Knowledge Graphs for Real-World Rumour Verification

John Dougrez-Lewis[1], **Elena Kochkina**[2], **Maria Liakata**[1,2,4], **Yulan He**[1,3,4]
[1]University of Warwick [2]Queen-Mary University of London [3]King's College London
[4]The Alan Turing Institute
j.Dougrez-Lewis@warwick.ac.uk, e.kochkina@qmul.ac.uk
m.liakata@qmul.ac.uk, yulan.he@kcl.ac.uk

Abstract

Despite recent progress in automated rumour verification, little has been done on evaluating rumours in a real-world setting. We advance the state-of-the-art on the PHEME dataset, which consists of Twitter response threads collected as a rumour was unfolding. We automatically collect evidence relevant to PHEME and use it to construct knowledge graphs in a time-sensitive manner, excluding information post-dating rumour emergence. We identify discrepancies between the evidence retrieved and PHEME's labels, which are discussed in detail and amended to release an updated dataset. We develop a novel knowledge graph approach which finds paths linking disjoint fragments of evidence. Our rumour verification model which combines evidence from the graph outperforms the state-of-the-art on PHEME and has superior generisability when evaluated on a temporally distant rumour verification dataset.

Keywords: rumour verification, veracity assessment, PHEME, knowledge graph, real-world dataset

1. Introduction

Online misinformation remains highly prevalent, with the potential for harm increasing with the advent of advanced language models that can easily generate highly plausible false content. There is a growing need to create tools which can automatically debunk such texts. In this work, we focus on the task of identifying whether rumours circulating on social media are "*True*", "*False*", or "*Unverified*", where a rumour is defined as a check-worthy claim of unknown veracity (Zubiaga et al., 2018).

Although good progress has been made towards verifying the veracity of simple factual statements (Lin et al., 2019; Zhong et al., 2020), algorithmically generated fake news (Wu et al., 2022; Fung et al., 2021), and rumours from previously seen contexts, relatively little has been done to address the debunking of rumours in real-time. To better reflect a real-world scenario, we use the PHEME dataset (Zubiaga et al., 2016) which is specifically designed with this in mind. The dataset is divided into 9 folds, each containing rumours pertaining exclusively to an isolated *event* such as the 'Charlie Hebdo shooting'. To closely reflect a real-world setting, we conduct rumour veracity classification in accordance with the following rules:

- Rumours must be genuine posts by human users in a social media context;
- When evaluating a rumour in the test set, models must not have encountered related rumours or background information;
- Evidence must have been available around the time that a rumour emerged, to avoid contamination by future revelations.

To the best of our knowledge, our work is the first to produce a knowledge graph approach to use external evidence for rumour verification. The closest work, by Dougrez-Lewis et al. (2022), provides a mechanism for retrieving external evidence and baselines for utilising the evidence in verifying the rumour. In this work, we use our retrieved evidence to construct a knowledge graph for each PHEME event. We combine subject-predicate-object triples from the most useful sentences of that event's evidence, as determined by a novel sequence-matching metric. The resulting graph allows us to obtain refined evidence by discovering paths linking sources which would often otherwise be disjoint. The paths are linearised and fed as input to the rumour verification system along with the post originating the rumour. Our framework outperforms the state-of-the-art on PHEME and demonstrates high generalisability compared to previous models by training on PHEME but evaluating on the Covid-RV dataset (Kochkina et al., 2023), consisting of Covid-related tweets and associated response threads distant in time from PHEME and tackling highly disparate topics.

The paradigm of evaluating rumourous claims as they would emerge in real-time is a challenging one, with existing results on PHEME being relatively low (see Table 5). This has been attributed to the difficulty of dealing with emerging and unseen rumours, alongside class imbalance (Kotteti et al., 2020).

Using the evidence in our knowledge graphs, during error analysis, we discovered some labelling discrepancies affecting the *Unverified* class, assigned when there was not enough evidence available to verify a rumour. Whilst the vast majority of *True* and *False* annotations still hold, many of these *Unverified* rumours are readily verifiable using currently indexed evidence, even when the specific

LREC-COLING 2024, pages 9843–9853
20-25 May, 2024.

articles are viewed via WayBackMachine[1] the day before the rumour was posted. We amend the *Unverified* portion of the dataset and provide various experimental settings.

Specifically, we make the following contributions:

- We introduce a knowledge-graph-based approach, linking disjoint pieces of evidence, which achieves state-of-the-art results on the original and evidence-enhanced PHEME datasets under various training regimes § 3.2.
- We demonstrate the generalisability of our approach using a temporally distant dataset of rumour response threads, namely the Covid-RV dataset (Kochkina et al., 2023) § 4.4.
- We introduce a new sequence-matching metric which outperforms both traditional and autoencoder-based approaches to fake news information retrieval § 3.2.2.
- We amend the labelling of the PHEME dataset with respect to our retrieved evidence, and release an updated version, enhanced with external evidence § 3.1.

2. Related work

2.1. Datasets for Rumour Verification

Rumour verification (RV) is the task of classifying rumourous posts circulating on social media as being *True*, *False*, or *Unverified*. This has been tackled as single post RV (Zhao et al., 2015) and in the context of conversation threads, as in the PHEME (Zubiaga et al., 2016) dataset, Twitter 15, Twitter 16 (Ma et al., 2017), and Covid-RV (Kochkina et al., 2023). With the exception of PHEMEPlus (Dougrez-Lewis et al., 2022), which provides some retrieved external evidence for PHEME and Covid-RV, these datasets do not contain external evidence linked to the rumour. We leverage knowledge graphs to enhance the popular PHEME dataset with external evidence, finding links between disjoint sources (§ 3.1).

Yue et al. (2023) use a meta learning approach to transfer knowledge between various rumour datasets, although this does not work so well for novel happenings such as those of PHEME where the sequence of events is unrelated to prior information. It would be interesting to see their model modified to allow learning from not just rumours + labels, but also from relevant evidence.

2.2. Graph Based Rumour Verification

Evidence Based Approaches The FACE-KEG model by Vedula and Parthasarathy (2021) verifies claims by using DBPedia[2] as a knowledge graph embedded via a graph neural network. Whilst this approach works well for claims old enough to have reliable DBPedia entries, it is unsuitable for the evaluation of misinformation pertaining to current events. Lin et al. (2021) draw upon evidence found in the Twitter response threads, building interaction graphs between users. Although this is useful in a real-world context against newfangled rumours, and indeed predictive of their veracity (Dungs et al., 2018; Dougrez-Lewis et al., 2021), performance can be substantially improved using external sources.

Rule-Based Approaches There are several successful rule-based approaches such as Lin et al. (2019) which mines for the optimal fact-checking rules and Wang and Pan (2021) which combines both rule-based and neural approaches. Zhong et al. (2020) aggregate tuples into a graph, using it to re-define relative distances between words for later embedding. Our approach also involves the building of a knowledge graph, although we use it for the discovery of evidence paths prior to their subsequent linearization for model input (§ 3.2.4).

Neural Approaches with Synthesised Evidence Wu et al. (2022) and Fung et al. (2021) classify the veracity of (potentially) fake news articles via graph neural networks, synthesising the fake class due to a lack of data. Our updated version of the PHEME dataset which tries to align rumour veracity labels with the external available evidence (§ 3.1) similarly lacks *Unverified* rumours (Table 1, right), so for some experiments we use the more balanced original labels combined with synthetic evidence (§ 4.2).

2.3. Rumour Detection

Rumour detection is related to RV, but substantially different and largely solved. DDGCN (Sun et al., 2022) fuse the propagation of comments with a knowledge graph in an approach which is heavily geared towards rumour detection but of relatively little use for verification. DDGCN builds graphs on a per-thread rather than per-event basis, and its knowledge graph would not benefit from our retrieved evidence. Unlike DDGCN, our approach does not take into account the response thread for each rumour, although in the future there is scope for its incorporation.

MTLTS (Mukherjee et al., 2022) incrementally improve the results of rumour detection and do not evaluate on verification. The 'Ferguson' class, often the trickiest and most imbalanced in PHEME (at least for verification), is inexplicably omitted leading us to believe that comparisons against their work may be invalid.

[1] https://archive.org/web
[2] https://www.dbpedia.org

Events	Total	True	False	Unv.	Total	True	False	Unv.
Charlie Hebdo	458	193	116	149	458	320	119	19
Sydney Siege	522	382	86	54	522	421	90	11
Ferguson	284	10	8	266	284	147	16	121
Ottawa Shooting	470	329	72	69	452	340	73	39
Germanwings Crash	238	94	111	33	238	124	111	3
Total Threads	1972	1008	393	571	1954	1352	409	193

Table 1: Statistics of the PHEME-5 dataset, featuring both the original (left) and our amended (right) versions.

3. Methodology

3.1. Dataset: Updating PHEME

The PHEME dataset (Zubiaga et al., 2016) consists of Twitter conversations surrounding rumours taken from nine real-world events, such as the Germanwings crash of 2015. It is designed to closely reflect a real-world setting of rumour verification and was annotated by journalists days after the events unfolded. An overview of PHEME can be found in Table 1.

Unverified class was assigned when there was insufficient evidence available to annotators at the time. While traversing the evidence in our knowledge graphs we identified discrepancies regarding the labelling of these rumours. We hypothesise that these discrepancies are due to search engine algorithms and indexing being different at the time of annotation. Table 2 contains two examples, with evidence as it appeared before the time of rumour posting via WayBackMachine.

Relabelling Approach Given the contradictory evidence, we decided to relabel the *Unverified* rumours. Two annotators took two passes through each PHEME event. On the first pass, they became familiar with the evidence retrieved across all rumours of an event via three retrieval methods (see §3.2.4). This was to ensure consistency when labelling related rumours, irrespective of the correct evidence being retrieved for a particular instance. On the second pass, labels were assigned to the rumours.

Amendments Statistics of the relabelled dataset are in Table 1. The *True* label was the most frequently assigned. If the evidence was inconclusive the rumour retained the *Unverified* label. The two annotators agreed on 471/570 rumours, with all disparities resolved via discussion. 18 rumours relating to "an active shooter" were removed from the dataset entirely due to being either *True* or *False* depending on the time of evidence publication.

3.2. Knowledge-Graph Based Rumour Verification with External Evidence

An overview of our proposed evidence retrieval framework is shown in Figure 1.

3.2.1. Article Retrieval

Given a tweet initiating a rumour, relevant articles are retrieved using Google Search, following Dougrez-Lewis et al. (2022). Google Search is used as opposed to Wikipedia or other knowledge bases because our model verifies never-seen-before ongoing/recent rumours, so the latest information is needed. Furthermore, the vast majority of results are from high-quality journalistic sources.

Importantly, for each rumour, only articles from the same day or earlier are retrieved, enforced by putting a "BEFORE:Date" at the beginning of each search query. Queries are preprocessed using the "Preprocessed" strategy from (Dougrez-Lewis et al., 2022), which helps obtain more relevant search results. Up to 10 articles are retrieved for each rumour although there would be no issues with scaling, the possibility of improving performance by increasing this number remains to be explored.

Articles retrieved from Google are pooled by their corresponding fold of the PHEME dataset, ensuring that the model is not exposed to evidence from the unseen test event when evaluated under leave-one-out cross-validation. Entities from the retrieved articles are disambiguated on a per-article basis using CorefBERT (Ye et al., 2020).

3.2.2. Sentence Selection

Given a tweet initiating a rumour and articles retrieved via Google Search, we select the most relevant sentences via a novel sequence-matching approach.

The motivation for sequence matching stems from how rumourous events tend to take on highly specific meanings in the context of that event. For

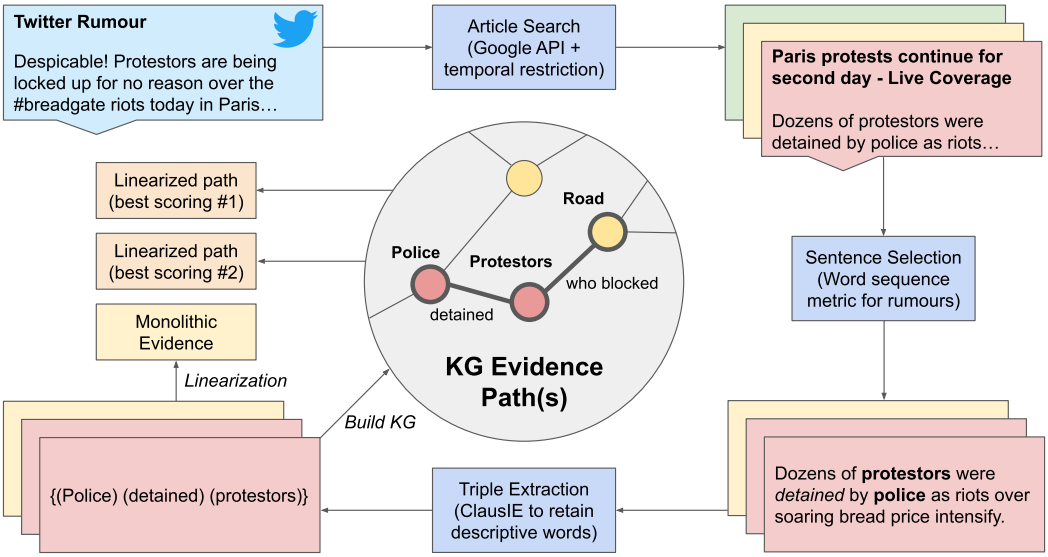

Figure 1: Overview of our evidence retrieval pipeline. Each of PHEME's events is handled separately to better reflect reality, as described in Section 1. Evidence is used by the model in Figure 2 to obtain the final prediction.

instance, "bank" is likely to refer to a specific building for a bank robbery, just as "vaccine" would likely refer to a Covid-specific vaccine in the context of the pandemic[3]. We hypothesize that the specificity of these specialized event words further increases when an entity is referenced using a sequence of words as opposed to a single word, and this is a frequent occurrence.

Thus, sentences are selected using a **sequence-matching metric** which rewards longer sequences of matching words between the sentence and the rumour. A sentence of *w* words containing *n* matching sequences each of length m_i is scored as follows:

$$\frac{\sum_{i=1}^{n} \frac{m_i(m_i+1)}{2}}{w} \quad (1)$$

Equation 1 is based on the sequence sum from 1 to m_i. For example, if there were a sentence of length 20 with matching fragments of lengths 2 and 3, the score would be (1+2 + 1+2+3)/20. Sequences in the candidate sentence are deemed matching irrespective of the position of their words in the rumourous tweet, exclude stopwords, and are uninterrupted by them. The top 5 sentences per rumour are used build the knowledge graph.

3.2.3. Knowledge Graph Construction

The chosen sentences are converted into subject-predicate-object triples using ClausIE (Del Corro and Gemulla, 2013), in preparation for knowledge graph construction. ClausIE is chosen here because it often brings many additional words into its triples, a useful property when working with recurrent event-specific phrases.

Triples are combined into a knowledge graph by creating a node for each subject or object, and an edge for each predicate linking them. Note that a separate graph is constructed for each of the PHEME events and its corresponding fold. The resulting knowledge graphs are mostly fully connected, with all edges and most nodes being unique, making them a poor fit for embedding by graph neural networks.

Two different approaches were considered for knowledge graph construction: (1) Use all subject-predicate-object triples from retrieved sentences; (2) Use only triples which contain *event-defining words* (Dougrez-Lewis et al., 2022) - nouns which occur frequently in an event's rumours, and are often the main focus of a rumour. The former approach was chosen as it resulted in graphs with large areas of intricately interconnected nodes, ideal for our goal of finding paths linking disparate articles, and yielding better final results.

3.2.4. Evidence Selection Strategies

In the PHEME dataset, rumours are raw tweets posted by users, initiating a rumourous conversation thread, and present several challenges; they are often nuanced and usually comprised of multiple components which can be treated as separate claims. To address this, we retrieve the final evi-

[3]At the time of writing in 2023, mention of this word is *also* taken to be Covid-specific.

Rumour	Selected Evidence
UPDATE: Parts of Sydney locked down amid cafe hostage crisis; Sydney Opera house evacuated	(1) Sydney, Dec 15: The Opera House in Sydney cancelled Monday night's performances in light of the situation in the central... (2) The Sydney Opera House has also been evacuated in response to the situation unfolding on Martin Place. (3) Indian Consulate in Sydney has been locked down and evacuated as a security measure following the incident.
VIDEO Police release surveillance tape from Ferguson store related to MichaelBrown:	(1) Police released still images and were planning to release video from the robbery, at a QuikTrip store in Ferguson. (2) Police released still images and were planning to release video from the robbery, at a QuikTrip store in Ferguson. (3) Chief Thomas Jackson also released documents and surveillance video, alleging that Mr. Brown was tied to a robbery...

Table 2: Examples of sentences selected by the three evidence selection methods, (1) Monolithic evidence; (2) Highest scored KG evidence; (3) Unique KG evidence. Both of these rumours were originally labelled *Unverified*, and relabelled *True*.

dence used for verification via multiple strategies, resulting in a more diverse evidence set. The evidence is combined with the rumour by the downstream classification model (Figure 2).

The first strategy involves simply choosing the top-scored sentence from the Google Search, using the same sequence-matching metric as before (Eq. 1). We call it **Monolithic evidence** as it comes from a single source.

The second and third strategies are based on the knowledge graph, which provides a richer connection between different pieces of evidence. The key idea is to find paths between entities in the graph, which bring together otherwise disjoint discourse from multiple sources. For each rumourous tweet, we find paths in its corresponding knowledge graph between all entities in the tweet, each path capped at a length of 3 edges to reduce noise. This interaction mining technique assumes that for a given event if multiple entities have the same name, their contexts are sufficiently distinct to work with our sequence matching metric.

Specifically, we iterate through an event's rumours, counting the number of times each node in the graph appears in any rumour. Nodes with lower counts are deemed more important since they are likely more specific to a subset of rumours rather than the event as a whole. Similarly, for each rumourous tweet, we count how many nodes relate to it, which are also the nodes we aim to find paths between. A path may take the form "X did A to Y" or "X did A to Y which did B to Z" and so on, involving up to 3 edges. We propose two strategies to select paths linking disjoint evidence from the knowledge graph (KG):

1. **Highest scored KG evidence**. Paths in the KG are flattened into text and scored using the metric from Eq. 1.
2. **Unique KG evidence** – choose the most specific path in the knowledge graph: the path between nodes A and B such that max(occurrences(A), occurrences(B)) is minimised, where occurrences(X) is the number of appearances of node X in the event's rumours. Ties are broken using the above metric (1).

The evidence selected from KG is essentially a path linking multiple entities found in a given rumour. Specifically, it can contain multiple linked triples in the form of (subject, predicate, object).

The retrieved sentences from which the triples originated from are concatenated to form the final evidence. Obtaining evidence by the above three separate strategies greatly improves the likelihood of finding relevant evidence. In particular, the graph-based approaches are often able to retrieve relevant evidence even when the Monolithic approach fails, due to their ability to find chains of logic linking multiple articles. The Unique approach is intended to find evidence in the case that the entities of a rumour are highly specific or obscure. See Table 2 for examples of the evidence retrieved by each strategy.

3.2.5. Rumour Verification Model

An overview of our model is shown in Figure 2. The rumourous tweet (excluding the corresponding response thread) and three retrieved pieces of evidence are encoded by a RoBERTa encoder (Liu et al., 2019), with joint loss. To best exploit the relationship between the rumour, evidence, and various retrieval methods, we use multi-headed attention with K = evidence, Q = rumourous tweet, and V = evidence. The final encoding is fed into a linear classifier which assigns a final label of *False*, *True*, or *Unverified*.

Model	False	True	Unverified	Accuracy	MacroF1
Original labels + Unmodified evidence	0.298	0.337	0.462	0.384	0.366
New labels + Unmodified evidence	0.266	**0.692**	0.146	**0.525**	0.368
Original labels + Modified evidence	**0.366**	0.575	**0.536**	0.523	**0.489**

Table 3: Results of our best model on variations of the PHEME dataset. New labels replace the original ones for the *Unverified* class, due to inconsistencies found during error analysis. Modified evidence replaces that which the model retrieved for the *Unverified* class, without necessitating its relabelling due to inconsistencies.

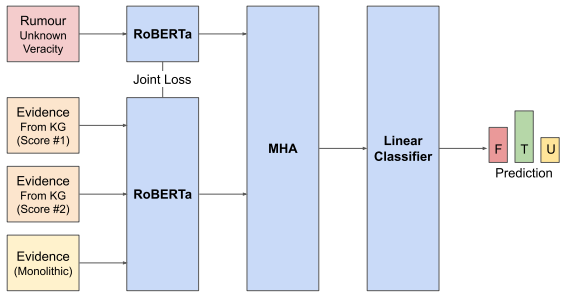

Figure 2: Overview of our model. Two pieces of evidence are retrieved from the Knowledge Graph and one (Monolithic) from the articles retrieved by Google Search. MHA = multi-headed attention.

4. Experiments

4.1. Experimental Setup

Model Overview Our model makes use of a RoBERTa encoder, with hidden layer size 768, 12 self-attention heads, and 125M parameters. RoBERTa outperforms both ALBERT and BERT. We use 8 self-attention heads in the multi-headed attention layer with dropout 0.15. The optimiser is AdamW with 1e-7 learning rate. 30 training epochs per fold. The inputs are the rumourous tweet (excluding its response thread) and our refined evidence, obtained using the three strategies from § 3.2.4.

Evaluation metrics We evaluate models using Accuracy and Macro F1 metrics, with F1 scores calculated by pooling results of the individual events before macro averaging.

Baselines and State-of-the-art We compare our proposed method with a variety of approaches, chosen to account for the state-of-the-art as well as popular architectures for rumour verification. These include multi-task approaches, variational autoencoders and a knowledge graph approach. All experiments are done using PHEME-5 for comparability, and results can be found in Table 5:

- MTL (Kochkina et al., 2018a) uses a multi-task learning approach to rumour classification, jointly learning to perform verification and classify the stances of the user's responses to the tweet originating the rumour.
- sLSTM (Li et al., 2019) also uses a multi-task learning approach, in addition to tweet metadata such as the verification status of a user's account.
- VRoC (Cheng et al., 2021) represents rumours via variation autoencoder, incorporating four related tasks in the multi-task paradigm.
- SAVED (Dougrez-Lewis et al., 2021) uses variational autoencoders to disentangle the topics discussed in rumours from their mannerisms, using the resulting latent vectors for prediction.
- KGAT (Wang et al., 2019) is a popular graph-embedding-based system we run using the same evidence as our model and the code publicly available on GitHub[4].

Traditionally, generic baselines underperform on PHEME due to the challenging structure of the rumours and the no-prior-knowledge paradigm (Kochkina et al., 2023).

Model	Accuracy	MacroF1
BERT	**0.685**	0.516
SAVED	0.592	0.558
Ours	**0.682**	**0.595**

Table 4: Binary classification results without the Unverified class. These results are essentially the same for both PHEME-5 and PHEME-9 datasets, the latter of which was used by Kochkina et al. (2023).

4.2. Results on PHEME

Results without *Unverified* class Table 3 (below) shows the results of our model under different settings concerning the *Unverified* class and its discrepancies with the evidence. Having noticed these discrepancies, we hypothesised that they would affect model performance. Running the model without the *Unverified* class as in Table 4 indeed yields increased performance. The relabelling of *Unverified* rumours, using the methodol-

[4] https://github.com/xiangwang1223

Model	False	True	Unverified	Accuracy	MacroF1
Majority Baseline	0	**0.818**	0	**0.692**	0.273
MTL	0.212	0.647	0.330	0.492	0.396
VRoC	**0.504**	0.480	0.465	0.521	0.484
sLSTM (Li et al., 2019)	-	-	-	0.483	0.418
SAVED	0.164	0.642	0.531	0.528	0.434
KGAT	0.258	0.541	0.216	0.404	0.338
Ours	0.366	0.575	**0.536**	0.523	**0.489**

Table 5: Comparison of our results with baselines and previous approaches, with F1 scores calculated by pooling results of the individual events before macro averaging. The difference between the first 2 rows is significant (p=0.04), for the others p<0.001.

Model	False	True	Unverified	Accuracy	MacroF1
Multiple Strategies					
Score + Score + Monolithic	**0.366**	**0.575**	**0.536**	**0.523**	**0.489**
Score + Unique + Monolithic	0.332	0.571	0.522	0.512	0.475
Single Strategies					
Score	0.318	0.386	0.505	0.412	0.403
Unique	0.275	0.404	0.430	0.389	0.370
Monolithic	0.316	0.528	0.523	0.485	0.456
Search Strategies					
Monolithic (Our Scoring)	0.316	0.528	0.523	0.485	0.456
Monolithic (BM25 Scoring)	0.306	0.513	0.339	0.420	0.386
Monolithic (DPR Scoring)	0.149	0.441	0.316	0.343	0.302

Table 6: Results of ablation studies using different evidence selection strategies, and different sentence scoring algorithms. Evidence is scored at the sentence selection stage using our scoring approach unless otherwise specified. In the case where the Score strategy is used twice, the two highest scoring sentences are used.

ogy in Section 3.1, while addressing the discrepancy issues between labels and evidence, caused extreme class imbalance which was also detrimental to performance (see second row of Table 3). In this case, whilst the overall accuracy was good due to the correct identification of *True* rumours, results for *Unverified* were poor due to the small class size after relabelling - almost non-existent in 3 of the folds.

Results on original PHEME, with more suitable evidence for *Unverified* To fix the imbalance issue, whilst maintaining comparability with previous approaches, we run our model on the original PHEME dataset but construct more suitable evidence for the *Unverified* class. The aim is to have useful evidence which pertains to the relevant event but is not helpful to assign either a *True* or *False* label.

To achieve this, we use some of the least relevant sentences retrieved by our score-based approach, modifying the metric not to normalize by length to prevent the selection of unduly long sentences. Using this new evidence for *Unverified* rumours yields strong performance gains for all 3 classes, which can be seen in the third row of Table 3. Our replacement evidence does not appear to unduly facilitate the classification of the *Unverified* class, as explained in the Limitations Section.

Comparison with baselines A comparison of our knowledge-graph-based approach with baselines is shown in Table 5. Our model is second best at the *False* class, beaten only by VROC (Cheng et al., 2021), the sole model which does adequately here. Whilst there seems to be some systematic difficulty with predicting this class, in this case, we suspect the main contributor is lack of evidence to reliably deem a rumour *False*. This is reflected in the manual annotation of the *Unverified* rumours, in which far more were deemed *True* than *False*. KGAT, the knowledge graph baseline trained under the same conditions and running on evidence retrieved via our pipeline, similarly struggles with *False* rumours. It is clear that our system for rumour verification, despite having a simpler architecture and no access to the response thread, is benefiting from the evidence retrieved through the knowledge graph compared to the other approaches.

Model	False	True	MacroF1
branchLSTM (Kochkina and Liakata (2020))	0.01	0.29	0.15
TD-RvNN (Ma et al. (2018))	0.01	0.45	0.23
BiGCN (Bian et al. (2020))	0.13	0.28	0.21
SAVED (Dougrez-Lewis et al. (2021))	0.30	0.39	0.35
BERT (Devlin et al. (2019)	0.06	0.34	0.20
Ours	**0.46**	**0.59**	**0.53**

Table 7: Results on training on PHEME (3-class) and evaluating on Covid-RV (2-class) in a zero-shot setting from Table 7 of Kochkina et al. (2023). MacroF1 scores are the average of those for the *False* and *True* classes. Whilst most models make use of Twitter response threads to rumours, only ours uses evidence from web search. Evidence for Covid-RV was gathered using the same Google-based pipeline as was used for PHEME.

4.3. Ablation Studies

We perform ablation studies to assess the contribution of the different components of our approach to the performance of the rumour verification model.

Scoring Metric Using the Monolithic evidence retrieval method from Section 3.2.4, we replace our scoring metric with BM25 and Dense Passage Retrieval (Karpukhin et al., 2020), with results in Table 6. Our consecutive word matching retrieval metric outperforms both of these approaches, in line with our hypotheses regarding event-specific phrases from Section 3.2.3. BM25 has no regard for consecutive matches. DPR performs surprisingly poorly, perhaps in part due to the rumours being different to the questions on which it was trained, despite finetuning on PHEME under the usual leave-one-out cross-validation paradigm. Kochkina et al. (2023) similarly find that BERT-based approaches such as DPR tend to perform worse than BM25 on the PHEME dataset.

Evidence Retrieval Methods We consider each evidence retrieval strategy from Section 3.2.4 individually, performing experiments on the original dataset with replacement evidence for Unverified, with results in Table 6. Although the use of sentences from the Monolithic retrieval strategy performs better than the knowledge graph based approaches, the best performance is achieved by combining the Monolithic and Score based approaches. Combinations of more than 3 pieces of evidence were not attempted due to computing limitations.

4.4. Cross Dataset Evaluation

To demonstrate the effectiveness of our model beyond PHEME, we use the challenging cross-dataset training paradigm (Li et al., 2019; Kochkina et al., 2023), training on PHEME and testing on the Covid-RV dataset (Kochkina et al., 2023). Covid-RV is comprised of rumours pertaining to Covid-19, temporally distant from PHEME and particularly challenging due to its wide variety of rumours, all considered to be from a single *event*. Evidence for Covid-RV was gathered using the same Google-based pipeline as was used for PHEME.

Results are in Table 7, with our model far outperforming previous attempts run by Kochkina et al. (2023) in a comparable manner, including good performance on the *False* class. Notably, this performance gain is heavily dependent on using our modified *Unverified* evidence when training on PHEME, lest the model essentially be trained on mismatched labels given the available evidence. Nevertheless, our knowledge graph based approach appears to be far more generalizable than its predecessors, substantially retaining its predictive power despite the change of dataset.

5. Conclusion

This work presents a knowledge-based graph approach to automatically retrieving evidence for rumour verification in a real-world setting. The evidence is fed to a rumour verification model which combines the original claim with the evidence, yielding state-of-the-art results on the PHEME dataset and also showing great generalisability to unseen rumours as they emerge as well as temporally distant datasets. Our work also provides an amended version of the PHEME dataset which can further the development of evidence-based approaches to rumour verification[5]. Future work aims to combine evidence-based approaches with approaches considering the context and discourse around a rumour as it unfolds.

Limitations

When re-annotating PHEME, due to the use of only model-retrieved evidence to save time, it proved difficult to annotate rumours as *False*. We erred on

[5]Code and data are available at https://github.com/johnnlp/

the side of caution for ambiguous cases, keeping their original *Unverified* label.

It is unlikely that our modified *Unverified* evidence made the class overly easy to predict, since under this regime the *Unverified* performance of the KGAT baseline in Table 5 is poor, as are the results of our ablation studies in Table 6. Training on PHEME with updated evidence was also a prerequisite to achieving any reasonable results when evaluating on the Covid-RV dataset.

Finally, previous approaches on the PHEME dataset have made use of the response thread initiated by each rumour, which can provide evidence as a rumour unfolds whilst access to more credible sources of evidence is still lacking. In the future, we would like to utilise this additional source of timely evidence.

Ethics Statement

We work with pre-existing rumour datasets the popular PHEME dataset, and a more recent dataset, COVID-RV, for both of which ethical approval had been obtained by the original research teams. All evidence we have used to augment the PHEME dataset is freely and readily available online via Google Search.

6. Acknowledgements

JDL was funded by the EPSRC Doctoral Training Grant. ML and YH are supported by Turing AI Fellowships (EP/V030302/1, EP/V020579/1, EP/V020579/2) funded by the UK Research and Innovation.

7. Bibliographical References

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):549–556.

Mingxi Cheng, Shahin Nazarian, and Paul Bogdan. 2021. Vroc: Variational autoencoder-aided multi-task rumor classifier based on text. *CoRR*, abs/2102.00816.

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. 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.

John Dougrez-Lewis, Elena Kochkina, Miguel Arana-Catania, Maria Liakata, and Yulan He. 2022. PHEMEPlus: Enriching social media rumour verification with external evidence. In *Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*, pages 49–58, Dublin, Ireland. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Sebastian Dungs, Ahmet Aker, Norbert Fuhr, and Kalina Bontcheva. 2018. Can rumour stance alone predict veracity? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3360–3370, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi Sil. 2021. InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Elena Kochkina, Tamanna Hossain, Robert L. Logan, Miguel Arana-Catania, Rob Procter, Arkaitz Zubiaga, Sameer Singh, Yulan He, and Maria Liakata. 2023. Evaluating the generalisability of

neural rumour verification models. *Information Processing Management*, 60(1):103116.

Elena Kochkina and Maria Liakata. 2020. Estimating predictive uncertainty for rumour verification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6964–6981, Online. Association for Computational Linguistics.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018a. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018b. PHEME dataset for Rumour Detection and Veracity Classification.

Chandra Mouli Madhav Kotteti, Xishuang Dong, and Lijun Qian. 2020. Ensemble deep learning on time-series representation of tweets for rumor detection in social media. *Applied Sciences*, 10(21).

Quanzhi Li, Qiong Zhang, and Luo Si. 2019. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1173–1179, Florence, Italy. Association for Computational Linguistics.

Hongzhan Lin, Jing Ma, Mingfei Cheng, Zhiwei Yang, Liangliang Chen, and Guang Chen. 2021. Rumor detection on Twitter with claim-guided hierarchical graph attention networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10035–10047, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peng Lin, Qi Song, Yinghui Wu, and Jiaxing Pi. 2019. Discovering patterns for fact checking in knowledge graphs. *J. Data and Information Quality*, 11(3).

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.

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)*, pages 1980–1989, Melbourne, Australia. Association for Computational Linguistics.

Rajdeep Mukherjee, Uppada Vishnu, Hari Chandana Peruri, Sourangshu Bhattacharya, Koustav Rudra, Pawan Goyal, and Niloy Ganguly. 2022. MTLTS: A multi-task framework to obtain trustworthy summaries from crisis-related microblogs. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. ACM.

Mengzhu Sun, Xi Zhang, Jiaqi Zheng, and Guixiang Ma. 2022. Ddgcn: Dual dynamic graph convolutional networks for rumor detection on social media. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4):4611–4619.

Nikhita Vedula and Srinivasan Parthasarathy. 2021. Face-keg: Fact checking explained using knowledge graphs. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 526–534, New York, NY, USA. Association for Computing Machinery.

Wenya Wang and Sinno Jialin Pan. 2021. Variational Deep Logic Network for Joint Inference of Entities and Relations. *Computational Linguistics*, 47(4):775–812.

Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. KGAT: knowledge graph attention network for recommendation. *CoRR*, abs/1905.07854.

Xueqing Wu, Kung-Hsiang Huang, Yi Fung, and Heng Ji. 2022. Cross-document misinformation detection based on event graph reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 543–558, Seattle, United States. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential reasoning learning for language representation. *CoRR*, abs/2004.06870.

Zhenrui Yue, Huimin Zeng, Yang Zhang, Lanyu Shang, and Dong Wang. 2023. Metaadapt: Domain adaptive few-shot misinformation detection via meta learning.

Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 1395–1405, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media. *ACM Computing Surveys*, 51(2):1–36.

Arkaitz Zubiaga, Geraldine Wong Sak Hoi, Maria Liakata, and Rob Procter. 2016. Pheme dataset of rumours and non-rumours.