PHEMEPlus: Enriching Social Media Rumour Verification with External Evidence

[1]**John Dougrez-Lewis,** [2,3]**Elena Kochkina,** [1,3]**Miguel Arana-Catania,** [1,2,3]**Maria Liakata,** [1,3]**Yulan He**

[1]Department of Computer Science, University of Warwick, UK
[2]Queen-Mary University of London, UK
[3]The Alan Turing Institute, UK

{j.Dougrez-Lewis,miguel.arana-catania,yulan.he}@warwick.ac.uk
{m.liakata,e.kochkina}@qmul.ac.uk

Abstract

Work on social media rumour verification utilises signals from posts, their propagation and users involved. Other lines of work target identifying and fact-checking claims based on information from Wikipedia, or trustworthy news articles without considering social media context. However works combining the information from social media with external evidence from the wider web are lacking. To facilitate research in this direction, we release a novel dataset, PHEMEPlus[1], an extension of the PHEME benchmark, which contains social media conversations as well as relevant external evidence for each rumour. We demonstrate the effectiveness of incorporating such evidence in improving rumour verification models. Additionally, as part of the evidence collection, we evaluate various ways of query formulation to identify the most effective method.

1 Introduction

The harm and prevalence of online misinformation made research into automated methods of information verification an important and active research area. This includes various tasks like fact-checking, social media rumour detection, stance classification and verification. In this work we are concerned with social media rumour verification, the task of identifying whether a rumour (i.e check-worthy claim circulating on social media whose veracity status is yet to be verified (Zubiaga et al., 2018)), is *True*, *False* or *Unverified*.

Although a significant amount of work has been done towards evaluating the veracity of social media rumours (Zubiaga et al., 2016; Ma et al., 2017; Song et al., 2018; Dougrez-Lewis et al., 2021), there is still a dearth of works and datasets combining the information from social media with external evidence from the wider web. While recent works focusing on rumours around the COVID-19 pandemic have been collecting data from a wide range of sources from news and social media to scientific publications (Cui and Lee, 2020; Zhou et al., 2020; Wang et al., 2020), these are not sufficient for the creation of generalisable verification models as they only focus on a single topic. At the same time works on fact-checking, which do not focus on social media content, but use claims from debunking websites (Lim et al., 2019; Ahmadi et al., 2019), as well as recent work by Li et al. (2021) have shown the benefits of utilising stance of evidence for verification.

Here we aim to further enable research in this direction and release an enriched version of a popular benchmark dataset PHEME (Zubiaga et al., 2016) with timely evidence for each of the rumours, obtained from a wide range of web sources.

Although a few works use web search for evidence retrieval (Popat et al., 2018; Lim et al., 2019), to our knowledge, only the work of Lim et al. (2017) touches upon the topic of the search query formulation. Here we analyse several query formulation strategies to find the most effective one.

In this work we make the following contributions:

- We collect and release the PHEMEPlus dataset of Twitter rumour conversations with the relevant heterogeneous evidence retrieved from the web to facilitate research on combining multiple sources of information for social media rumour verification.

- We investigate approaches towards search query formulation for evidence retrieval, together with evaluation metrics for the quality of evidence retrieved.

- We demonstrate the effectiveness of incorporating external evidence into rumour veracity classification models.

---

[1]https://github.com/JohnNLP/PhemePlus

2 Related work

2.1 Existing Veracity Classification Datasets

Among existing datasets for veracity classification we can broadly discern two categories: (1) focusing on claims arising from social media in the form of posts (Zubiaga et al., 2016; Ma et al., 2017) and (2) focusing on manually formulated claims, either created specifically for a task (Thorne et al., 2018), or consisting of titles from news or debunking websites (Wang, 2017; Alhindi et al., 2018; Lim et al., 2019; Ahmadi et al., 2019). These different types of claims present different challenges for verification models and evidence retrieval systems. In particular social media posts often use non-standard grammar, hashtags and have typos (intentional or otherwise). It can be crucial to process claims directly from social media to enable early-stage misinformation detection as rumours often start spreading on social media, later making it into the mainstream media. Only a few datasets incorporate both social media and evidence from the web, however these often focus on a very limited number of sources of evidence or a single topic (Dai et al., 2020; Cui and Lee, 2020). One of such datasets is FakeNewsNet (Shu et al., 2018) incorporating *fake* and *true* news articles from fact-checking websites PolitiFact[2] and GossipCop[3]. Articles are further augmented with users' posts on Twitter pertaining to them but not including full conversation structure. FakeHealth (Dai et al., 2020) is a similarly constructed dataset based on health-related news articles labelled by the Health News Review[4], including Twitter users' replies and profiles. Barrón-Cedeno et al. (2020) organised shared tasks for automatic identification and verification of claims in social media. Apart from tasks on check-worthiness estimation for tweets and verified claim retrieval, they also released tasks for supporting evidence retrieval and claim verification. However, the tasks mainly focused on misinformation about COVID-19 and the latter tasks were only offered in Arabic.

In light of the wave of misinformation associated with COVID-19 pandemic researchers have been collecting relevant datasets of scientific publications, news articles and their headlines, social media posts and claims about COVID-19 (Shaar et al., 2020; Dharawat et al., 2020; Zhou et al., 2020; Li et al., 2020; Memon and Carley, 2020; Hossain et al., 2020; Barrón-Cedeno et al., 2020). One of the most relevant work to ours is COAID (Cui and Lee, 2020), a large-scale dataset containing COVID-19 related news articles as well as social media posts. While these are rich resources, which enable further research against misinformation, they are insufficient for training generalisable models as they solely focus on one topic.

In this work we have augmented the PHEME dataset, a popular benchmark dataset for social media rumour verification, it contains rumours expressed via Twitter posts with full conversation threads from several news-breaking events on different topics. This dataset is set up to imitate realistic scenarios as (1) it was collected as the events were unfolding and then rumour stories were identified and annotated by a professional journalist as opposed to collecting tweets based on existing fact-checks as in Ma et al. (2017); and (2) the evaluation is performed in on events unseen during training. We augment it with evidence articles from across the web to give it access to an unlimited set of resources. To preserve the realistic scenario of verifying emerging rumours, all of our evidence is restricted to articles indexed by Google no later than the day on which the rumour was posted to Twitter.

2.2 Social Media Rumour Verification Models Using External Information

Social media rumour verification models use various types of information available on social media platform: text of rumourous posts and responses (Dougrez-Lewis et al., 2021), user information and connections (Khoo et al., 2020), propagation patterns (Ma et al., 2018). However, still only few works incorporate external evidence.

Lim et al. (2017) proposed the iFACT framework that extracts claims from tweets pertaining to major events. For each claim, it collects evidence from web search and estimates the likelihood of a claim being credible. To formulate the search query iFACT uses ClausIE (Del Corro and Gemulla, 2013) to extract (*subject, predicate, object*) triples from tweets. To determine the credibility of the claim iFACT uses features extracted from search results and dependencies between claims. Here we also experiment with using ClausIE to formulate the search query.

---

[2]https://www.politifact.com/
[3]https://www.suggest.com/
[4]https://www.healthnewsreview.org/

Events	Threads	True	False	Unverified	Relevant Articles
Charlie Hebdo	458	193	116	149	3941
Sydney Siege	522	382	86	54	4436
Ferguson	284	10	8	266	2473
Ottawa Shooting	470	329	72	69	4020
Germanwings Crash	238	94	111	33	2057
Total Threads	1972	1008	393	571	16927

Table 1: Statistics of the PHEMEPlus dataset by extending the PHEME-5 dataset with retrieved relevant articles. All but 2 rumours have at least 1 associated article.
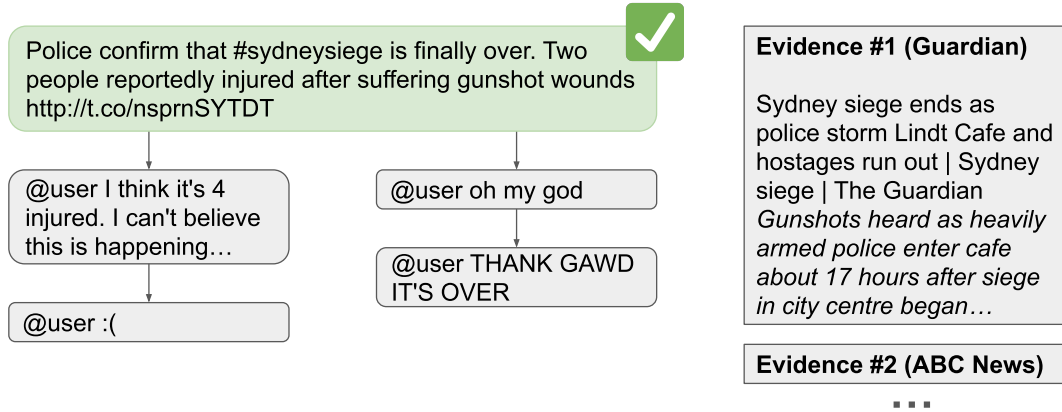

Figure 1: The PHEMEPlus dataset consists of labelled Twitter rumours, their conversation thread, and corresponding evidence retrieved from the web. This is an adapted example.

Li et al. (2021) propose to improve rumour detection on PHEME dataset by using evidence from Wikipedia. They first train the evidence extraction module on the FEVER dataset and then use it as part of a rumour detection system to get relevant sentences from a Wikipedia dump along with Twitter conversation around a rumour. While being limited by a single source of information, they demonstrate performance improvements over previous models not using external information.

In this work we use BERT-based models as strong baselines to demonstrate the effectiveness of incorporating the evidence for social media rumour verification. In future work we will be experimenting with various ways of incorporating it to maximise the benefits.

3 Augmenting PHEME dataset with External Evidence

3.1 Base dataset

We chose to extend the PHEME-5 dataset (Zubiaga et al., 2016), which consists of Twitter conversations discussing rumours around five real-world events including the Lindt Cafe siege in Sydney and the 2015 Charlie Hebdo terrorist attack. This dataset is a popular benchmark for rumour verification, it is particularly challenging due to class imbalance and evaluation using leave-one-event-out cross-validation, reflecting a real-world evaluation scenario. Table 1 shows the statistics of the PHEMEPlus dataset by extending the original PHEME-5 dataset with retrieved relevant articles. The first four columns show the number of conversation threads in each of the event and each of the classes in the orignal PHEME-5 dataset. Figure 1 shows an example entry in the PHEMEPlus dataset, comprised of a rumorous tweet, veracity label, its conversation thread, and relevant evidence retrieved from the web. It is notable that tweets in the conversation thread (and the rumour itself) often contain URLs provided by users which may be useful as a further source of evidence, and that the corresponding evidence is not a part of the original PHEME dataset. Kochkina (2019) has shown that *True* rumours in PHEME have a higher percentage of URLs attached (55%) than for *False* (48%)

and *Unverified* (48%) rumours. For the portion of PHEME with comments annotated for stance, these supplementary URLs were overwhelmingly found in comments *supporting* the source tweet's claim (33%) as opposed to those, *denying* (8%), *querying* (6%), or *commenting* (9%) on it.

3.2 Evidence Retrieval through Web search

In order to obtain evidence from the unlimited number of sources we chose to use Web search for evidence retrieval. We choose Google Search as it is one of the most established search engines, and, importantly, allows us to filter results by date. This is crucial as rumours are often resolved and widely debunked in some time following their originating event and the rumourous post, but this information would not be available to the model in a real time evaluation scenario.

Furthermore, the evidence we retrieve from Google appears robustly reputable, with popular news sources consistently ranking highly in the search results. This is to be expected, since their PageRank system weights heavily websites which are highly cited/referenced by others. Web search results are also more likely to be up-to-date than any corresponding Wikipedia pages regarding a current real world happening, which may not be updated nor appropriately checked for correctness.

For every search we include the term (before: date) at the start of the query to restrict results to articles from before the date the rumourous tweet was posted. For each query we collect the top 5 non-empty results from the web search.

While Google search is able to process various types of queries, from keywords to natural language utterances, we performed a set of experiments to identify the most suitable method of query formulation for our particular task of evidence retrieval for rumours conveyed in Twitter posts. We experiment with queries formulated as (1) natural language sentence, (2) keywords, and (3) *(subject, object, predicate)* triples. For each experiment, we include around 99% of the PHEME dataset since a few queries did not yield enough non-empty results. Although we are aware of some more advanced studies into query expansion and formulation (Tamannaee et al., 2020; Scells et al., 2020), contributing to these fields is beyond the scope of this paper. Here we aim to demonstrate gains from relatively simple approaches described below towards evidence retrieval.

3.2.1 Search Strategies

We experiment with the following search strategies:

Preprocessed The search query is the source rumour, obtained from the preprocessed tweet. Our preprocessing entails removing URLs, replacing user mentions with "user" (so as to retain lexical structure), removing hashtags from the end but not the middle (also for lexical structure) and segmenting any compound hashtags. URLs are saved aside since they may have future use as evidence. Hashtags at the end of the tweet (but not others) are also retained, placed in brackets for an "OR" search with the rest of the query. These hashtags in particular are expected to be highly telling of the topic/theme of the tweet, especially when it is otherwise lacking in contextual words.

Shortening with StanfordNLP We use Stanza (Qi et al., 2020) to parse preprocessed tweets. Having obtained a parse tree, words in the following constructs are retained in-place: {*obl:npmod, compound, advcl, nummod, acl:relcl, nsubj:pass, acl, amod, aux:pass*}. This combination of constructs was iteratively finetuned until the resultant queries felt similar to the author's own search style, the idea being to replicate the search strategy of an experienced user. Hashtags at the end of tweets are handled as before.

Shortening with ClausIE We use ClausIE (Del Corro and Gemulla, 2013), a popular subject-relation-object extraction system in the same manner to find (*subject, predicate, object*) triples. These are kept in-place whilst the other words are removed. Hashtags at the end of tweets are retained as before.

Examples of the search queries formed can be found in Table 2.

3.2.2 Evaluation metrics

We devise evaluation metrics to compare the quality of evidence retrieved using different query types, without the need for a rumour verification model in advance.

URL Words Metric URLs frequently contain English words which are representative of the content on their webpage, which we can treat as gold-standard keywords as in (Ma et al., 2016). To get a goodness score in the range [0,1] we compute the cosine similarity between the words in URLs of retrieved articles and those posted in re-

Original Rumour:	*MORE: Massacre suspects believed to have taken hostage and holed up in small industrial town northeast of Paris: \<url\> #CharlieHebdo*

Query Strategy	Query Text
Preprocessed	before:2015-01-09 MORE : Massacre suspects believed to have taken hostage and holed up in small industrial town northeast of Paris :
StanfordNLP	before:2015-01-09 (Charlie Hebdo) Massacre suspects small industrial town northeast
ClausIE	before:2015-01-09 (Charlie Hebdo) Massacre suspects believed to have taken hostage holed up in small industrial town northeast of Paris

Table 2: Examples of search queries generated by the various search strategies, given the original rumour. In this case, the ClausIE strategy only removes the words "MORE" and "and".

sponse to the rumour. Specifically, for each retrieved article, its URL-words are compared with those of each URL in the Twitter comments. The final score is the average of all such cosine similarities across all retrieved articles in the dataset, encoded by Word2Vec (Mikolov et al., 2013).

GloVe Metric If an article is relevant to a rumour, they will be similar in content. We use GloVe (Pennington et al., 2014) to calculate the similarity between the first 3 paragraphs of an article and the source rumour, with the title also counting as a paragraph. We use only the first few paragraphs because they seem likely to contain the highest density of relevant information. Cosine similarity scores are calculated between each of these paragraphs and the source rumour, and are averaged to give the article a similarity score. Unknown words with zero vectors are ignored for this purpose, although there is a weakness that some of the most important event-specific words could be unknown.

BERTScore Metric This is calculated similarly to the GloVe metric, except that BERTScore (Zhang et al., 2020) is used in its place.

3.2.3 Evaluating Retrieval Results

Table 3 displays the performance of our search strategies when evaluated via the URL Words, GloVe, and BERT evaluation metrics. These results suggest that searching for the preprocessed tweet may be the best way to get relevant background information from the web, as opposed to extracting keywords from the tweet. This narrowly surpasses the performance of our ClausIE-based search strategy, which outperforms the StanfordNLP approach. The ClausIE strategy may retain a higher proportion of key grammatical constructs than the lat-

ter, which play an unexpectedly important role in Google's search algorithm. This is contrary to the authors' searching intuition, perhaps due to their recent integration of models such as BERT (Devlin et al., 2018).

Metric	Preprocessed	StanfordNLP	ClausIE
URL Words	**0.802**	0.777	0.795
GloVe	**0.661**	0.651	0.660
BERTScore	**0.826**	0.825	0.825

Table 3: Performance of the search strategies, evaluated by our evaluation metrics.

Although some of the values in Table 3 appear close together, it is notable that the results of the different query formulations land in the same order irrespective of the scoring metric used. Furthermore, the score differences between different query formulations become more substantial when taking into account their weak upper and lower bounds derived from using artificially generated 'target article' and 'random' queries (data not shown).

3.3 PHEMEPlus dataset

An example entry of the PHEMEPlus dataset can be found in Figure 1. The number of articles we retrieved using the *Preprocessed* method can be found in Table 1. All but two of the rumours have at least one associated evidence article, up to a maximum of 10.

We explore the overlap between the evidence in our resultant PHEMEPlus dataset and the URLs in the Twitter comments responding to the rumours. Table 4 shows the overlap between the articles retrieved from web search (using the Preprocessed Only strategy) and those from the Twitter comments. We observe little overlap between articles retrieved from web search and articles retrieved

	Overall pages	Unique pages
From web search	13255 (12008 not-empty)	3817 (3425 not-empty)
From rumour responses	2160 (1658 not-empty)	601 (457 not-empty)
Overlap	100	102

Table 4: Overlap of retrieved articles with articles from rumour responses.

from comments responding to rumours. The latter may thus be a substantially different, potentially less useful, source of evidence due to a high density of social media pages and the likelihood that some of the comments may not be directly responding to the source rumour.

A relatively large proportion of the articles retrieved from responses are deemed "empty", meaning they either have no body-text and/or no title. From this, and manual inspection, we infer that response-URLs are more likely to be social media posts or videos which are prone to missing titles or first paragraphs.

The overall:unique ratio being similar for both thread and web suggests that the Google results are indeed sensitive to the content of each thread, as opposed to repeatedly giving the same results for a given rumourous event. There is not much overlap between the search results and the Twitter thread, and a large proportion of existing overlap might be explainable by news websites tweeting their news URLs. This is not attributable to overly stringent overlap criteria as the discrepancy between the overall number of articles and the number of articles without duplicates acts as a positive control to this end.

Similar links nearly always result from the same thread, possibly due to the aforementioned news companies. Investigating further, the vast majority (if not all) of the overlap was news articles. Speculatively, it is plausible that most of this overlap came from news websites tweeting their stories, as there are some examples of this in the dataset.

4 Evaluating the Effectiveness of Evidence for Rumour Verification

We conduct experiments to evaluate the effectiveness of our retrieved evidence for Twitter rumour veracity classification.

4.1 Evidence Sentence Retrieval

In our PHEMEPlus dataset, each source tweet is paired with up to 10 most relevant retrieved articles. We follow the typical pipeline fact check-

ing approach to further select the 5 most relevant sentences from the articles associated with each source tweet. In order to do this, we use a simple novel approach based on ClausIE (Del Corro and Gemulla, 2013). The idea is to be able to reliably find relevant sentences whilst not being clobbered by the inevitably rare rumour-specific vocabulary which may not be recognised by many approaches. First, we use ClausIE to extract all relevant subject-predicate-object triples from the retrieved information. We assume these to be the words with the most potential for true relevance to the tweet. Any stop-words contained within are filtered out. For each sentence, a score is assigned based on how many of these important words are also contained in the tweet, penalising both overly long (>20 token) and short (<5 token) sentences as are likely to be either uninformative or unconcise and work poorly with the BERT models. In particular, short sentences are ignored, whereas long sentences lose 2% of their score for each additional word. Only rumours with enough evidence to extract 5 sentences as above are used (99% of them) in our experiments. The top 5 such sentences are paired with each source tweet and are fed into a rumour classification model for veracity assessment.

4.2 Veracity Classification Models

We compare the performance of several veracity classification models in three input scenarios: (1) rumour (i.e., source tweet) alone, (2) evidence (i.e., extracted sentences) alone and (3) rumour concatenated with the evidence (extracted sentences). The classification models chosen include pre-trained language models such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), and a model making use of natural language inference results between a source rumour and its related evidence sentence.

BERT-based approaches We train BERT-based models including BERT and RoBERTa followed by a single softmax layer for rumour verification. Each pair of a rumour and a piece of relevant evidence sentence is concatenated as input to the model. The

BERT	Ch	Fe	Ge	Ot	Sy	False	True	Unv	MacroF1
Rumour + Ev.	**0.317**	**0.174**	0.213	**0.406**	0.318	**0.221**	0.549	**0.265**	**0.345**
Rumour	0.306	0.134	**0.315**	0.345	**0.320**	0.209	0.562	0.242	0.338
Evidence	0.268	0.045	0.264	0.370	0.307	0.140	**0.645**	0.099	0.295
RoBERTa									
Rumour + Ev.	**0.306**	**0.183**	**0.383**	0.368	**0.347**	**0.384**	0.600	**0.279**	**0.421**
Rumour	0.290	0.113	0.260	**0.420**	0.309	0.211	0.549	0.232	0.331
Evidence	0.288	0.028	0.252	0.335	0.327	0.145	**0.611**	0.144	0.301
NLI-SAN									
Rumour + Ev.	**0.354**	**0.256**	**0.365**	**0.591**	**0.458**	0.186	0.480	0.250	**0.405**

Table 5: Per-event and per-fold F1 scores from the BERT, RoBERTa, and NLI-SAN models. The 2-letter column headings abbreviate the names of individual rumourous events in PHEME (as in Table 1).

final predictions were determined by majority voting. These particular models are chosen because flavours of BERT have previously achieved state-of-the-art results in many natural language processing tasks.

Self-Attention Network based on Natural Language Inference (NLI-SAN) This method uses not only the representation of rumour and evidence like the previous methods, but also the Natural Language Inference (NLI) relationship between them.

First each rumour is paired with each of the evidence sentences and is fed into the RoBERTa-large-MNLI[5] model to generate the NLI relation triplet representing the *contradiction*, *neutrality*, and *entailment* probabilities. The rumour-sentence pair is also fed into the RoBERTa-large[5] model to generate the contextual representation. Both outputs are then combined using a self-attention network in which the NLI relation triplet is used as the query, while the contextual representation is used as the key and value. Afterwards, all the outputs are concatenated into a single output that is passed through a Multi-Layer Perceptron (MLP) and a Softmax layer that generates the final veracity classification value.

Since this approach relies on the inference relationship between rumour and evidence, we will only compare it with the other models if both elements are available, and thus only one result is shown in Table 5.

4.3 Experimental Setup

Experiments were performed using 5-fold leave-one-out-cross-validation with each of PHEME's rumourous events being a fold, as is customary for

this dataset (see Section 3.1). We will release the code used to collect the evidence and to perform experiments on GitHub.

For the training of the aforementioned models, the inputs are padded and truncated to the longest sequence. Cross-entropy is used as the loss function. The optimizer used is AdamW (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.01. For the BERT-based models, the batch size is 20, the learning rate is 3×10^{-5}, and the training is performed for 25 epochs. For NLI-SAN, the size of the hidden layer is 50, the batch size is 30, the learning rate is 10^{-4}, and the training is performed for 200 epochs.

4.4 Results and Discussion

Table 5 presents the results of our experiments in terms of macro-averaged F1-score. Macro F1 score is a suitable metric to evaluate performance on this dataset due to class and fold size imbalance.

In these experiments it is not our goal to outperform state-of-the-art results on the PHEME dataset, but to demonstrate the effectiveness of incorporating the evidence for social media rumour verification. State-of-the-art results are obtained by more complex architectures, in which incorporating the evidence and evaluating its effects is a more challenging task. For instance, the VRoC model (Cheng et al., 2020) currently yields state-of-the-art F1 score of 0.484 on this task, it uses Variational Autoencoder for representation of the rumour as well as multitask learning set up incorporating four tasks.

The results in Table 5 suggest that there is indeed a benefit to using the evidence which we have retrieved for rumour veracity classification. This joint approach outperforms the other two, and the

use of the rumour alone generally outperforms the use of evidence alone, fitting with the idea that veracity can be classified to some extent by the writing style of the rumour alone.

In addition to the improvement in the results obtained by having evidence relevant to each rumour, our work opens the door to the use of more complex veracity classification models that consider additional attributes between both elements. The results obtained in the case of the NLI-SAN model show how this approach can be useful, obtaining better results than using the BERT model, although in this case inferior to the more simple use of RoBERTa.

A more detailed, per-class and per-fold, results breakdown for all of the models can be found in Table 5. For both BERT and RoBERTa, the combination of rumour together with evidence seems particularly useful for correct classification of the *False* class, with a mild gain also noted for *Unverified*. This could be the result of models inferring that there is disagreement between *False* rumours and their evidence, which would not be possible without the presence of both sources. It is noteworthy that existing rumour veracity classification models using the PHEME dataset have often found the False and Unverified classes to be problematic (Dougrez-Lewis et al., 2021). *True* class also benefits from incorporating evidence in RoBERTa model comparing to using rumour only. The results breakdown for the NLI-SAN model can also be found in Table 5, for which a similar pattern of per-class results can be observed. Most of the per-fold results for both BERT and RoBERTa also show the best performance when using a combination of rumour and evidence, only with exception of Germanwings Crash event (dominated by *False* class) for BERT and Ottawa shooting event (dominated by *True* class) for RoBERTa.

5 Conclusions and Future Work

After experimentation with various searching strategies for retrieving evidence from the web, we have constructed the PHEMEPlus dataset, which will facilitate further work on using evidence from wide range of sources for rumour veracity classification. The best such strategies, according to our evaluation metrics, are those which leave the grammatical structure of the claim relatively intact. There is much potential to improve existing rumour veracity classification systems by augmenting them with, or with a broader range, or better quality of evi-

dence. We plan to build upon these findings in the future, working on identifying ways of incorporating the evidence from heterogeneous sources into more complex rumour verification models to maximise the gains from this information and achieve state-of-the-art results.

References

Naser Ahmadi, Joohyung Lee, Paolo Papotti, and Mohammed Saeed. 2019. Explainable fact checking with probabilistic answer set programming.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: improving fact-checking by justification modeling. In *Proceedings of the first workshop on fact extraction and verification (FEVER)*, pages 85–90.

Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, and Fatima Haouari. 2020. Checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media. In *European Conference on Information Retrieval*, pages 499–507. Springer.

Mingxi Cheng, Shahin Nazarian, and Paul Bogdan. 2020. *VRoC: Variational Autoencoder-Aided Multi-Task Rumor Classifier Based on Text*, page 2892–2898. Association for Computing Machinery, New York, NY, USA.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.

Enyan Dai, Yiwei Sun, and Suhang Wang. 2020. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 853–862.

Luciano Del Corro and Rainer Gemulla. 2013. Clausie: Clause-based open information extraction. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 355–366, New York, NY, USA. Association for Computing Machinery.

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

Arkin Dharawat, Ismini Lourentzou, Alex Morales, and ChengXiang Zhai. 2020. Drink bleach or do what now? covid-hera: A dataset for risk-informed health decision making in the presence of covid19 misinformation. *arXiv preprint arXiv:2010.08743*.

John Dougrez-Lewis, Maria Liakata, Elena Kochkina, and Yulan He. 2021. Learning disentangled latent topics for twitter rumour veracity classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3902–3908.

Tamanna Hossain, Robert L Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. Covidlies: Detecting covid-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Ling Min Serena Khoo, Hai Leong Chieu, Zhong Qian, and Jing Jiang. 2020. Interpretable rumor detection in microblogs by attending to user interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8783–8790.

Elena Kochkina. 2019. Rumour stance and veracity classification in social media conversations,".

Jiawen Li, Shiwen Ni, and Hung-Yu Kao. 2021. Meet the truth: Leverage objective facts and subjective views for interpretable rumor detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 705–715.

Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation. *arXiv preprint arXiv:2011.04088*.

Wee Yong Lim, Mong Li Lee, and Wynne Hsu. 2017. Ifact: An interactive framework to assess claims from tweets. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 787–796, New York, NY, USA. Association for Computing Machinery.

Wee Yong Lim, Mong Li Lee, and Wynne Hsu. 2019. End-to-end time-sensitive fact check,". In *ACM SIGIR Workshop on Reducing Online Misinformation Exposure (ROME)*.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the*

Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16, page 3818–3824. AAAI Press.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717, Vancouver, Canada. Association for Computational Linguistics.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1980–1989.

Shahan Ali Memon and Kathleen M Carley. 2020. Characterizing covid-19 misinformation communities using a novel twitter dataset. In *CEUR Workshop Proceedings*, volume 2699.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2018. Credeye: A credibility lens for analyzing and explaining misinformation. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 155–158, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Harrisen Scells, Guido Zuccon, Bevan Koopman, and Justin Clark. 2020. Automatic boolean query formulation for systematic review literature search. In *Proceedings of The Web Conference 2020*, pages 1071–1081.

Shaden Shaar, Alex Nikolov, Nikolay Babulkov, Firoj Alam, Alberto Barrón-Cedeno, Tamer Elsayed, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Giovanni Da San Martino, et al. 2020. Overview of checkthat! 2020 english: Automatic identification

and verification of claims in social media. In *CLEF (Working Notes)*.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

Changhe Song, Cunchao Tu, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2018. Ced: Credible early detection of social media rumors.

Mahtab Tamannaee, Hossein Fani, Fattane Zarrinkalam, Jamil Samouh, Samad Paydar, and Ebrahim Bagheri. 2020. Reque: A configurable workflow and dataset collection for query refinement. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, CIKM '20, page 3165–3172, New York, NY, USA. Association for Computing Machinery.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. 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.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, et al. 2020. Cord-19: The covid-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. Recovery: A multimodal repository for covid-19 news credibility research. In *International Conference on Information and Knowledge Management, Proceedings*.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, 51(2).

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.