. 2020. Opus-mt–building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

Sarah Uhrig, Yoalli Rezepka Garcia, Juri Opitz, and Anette Frank. 2021. Translate, then parse! a strong baseline for cross-lingual amr parsing. *arXiv preprint arXiv:2106.04565*.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, pages 1–12. ACM New York, NY, USA.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

David Wadden, Kyle Lo, Lucy Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. Multivers: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76.

Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, et al. 2018. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. *arXiv preprint arXiv:1812.10860*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020. Cord-19: The covid-19 open research dataset. *ArXiv*.

Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019a. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 950–958.

Yunjuan Wang, Kezhen Zhang, and Yong Deng. 2019b. Base belief function: an efficient method of conflict management. *Journal of Ambient Intelligence and Humanized Computing*, 10:3427–3437.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Fuyuan Xiao. 2019. Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy. *Information Fusion*, 46:23–32.

Xia Zeng, Amani S Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Language and Linguistics Compass*, 15(10):e12438.

Zhiwei Zhang, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. 2021. Abstract, rationale, stance: a joint model for scientific claim verification. *arXiv preprint arXiv:2110.15116*.