Information Processing and Management 60 (2023) 103116

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

# Evaluating the generalisability of neural rumour verification models

Elena Kochkina [a,d,*], Tamanna Hossain [b], Robert L. Logan IV [b], Miguel Arana-Catania [c], Rob Procter [c,d], Arkaitz Zubiaga [a], Sameer Singh [b], Yulan He [c,d], Maria Liakata [a,d]

[a] Queen Mary University of London, London, UK
[b] University of California, Irvine, USA
[c] University of Warwick, Coventry, UK
[d] The Alan Turing Institute, London, UK

ARTICLE INFO

*Keywords:*
Rumour verification
Generalisability
Rumour dataset
Deep learning

ABSTRACT

Research on automated social media rumour verification, the task of identifying the veracity of questionable information circulating on social media, has yielded neural models achieving high performance, with accuracy scores that often exceed 90%. However, none of these studies focus on the real-world generalisability of the proposed approaches, that is whether the models perform well on datasets other than those on which they were initially trained and tested. In this work we aim to fill this gap by assessing the generalisability of top performing neural rumour verification models covering a range of different architectures from the perspectives of both topic and temporal robustness. For a more complete evaluation of generalisability, we collect and release COVID-RV, a novel dataset of Twitter conversations revolving around COVID-19 rumours. Unlike other existing COVID-19 datasets, our COVID-RV contains conversations around rumours that follow the format of prominent rumour verification benchmarks, while being different from them in terms of topic and time scale, thus allowing better assessment of the temporal robustness of the models. We evaluate model performance on COVID-RV and three popular rumour verification datasets to understand limitations and advantages of different model architectures, training datasets and evaluation scenarios. We find a dramatic drop in performance when testing models on a different dataset from that used for training. Further, we evaluate the ability of models to generalise in a few-shot learning setup, as well as when word embeddings are updated with the vocabulary of a new, unseen rumour. Drawing upon our experiments we discuss challenges and make recommendations for future research directions in addressing this important problem.

## 1. Introduction

### 1.1. Automated misinformation detection

The proliferation of misinformation on social media poses a serious threat to the functioning of society and the health and wellbeing of its citizens (Islam, et al., 2020). This issue has motivated efforts by fact checkers, journalists, social media platforms and researchers to develop ways to identify and debunk misinformation so as to mitigate its impact (Graves & Mantzarlis, 2020; Karafillakis, Van Damme, Hendrickx, & Larson, 2022; Shu, et al., 2020). A range of Natural Language Processing (NLP) techniques have been developed to address the challenges of identifying the veracity of content circulating on social media. A usual first step

* Corresponding author at: Queen Mary University of London, London, UK.
*E-mail address:* e.kochkina@qmul.ac.uk (E. Kochkina).

https://doi.org/10.1016/j.ipm.2022.103116
Received 19 April 2022; Received in revised form 7 October 2022; Accepted 9 October 2022
Available online 26 October 2022
0306-4573/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

is rumour detection, which consists in distinguishing check-worthy, widely spreading, unverified claims (*rumours*) from other kinds of content in social media posts (*non-rumours*) (Zubiaga, Aker, Bontcheva, Liakata, & Procter, 2018). Next, rumour verification characterises a rumour as *True*, *False* or *Unverified*, given a conversation thread, which consists of a sequence of posts discussing the rumour, linked through a reply relationship. Often rumour detection and rumour verification are operationalised as a single, combined, 4-way classification task.[1]

### 1.2. Motivation for generalisability study

In recent years a significant body of work has shown that utilising aspects of conversations around rumours such as structure of the conversation, and/or stance expressed in the responses when developing social media rumour verification models leads to important improvements in performance (Dougrez-Lewis, Liakata, Kochkina, & He, 2021; Huang, Zhou, Wu, Wang, & Wang, 2019; Khoo, Chieu, Qian, & Jiang, 2020; Kochkina, Liakata, & Zubiaga, 2018). In fact, some of these models achieve accuracy scores above 90% (Bian, et al., 2020; Huang, Yu, Wu, & Wang, 2020; Yuan, Ma, Zhou, Han, & Hu, 2019); However, these high performance scores may have been overstated as they have not been tested across datasets. Testing on a different dataset from the one used for training is a more realistic way to evaluate model robustness and generalisability (Marasović, 2018). Generalisability is increasingly becoming an important factor within NLP research, as is evident in recent work across different tasks, e.g., sentiment analysis (Moore & Rayson, 2018) and hate speech detection (Yin & Zubiaga, 2021). Despite the real-world importance of rumour verification as a task and the need for models to be able to apply to unseen rumours to be effective (Lukasik, et al., 2019), model generalisability in this area remains unexplored. New rumours always introduce unseen topics, thus leading to the usage of different vocabulary and new social media users contributing to them in distinctive ways. This poses a major challenge to model generalisability. However, social media rumour verification models are built on the hypothesis that conversations around rumours follow similar patterns across events and datasets, such as common stance and propagation patterns. Indeed, the existence of such patterns has been demonstrated in earlier work (Zubiaga, Liakata, Procter, Hoi, & Tolmie, 2016) and has been exploited to improve performance on veracity assessment (Dougrez-Lewis et al., 2021; Kochkina et al., 2018; Kumar & Carley, 2019). As a result one would expect that models that capture such patterns will be less susceptible to changes in vocabulary. When rumour verification models fail to generalise it is therefore not clear if they are mostly affected by a change in vocabulary, inability to exploit generic patterns of conversations around rumours or the lack of such patterns in some types of rumours. This makes the study of rumour verification model generalisability particularly interesting and challenging. Understanding the extent or the lack of model generalisability will help inform limitations of current approaches and identify future research directions to improve the state-of-the-art.

### 1.3. Our work

In this work we assess and analyse the ability of several top-performing rumour verification models to generalise to unseen rumours and events from the perspectives of both topic and temporal robustness. Our focus here is on rumour verification leveraging conversational threads discussing the rumours, following a line of research that has been widely studied in recent years (Bian, et al., 2020; Gao, Han, Song, & Ciravegna, 2020; Khoo et al., 2020).

To enable an up to date evaluation of generalisability across time, we also collect and release a novel dataset containing social media conversations around rumours involving COVID-19, a topic that sparked a high volume of posts and controversy (Pian, Chi, & Ma, 2021). Despite numerous efforts at producing COVID-19 misinformation datasets (Cui & Lee, 2020; Hossain, et al., 2020; Zhou, Mulay, Ferrara, & Zafarani, 2020), existing datasets do not contain conversations around rumours, along with their associated thread structures. They contain claims (Dharawat, Lourentzou, Morales, & Zhai, 2020), tweets (Hossain, et al., 2020), news headlines (Cui & Lee, 2020) and scientific publications (Wang, Lo, et al., 2020). The absence of social media conversations in these datasets currently prevents evaluation of verification models operating on the conversation structure (Bian, et al., 2020; Ma, Gao, & Wong, 2018). Following the structure and format of the PHEME and Twitter15/16 datasets, we therefore collect and release a novel, carefully curated dataset of rumours and corresponding Twitter conversations discussing them, which we use to evaluate the effectiveness of state-of-the-art rumour verification models.

### 1.4. Research questions

In evaluating generalisability of social media rumour verification models we are interested both in the performance gap when operating across datasets and also, importantly, in understanding what aspects of models and datasets affect this, so that verification techniques developed in future work may leverage these findings.

Specifically, we address the following research questions:

- *RQ1*: How well can rumour verification models generalise to unseen rumours across datasets from similar time periods and those more distant in time?
- *RQ2*: If the models experience performance drop on unseen datasets, will the ranking of model performance evaluated *across* datasets align with the ranking of these models when evaluated *within* one dataset?

[1] In this work we consider social media rumour verification as a standalone, 3-class classification problem as well as its combination with rumour detection into a 4-class problem; for brevity we will refer to both as rumour verification.

- *RQ3*: Which models or groups of models show better generalisability?
- *RQ4*: Which training datasets lead to better model generalisability? Does increasing the size of the training data improve generalisability?
- *RQ5*: Which data properties are related to performance drop?
- *RQ6*: Does the evaluation strategy used on the original training dataset play a role in training more generalisable models or estimating their future performance more realistically?
- *RQ7*: How receptive are different models to strategies such as few-shot learning and using embeddings updated with data from a new rumour?

### 1.5. Contributions

We make the following contributions:

- We are the first to test generalisability of top performing rumour verification models across datasets in two settings: (1) between datasets from the same time period; and (2) on test data collected at a later time period. We show that rumour verification models fail to generalise, with a sizeable performance drop when applied across datasets for all models.
- We provide extensive analysis of similarities and differences in model performance for five different models in several training settings.
- We release COVID-RV, a novel COVID-19 dataset of false claims and corresponding Twitter conversations to facilitate generalisability analysis.
- We investigate the difference in vocabulary between datasets and find that with the rise in distance between vocabularies we observe a higher performance drop.
- We discuss challenges and provide suggestions on ways to improve the generalisability of verification models.

## 2. Related work

In this section we describe the landscape of relevant works on rumour verification in terms of models and benchmark datasets, as well as novel datasets focusing specifically on COVID-19 rumours, and discuss how our work fits within the model generalisability domain.

### 2.1. Rumour verification models

Automated social media rumour verification is an active research area in NLP. Models have achieved high performance by leveraging linguistic, network- and user-related features (Chen, Zhou, Zhang, & Bonsangue, 2021; Kumari, Ashok, Ghosal, & Ekbal, 2022; Li, Fan, Yuan, & Zhang, 2022), propagation patterns, stance of the responses and conversation structures (Bian, et al., 2020; Dougrez-Lewis et al., 2021; Ma et al., 2018). Generalisability to new, unseen rumours is a crucial requirement for these models to be useful in real world settings. The fact that each new rumour introduces a new topic associated, which may be linked with fast-paced, ongoing and new real-life events, and attract discussions from diverse groups of individuals, makes this a challenging task and a problem inherent to rumour verification. Social media rumour verification models aim to resolve rumours at an early stage, and so they cannot always rely on the availability of confirmation from a specific, reliable source to determine their veracity (unlike the setup in Hossain, et al. (2020)). Thus, these models have to find generalisable signals among linguistic, network- and user-related features of Twitter conversations, rather that memorise information from the training set. Twitter15, Twitter16 (Ma, Gao, & Wong, 2017) and PHEME (Kochkina et al., 2018; Zubiaga, et al., 2016) are widely-used benchmark datasets for the tasks of rumour detection and verification. Many of the proposed models achieve high accuracy, such as 88% on Twitter15 in Bian, et al. (2020), 85% in Khoo et al. (2020), and 90.5% in Yuan, et al. (2019). This outperforms the performance of a non-professional human annotator on this task, which was estimated at around 60–65% in Kochkina et al. (2018). Every year, new approaches advancing the state-of-the-art performance appear, however, the question on how ready these are for application in a real-world setting remains unexplored. Papers proposing rumour verification models generally focus just on either Twitter15, Twitter16 (Huang et al., 2020; Huang, et al., 2019; Tu, et al., 2021; Wakamiya & Aramaki, 2020; Yuan, et al., 2019; Zhang, Cook, & Yilmaz, 2021) or PHEME (Kumar & Carley, 2019; Lee, et al., 2021; Roy, Bhanu, Saxena, Dandapat, & Chandra, 2022), with few papers using all three datasets (Khoo et al., 2020; Kochkina & Liakata, 2020) but still evaluating performance on each dataset separately. Ni, Li, and Kao (2021) perform cross-dataset performance comparison for several rumour datasets but their work is limited to a single BERT model operating on individual tweets. Furthermore, the work is limited to a binary classification setup, and the paper confuses the label definitions in the tasks, e.g. they perform rumour detection on PHEME using only rumour–non-rumour labels (meaning check-worthiness rather than veracity), but they are actually using veracity labels 'True' and 'False' from Twitter 15 and Twitter 16. Hence, we are the first to evaluate generalisability of a selection of top performing models across these popular datasets and also on a novel dataset of Twitter rumour conversations that is distant in time from the training data.[2] We expect a reasonable level of

[2] We note that separating domain/topic and temporal generalisability is not feasible in this case (unlike other tasks like sentiment analysis or stance classification) as new, unseen rumours appearing at a later date always concern new topics.

generalisability across the Twitter15/16 and PHEME datasets because data were collected around the same time period and share some of the topics. Generalisability to new rumours is currently addressed by using Leave-One-Event-Out (LOEO) cross-validation (CV) evaluation in the PHEME dataset. It has been shown that the performance difference between LOEO CV and random split CV is sizeable, and can reach around 38% (Khoo et al., 2020; Kochkina et al., 2018). Similarly, Zhang, Cao, et al. (2021) used temporal splits on the Weibo dataset (dataset of rumour conversations in Chinese from the Weibo microblog platform). However, temporal or event level splits are not addressed in research using the Twitter15/16 datasets. Thus, we hypothesise that due to the more strict LOEO CV evaluation, performance on PHEME is less likely to be overestimated than on Twitter15/16.

### 2.2. Generalisability

Generalisability is the capacity of a model to perform well on new, unseen data. Models are often evaluated on the test set with the assumption that future cases come from the same distribution as the training data. However, various NLP studies have reported a lack of generalisability among state-of-the-art models when tested on out-of-distribution data (Ettinger, Rao, Daumé, & Bender, 2017; Thakur, Reimers, Rücklé, Srivastava, & Gurevych, 2021), e.g., for hate speech (Yin & Zubiaga, 2021) and sentiment analysis (Moore & Rayson, 2018). Furthermore, recent studies (Alkhalifa, Kochkina, & Zubiaga, 2021; Röttger & Pierrehumbert, 2021) raise the issue of temporal robustness, i.e. performance drops when models are evaluated on data from the same domain but distant in time. This draws attention to the importance of generalisability and reveals domain-specific reasons for the lack of generalisability to inform future research.

To fulfil their purpose in the real-world, rumour verification models need to be able to deal with the constant growth in and evolution of rumours. However, existing research has not assessed the generalisability of rumour verification models. We fill this gap by evaluating the generalisability of rumour verification models across datasets, within a similar time period (between PHEME and Twitter15, Twitter16) and distant in time (testing on COVID-RV).

Domain adaptation (DA) and domain generalisation (DG) focus on models that learn from one or several different but related domains that will generalise well on unseen testing domains. In DA one can leverage unlabelled target information, but this is not the case for DG. Ramponi and Plank (2020) and Wang, et al. (2021) provide comprehensive recent surveys of DA and DG areas. One important point that Ramponi and Plank (2020) highlight is that there is no common ground on what constitutes a domain in NLP. For example, in the case of social media rumour verification, we could call each new rumour a new domain; or we could aggregate rumours into larger groups such as politics, celebrity or football, and treat these groups as domains; another option yet would be to treat the source of the data (e.g., Twitter, Facebook, News) or even each individual user as a domain.

Thus, in this paper, we do not consider generalisability of rumour verification models to be strictly a domain adaptation or domain generalisation problem. Rumours are always concerned with new, unexpected topics and events of various scales, and thus rumour verification models are designed and trained to identify features that are common/inherent to rumours or misinformation (such as writing style, stance and propagation patterns). These are expected to perform well across different topics and events. Depending on one's perspective of what constitutes a domain (e.g., each rumour or Twitter overall) we may or may not need to perform domain adaptation. Furthermore, as we are unable to foretell the topics of future rumours, we are also unable to accurately predict the scale of events that trigger rumours and the degree of their propagation. Therefore, it is impossible to tell whether new unsupervised event/rumour data will be accessible in time to support rumour resolution, and it is unclear whether DA or DG methodologies should be applied. Thus, contributing to DA or DG methodology is out of the scope of the current paper. Instead our goal is to evaluate the extent of generalisability of SOTA rumour verification models and what could be future fruitful avenues of research to improve it. While we study two setups from DA that can improve generalisability (few-shot learning and updated word embeddings), these have been chosen to serve as a baseline for future work in this direction.

In addition to providing a much needed generalisability investigation of rumour verification models,[3] we cater for multiple training data setups and compare leave-one-event-out with random cross-validation splits as evaluation approaches. This is in line with Wang, et al. (2021) and Zhou, Elfardy, Christodoulopoulos, Butler, and Bansal (2021) who stress the importance of train data quality and diversity, as well as the importance of evaluation setups, where the testing domain is unseen during training, such as leave one-domain-out cross-validation.

Another important question is what has a more significant effect on the generalisability of a rumour verification system: choice of model architecture or choice of training dataset. Gröndahl, Pajola, Juuti, Conti, and Asokan (2018) evaluate cross-dataset performance of hate speech detection models and report that for successful hate speech detection, model architecture is less important than the type of data and labelling criteria. This is somewhat expected as in their study different datasets include different types of hate speech, such as racist, sexist, offensive; furthermore, the notion of hate and offensive speech has a subjectivity component. Unlike Gröndahl, et al. (2018), we use datasets that have a consistent definition of a rumour and the verification process is arguably less subjective than that of identifying hate speech. Therefore, it remains to be seen whether the influence of the model architecture or the training data will be more significant, and we investigate this in this study.

### 2.3. COVID-19 datasets

The COVID-19 pandemic has been accompanied by a so-called *misinfodemic*: the wide spread of rumours and conspiracy theories about the coronavirus. To address this, the scientific community has been collecting datasets of true and false information on this

[3] Ramponi and Plank (2020) point out the over-representation of sentiment analysis task in generalisability studies and the need for other tasks to be addressed.

**Table 1**
Overview of existing COVID-19 datasets.

| Dataset | Claim | Tweet | Twitter conv. | News/ webpages | Manual annot. | Annotation Categories | Size | Language | Data dates |
|---|---|---|---|---|---|---|---|---|---|
| Poynter | ✓ | – | – | ✓ | ✓ | Various truthfulness categories | – | Multilingual | Dec 2019 - Present |
| FakeCOVID | ✓ | – | – | ✓ | ✓ | 11 categories of truthfulness | 7,623 fact-checks | Multilingual | January, 2020 - July, 2020 |
| PANACEA | ✓ | – | – | ✓ | – | Veracity: True, False | 5,143 claims | English | December, 2019 - September, 2020 |
| ReCOVery | – | ✓ | – | ✓ | – | Source credibility: Reliable, Unreliable | 2,029 news, 140,820 tweets | English | January, 2020 - May, 2020 |
| COAID | ✓ | ✓ | – | ✓ | – | Claim veracity: True, False | 482 claims, 3,769 news, 160,667 tweets | Mostly English | December, 2019 - September, 2020 |
| COVID-HERA | ✓ | ✓ | – | ✓ | ✓ | Claim health risk assessment: Possibly severe, Highly severe, Refutes/Rebuts, Other, Real News/Claims | 61,286 tweets | English | December, 2019 - September, 2020 |
| MM-COVID | – | ✓ | – | ✓ | – | Real and Fake News | 11,565 news, 105,300 tweets | English, Spanish, Portuguese, Hindi, French and Italian | February, 2020 - July, 2020 |
| CMU-MisCOV19 | – | ✓ | – | – | ✓ | 17 categories of tweets: Irrelevant, Conspiracy, True treatment, True prevention, Fake Cure, etc. | 4,573 tweets | English | weeks beginning 29th March, 15th June and 24th June 2020 |
| CheckThat! 2020 | – | ✓ | – | – | ✓ | Check-worthiness: Rumour, Non-rumour | 7,500 tweets in Arabic 628 tweets in English | English, Arabic | March, 2020 |
| COVID-19 Rumour | – | ✓ | – | ✓ | ✓ | Stance: Support, Deny, Question, Comment, Sentiment: Very negative, Negative, Neutral, Positive, Very positive, Veracity: True, False, Unverified | 4,129 news, 2,705 source tweets + additional responses | English | January, 2020 - April, 2020 |
| COVID-19-Stance | – | ✓ | – | – | ✓ | Stance: In-favour, Against, Neither, Sentiment: Negative, Positive, Opinion: Implicit, Explicit | 7,122 tweets +automatically annotated | English | March, 2020 - August, 2020 |
| COVMis-Stance | ✓ | ✓ | – | – | ✓ | Stance: Favour, Against, Neither | 2,631 tweets | English | – |
| CovidLies | ✓ | ✓ | – | – | ✓ | Relevance: Relevant, Non-relevant, Stance: Agree, Disagree, No stance | 6,761 tweets (1013 relevant) | English | March, 2020 - April, 2020 |
| COVID-RV (ours) | ✓ | ✓ | ✓ | – | ✓ | Relevance: Relevant, Non-relevant, Stance: Agree, Disagree, No stance Veracity: True, False | 2,445 tweets (907 relevant) | English | January, 2020 - November, 2020 |

topic from various sources. These include scientific publications, news articles and their headlines, social media posts and claims about COVID-19. Table 1 summarises existing datasets and some of their properties. Datasets relevant to our work include:

- Poynter:[4] The CoronaVirusFacts/DatosCoronaVirus Alliance Database that gathers all of the falsehoods that have been detected by the CoronaVirusFacts/DatosCoronaVirus alliance. This database unites fact-checkers from more than 70 countries and includes articles published in at least 40 languages.
- FakeCOVID (Shahi & Nandini, 2020): multilingual cross-domain dataset of 7623 fact-checked news articles about COVID-19 from 92 fact-checking websites after obtaining references from Poynter and Snopes. Manually annotated into 11 categories of the fact-checked news according to their content.
- PANACEA (Arana-Catania, et al., 2022): dataset consisting of heterogeneous claims on COVID-19 and their respective information sources.
- ReCOVery (Zhou et al., 2020): news articles on coronavirus annotated using the level of credibility of their source, along with tweets that reference these news articles up to May 2020.
- COAID (Cui & Lee, 2020): COVID-19 fake news on websites between December, 2019 and September, 2020 and social media platforms, users' social engagement with such news; tweets automatically identified as relevant to the claims.
- COVID-HERA (Dharawat et al., 2020): individual COVID-19 tweets from COAID (Cui & Lee, 2020) and their health risk assessment.
- MM-COVID (Li, Jiang, Shu, & Liu, 2020): multilingual, multimodal dataset containing fake news and the relevant social context from February to July 2020.
- CMU-MisCOV19 (Memon & Carley, 2020): communities of Twitter users, with their posts collected over three weeks beginning 29th March, 15th June and 24th June 2020 and classifying them as either informed or misinformed.

[4] https://www.poynter.org/ifcn-covid-19-misinformation/

- CheckThat! 2020 (Shaar, et al., 2020): individual COVID-19 tweets from March 2020 labelled as either rumours or non-rumours, depending on the check-worthiness of the content.
- COVID-19 Rumour (Cheng, et al., 2021): news articles and tweets discussing rumours about COVID-19 from the period between January and April 2020, annotated for stance, sentiment and veracity.
- COVID-19-Stance (Glandt, Khanal, Li, Caragea, & Caragea, 2021): stance detection dataset towards four targets 'Wearing a Face Mask', 'Keeping Schools Closed', 'Anthony S. Fauci, M.D', 'Stay at Home Orders'.
- CovidLies (Hossain, et al., 2020): tweets posted between March and April 2020, discussing false claims about COVID-19, annotated for their stance.
- COVMis-Stance (Hou, van der Putten, & Verberne, 2022): stance dataset consisting of 2631 tweets annotated with the stance towards COVID-19 misinformation.

Poynter, FakeCOVID (Shahi & Nandini, 2020) and PANACEA (Arana-Catania, et al., 2022) datasets focus on collecting fact-checked claims about COVID-19, rather than related social media posts. CheckThat! 2020 (Shaar, et al., 2020), ReCOVery (Zhou et al., 2020), CMU-MisCOV19 (Memon & Carley, 2020), and COVID-19-Stance (Glandt, et al., 2021), while addressing relevant tasks, do not provide or focus on veracity annotations. Datasets such as COAID (Cui & Lee, 2020) and MM-COVID (Li et al., 2020) (also COVID-HERA (Dharawat et al., 2020) as a subset of COAID) are collected and/or annotated using automated means (fully or partially) and contain potential noise in labels when matching between rumourous claims and social media content, and/or duplicate claims. One of the main distinctive properties of our dataset compared to those listed above is that it is carefully curated and manually annotated in two stages.

Our dataset COVID-RV (Section 3) extends CovidLies (Hossain, et al., 2020) by associating social media conversations with claims in CovidLies. The set of claims has been further refined compared to CovidLies to remove time-dependent, multi-part, ambiguous and duplicate claims. Furthermore, unlike other datasets, which either use individual posts (Dharawat et al., 2020; Hossain, et al., 2020; Hou et al., 2022; Shaar, et al., 2020) or do not provide the connections between posts (Cheng, et al., 2021), in our new dataset we focus on finding relevant tweets that are sources of conversations around a rumour and then collect the relevant conversations, which are then used in rumour verification models. The conversations in COVID-RV are of the same type as those in PHEME (Zubiaga, et al., 2016) and Twitter15, Twitter 16 (Ma et al., 2017), which led to the creation of a plethora of automated verification models (see Section 2.1), the robustness of which we test in Sections 4 and 5. It is also worth noting that our dataset covers a wider time period than the above-mentioned prior datasets, that is, between January and November 2020.

## 3. Creating the COVID-RV dataset

In this section we describe the creation process of the COVID-RV (COVID-Rumour Verification) dataset, which consists of: (1) matching claims to tweets; (2) relevance annotation; (3) stance annotation; and (4) collecting conversation threads.

*Claim–tweet matching* We start with a set of 62 false claims (misconceptions) about COVID-19 refined from CovidLies (Hossain, et al., 2020). The claims in CovidLies were sourced from Wikipedia and manually re-written to be a positive expression of a misconception.[5] Claims pertaining to the actions of particular political parties, governments, religious groups, or ethnicities were removed as these do not usually pertain to the topic of the COVID-19 pandemic but rather events happening during the same time period; claims referring to photos or videos (multi-modal) were also removed, as they require different approaches to verification, which involve taking other modalities into account. Compound claims were split into atomic ones, some claims were corrected and some edited for brevity and duplicate claims were removed (see Appendix A).

We use the COVID-19-TweetIDs collection (Chen, et al., 2020) to identify tweets matching the 62 claims so as to collect associated tweet threads. In line with previous work (Zubiaga, et al., 2016), we only use tweets in English with over 100 retweets. These are dated between January and November 2020, resulting in a total of 424,073 tweets. We do not edit the content of the tweets, and each model defines its own input representation based on the text of the tweet. Models that use multimodal aspects of the tweets are not represented in the current study; this is left for future work. To match claims and tweets, Hossain, et al. (2020) used BERTScore (Zhang, Kishore, Wu, Weinberger, & Artzi, 2020), which resulted in only 15% of the pairings being actual matches. We found that BM25-based (Robertson, et al., 1995) re-ranking methods are better at retrieving relevant tweets compared to BERT-based methods, with BM25+Mono T5 re-ranking (Nogueira, Lin, & Epistemic, 2019) being the most effective in our preliminary experiments.

*Tweet retrieval with BM25+Mono T5* To match claims with relevant tweets we index the set of tweets and use claims as queries. We retrieve the top 100 matches per claim.[6] according to their BM25 score and then re-rank these retrieved pairs using the Mono T5 model. We use the implementation of BM25 in Pyserini and Mono T5 re-ranking in Pygaggle Python packages.[7] The scores returned by the re-ranking algorithm are converted to relevance probabilities, and only claim–tweet pairs with probability threshold > 90%[8] are kept for annotation (resulting in 1215 instances to annotate).

[5] https://github.com/ucinlp/covid19-data

[6] We found in preliminary experiments that 100 matches per claim gives us enough good matches but that also that the amount of relevant instances decreases the closer we get to the 100th instance.

[7] https://github.com/castorini/pyserini, https://github.com/castorini/pygaggle

[8] In our preliminary experiments we found this to be a strong threshold that minimises the number of irrelevant instances while providing a good number of relevant claim–tweet pairs.

**Table 2**
Relevance and stance annotation outcomes for the two claim–tweet matching methods.

| | BM25+Mono T5 re-ranking | DPR |
|---|---|---|
| Relevant | 671 | 236 |
| Non-relevant | 544 | 994 |
| Agree | 378 | 103 |
| Disagree | 205 | 89 |
| No stance | 14 | 1 |
| Tie/No label | 10 | 2 |

*Tweet retrieval with DPR* In addition to BM25+Mono T5 we apply the Dense Passage Retrieval[9] (DPR) (Karpukhin, et al., 2020) method to the same set of tweets and claims as above. We use DPR as it represents a different type of approach compared to BM25, relying on contextual language models rather than exact word matching. We are interested to see whether it would be a complementary approach, which would allow us to retrieve instances missed by BM25. DPR employs a dual-encoder framework to produce dense representations of queries and passages using a neural encoder, e.g., BERT (Devlin, Chang, Lee, & Toutanova, 2019). Once representations are obtained, retrieval is performed using cosine similarity. We obtain our encoder by further fine-tuning on COAID (Cui & Lee, 2020) the encoder originally trained on Natural Questions, using the default settings (Kwiatkowski, et al., 2019). For annotation we take the top 20 instances for each claim. We chose this number of instances to approximately match the number of instances manually annotated for the BM25+Mono T5 matching method. Among those instances we found that the overlap between results returned by the two methods is very low (only 8%), showing that the two methods can indeed be used in a complementary manner.

*Relevance annotation* We annotated the tweets identified by the BM25+Mono T5 and DPR methods for relevance to the claim they were paired with. For this we used Amazon Mechanical Turk (MTurk) and recruited 3 annotators per claim–tweet instance. As per our annotation guidelines the relevance of each tweet is judged based on its connection to the specific claim, rather than the general topic of COVID-19. Annotators were given the option to open any links present in the tweet if they judged it useful for determining relevance. Annotators could also flag issues with a tweet (see Appendix B for details on annotator recruitment and guidelines). Out of 2445 annotated pairs: 1969 received 100% annotator agreement, i.e., all 3 annotators selected the same label (relevant or non-relevant), 21 pairs were flagged as having an issue by at least one of the annotators and in 681 tweets links were opened by at least one of the annotators. The top two rows of Table 2 shows the results of manual annotation of relevance on claim–tweet pairs returned by both matching methods. We found that BM25+Mono T5 re-ranking returns many more relevant claim–tweet pairs than DPR.

We then dropped any accidental duplicates, instances flagged as having an issue and included only relevant tweets with 100% annotator agreement in the next round of annotations for stance.

*Stance annotation* Tweets annotated as relevant were subsequently annotated for stance, which is known to be particularly useful for rumour verification from social media conversations (Zubiaga, Kochkina, et al., 2018). The stance annotation interface was similar to the relevance annotation one but, rather than using MTurk workers, we employed students and university staff volunteers from the United States with English proficiency. Annotation guidelines (Appendix C) provide instructions and examples for labelling the stance of a tweet towards a claim as either: *Agree* if the tweet agrees with the claim, *Disagree* if it disagrees, and *No Stance* if the tweet expresses no opinion towards the claim.[10] As before, annotators could flag any issues with tweets and three annotators worked on each claim–tweet pair. We labelled each claim–tweet pair as *Agree*, *Disagree*, or *No Stance* based on the majority agreement between annotators, or *as Tie/No label* if there was no consensus. The inter-annotator agreement was 88%. This was computed as the average percent agreement between annotators per instance. The four bottom rows of Table 2 present the results of the stance annotation for each of the matching methods. We observe that the majority of relevant tweets agree with the matched claim.

*Collecting conversations* As in previous work involving the resolution of social media rumours (see Section 2.1) we collect conversations consisting of tweets discussing the rumour. For tweets that are labelled as either *Agree* or *Disagree*, we collect the associated conversations using the Twitter API v2. The majority of tweets (97%) initiate the conversations (i.e., are source tweets), so we use their IDs to download the replies tree. When an annotated tweet is not a source tweet, we download the tweet object first, then get the conversation ID and finally get the conversation. For each conversation, we reconstruct the conversation tree using parent–child tweet pairs.

We notice that COVID-RV contains more tweets per conversation than other datasets (see Table 3), which may be due to several reasons: (1) we choose our source tweets to be the tweets that attracted at least 100 retweets. While this is also a criterion for source tweet selection in the PHEME dataset, the COVID-19 pandemic has attracted unprecedented attention from the public worldwide compared to events in PHEME and other datasets; and (2) the set of claims that COVID-RV contains are the ones that have attracted

[9] https://github.com/facebookresearch/DPR

[10] This stance annotation approach differs from the one in the PHEME dataset as here we are mainly interested in the *Agree* and *Disagree* categories for source tweets of the conversations so as to provide veracity labels to the conversations, rather than labelling all of the responses in order to gain explicit fine-grained support patterns (as in PHEME), which could be done as part of future work for this dataset.

**Table 3**
Number of posts, conversation trees and class distribution in the datasets (T – *True*, F – *False*, U – *Unverified*, NR – *Non-Rumour*).

| | # Posts | # Trees | Median N tweets | Median N branches | Median depth | T | F | U | NR |
|---|---|---|---|---|---|---|---|---|---|
| PHEME | 105 354 | 6425 | 10 | 7 | 3 | 1067 | 638 | 697 | 4023 |
| Twitter15 | 40 927 | 1374 | 17 | 16 | 3 | 350 | 336 | 326 | 362 |
| Twitter16 | 18 770 | 735 | 15 | 14 | 3 | 189 | 173 | 174 | 199 |
| COVID-RV | 133 335 | 775 | 53 | 36 | 5 | 294 | 481 | – | – |

the most attention out of all circulating claims since they are described in Wikipedia. This property of COVID-RV differentiates it from previous datasets. However, this a natural effect, mainly linked to the prominence and scale of a rumour and it can not be controlled for when collecting new unseen rumours and events. Furthermore Ma et al. (2018) and Bian, et al. (2020) show performance increase in time as more information becomes available.

*Veracity labels* Rumour verification models operating on conversation threads require veracity labels. We label conversations on the basis of the source tweet of the conversation introducing the rumour. The conversation is labelled as 'False' if the corresponding source tweet agrees with the rumour and 'True' otherwise. Table 3 shows the resulting number of tweets contained in conversations around the claims, as well as the statistics for other rumour datasets we use for training in Section 4. Fig. 1 shows an example of an instance from our dataset.

## 4 Evaluating rumour verification models

### 4.1 Models

Here we describe the models whose generalisability we test across datasets. These models were selected among the top-performing rumour verification models with publicly available code that enabled reproducibility. In all cases, we keep the original model parameters proposed in the corresponding articles. This selection includes comparison of models of different types in terms of: (1) the input word-level representation, including models that take as input one-hot embeddings and Bag-of-Words representations (TD-RvNN, BiGCN, SAVED), ones that take Word2vec (Mikolov, Chen, Corrado, & Dean, 2013) embeddings (branchLSTM) and large contextual LMs (BERT, CT-BERT); (2) the representation of the conversation structure including models that use trees (TD-RvNN,BiGCN), linear sequences (branchLSTM, SAVED) or individual tweets only (BERT, CT-BERT); (3) vocabulary associated with the rumour, with most models unaware of COVID vocabulary and a couple exposed to unannotated COVID tweets (CT-BERT, CT-SAVED). See summary of model properties in Table 4. Here we focus our study on single-task learning models. Since multitask learning has been a very successful method for rumour verification (Lee, et al., 2021), we plan to explore this setting in future work. An overview of the models chosen is provided below:

*branchLSTM* Kochkina and Liakata (2020), Kochkina et al. (2018) uses linear tweet branches from rumour conversations as input and an LSTM-based model to process them.[11] An average of per-branch predictions is used to obtain the final veracity prediction for the full tree. This model was originally proposed for RumourEval-2017 dataset (Kochkina, Liakata, & Augenstein, 2017) and then was tested on PHEME (Kochkina et al., 2018) and Twitter15/16 (Kochkina & Liakata, 2020). It was also a strong baseline for RumourEval-2019 (Gorrell, et al., 2019).

*Top-Down Recursive Neural Networks (TD-RvNN)* Ma et al. (2018) are top-down tree-structured neural networks for rumour representation learning and classification, which naturally conform to the propagation layout of tweets or a tree structure of a conversation.[12] This model was originally proposed and tested on Twitter15/16 datasets (Ma et al., 2018).

*Bidirectional Graph Convolutional Neural Network (BiGCN)* Bian, et al. (2020) operates on both top-down and bottom-up propagation of rumours.[13] It leverages a GCN with a top-down directed graph of rumour spreading to learn the patterns of rumour propagation, and a GCN with an opposite directed graph of rumour diffusion to capture the structures of rumour dispersion. This model was originally proposed on and tested on Twitter15/16 and Weibo datasets in Bian, et al. (2020).

*Stance-Augmented VAE Disentanglement framework (SAVED)* Dougrez-Lewis et al. (2021) is a two stage approach to rumour verification, the current state-of-the-art on the PHEME dataset.[14] First, a Variational Autoencoder is used to obtain representations of each rumour by disentangling the informational content of a tweet from the manner in which it is written. This is achieved by obtaining latent topic vectors in an adversarial learning setting using the auxiliary task of stance classification. The resulting latent vectors are then used to predict rumour veracity. This model was originally proposed and tested on the PHEME-5 dataset, i.e., using only the 5 largest events from PHEME. Here we use the full PHEME dataset with 9 events, therefore reported results differ from Dougrez-Lewis et al. (2021). We have also trained a CT-SAVED model which is a variant of SAVED, where the Variational

[11] https://github.com/kochkinaelena/Uncertainty4VerificationModels
[12] https://github.com/majingCUHK/Rumor_RvNN
[13] https://github.com/TianBian95/BiGCN
[14] https://github.com/JohnNLP/SAVED

**Table 4**
Summary of model properties.

| | Tweet representation | Conversation representation | COVID-19 vocabulary |
|---|---|---|---|
| branchLSTM | Word2vec | Linear sequences of tweets | – |
| TD-RvNN | One-hot encoding | Tree structure | – |
| BiGCN | One-hot encoding | Tree structure | – |
| SAVED | One-hot encoding | Linear sequences of tweets | – |
| BERT | Contextual LM | Individual tweets | – |
| CT-SAVED | One-hot encoding | Linear sequences of tweets | Yes |
| CT-BERT | Contextual LM | Individual tweets | Yes |

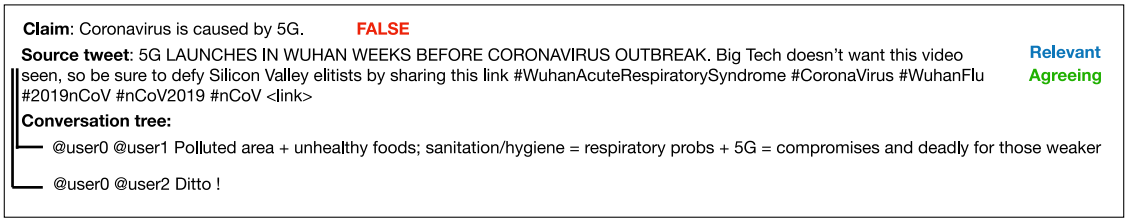


**Fig. 1.** Example instance from the COVID-RV dataset. Each instance contains a *False* claim, a *Relevant* source tweet annotated as *Agreeing* or *Disagreeing* with the claim, as well as tree-structured conversation around the source tweet conveying the rumour.

Autoencoder is trained using the data from the training set as well as additional unlabelled COVID-19 Twitter conversations.[15] Thus the topic-discourse module of the CT-SAVED model has been exposed to COVID-19 vocabulary and can therefore produce COVID-19-aware tweet representations.

*Large pre-trained language models BERT* Devlin et al. (2019) and CT-BERT (Müller, Salathé, & Kummervold, 2020).[16] We use these to obtain a representation and then classify rumours based purely on the source tweet of the conversation. We compare a general-purpose pre-COVID BERT model with a Twitter-specific one that includes a more up-to-date lexicon of COVID tweets (CT-BERT). BERT has been previously tested on the PHEME-5 dataset in Dougrez-Lewis et al. (2021) in a similar setting.

### 4.2 Data

To train the models we use publicly available datasets of Twitter rumour conversations (see Table 3 for details). These include:

*PHEME* Twitter conversations discussing rumours about nine breaking news events, which were labelled as *True*, *False* or *Unverified* by journalists (Zubiaga, et al., 2016). The PHEME dataset also contains conversations that discuss the same events but are labelled as *Non-Rumours* (Kochkina et al., 2018).

*Twitter15/16* The Twitter15 and Twitter16 datasets (Ma et al., 2017) were created using reference datasets from Ma, et al. (2016) and Liu, Nourbakhsh, Li, Fang, and Shah (2015). Claims were annotated using veracity labels on the basis of articles corresponding to claims found in rumour debunking websites such as *snopes.com* and *emergent.info*. These datasets merge rumour detection and verification into a single, four-way classification task containing *True*, *False* and *Unverified* rumours as well as *Non-Rumours*.

Both datasets are split into 5 folds for cross-validation and, contrary to the PHEME dataset, folds are of approximately equal size with a balanced class distribution. It is not possible to apply leave-one-event-out cross-validation to Twitter15 and Twitter16 datasets as the event split is not provided. The overall data format is practically equivalent in all of the datasets (as shown in Fig. 1), which enables their use within the same rumour verification models interchangeably.

### 4.3 Experiment setup

We split our experiments into 4 groups:

1. **In-dataset evaluation** of models (on PHEME, Twitter15, Twitter16) using the evaluation strategy per dataset as published in previous works (leave-one-event-out cross-validation for PHEME and 5-fold cross-validation for Twitter15, Twitter16);
2. **Cross-dataset evaluation** of models *on datasets from a similar time period* (training on PHEME, testing on Twitter15/Twitter16 and vice versa);

[15] We have collected approximately 19,000 conversations discussing COVID-19 using a subset of tweets from our filtered COVID-19 set of tweets.
[16] https://huggingface.co

**Table 5**
Performance of the models evaluated within each dataset and across the datasets from similar time period. Abbreviations: PH4 – PHEME4, Tw15 – Twitter15, Acc. – accuracy, MF – macro-averaged F1-score.

| | In-dataset | | | | | | Cross-dataset | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PHEME4 | | Tw15 | | Tw16 | | PH4 → Tw15 | | PH4 → Tw16 | | PH4 → Tw15+16 | | Tw15 → PH4 | |
| | Acc. | MF | Acc. | MF | Acc. | MF | Acc. | MF | Acc. | MF | Acc. | MF | Acc. | MF |
| branchLSTM | 0.610 | 0.299 | 0.591 | 0.597 | 0.788 | 0.787 | 0.300 | 0.206 | 0.369 | 0.279 | 0.324 | 0.232 | **0.251** | **0.244** |
| TD-RvNN | 0.555 | 0.297 | 0.707 | 0.723 | 0.718 | 0.737 | 0.345 | 0.270 | 0.388 | 0.304 | 0.360 | 0.284 | 0.180 | 0.201 |
| BiGCN | 0.555 | 0.314 | **0.886** | 0.860 | **0.880** | **0.862** | **0.400** | **0.376** | 0.407 | **0.351** | **0.402** | **0.373** | 0.179 | 0.194 |
| SAVED | 0.468 | 0.298 | 0.667 | 0.661 | 0.623 | 0.631 | 0.341 | 0.307 | 0.363 | 0.331 | 0.348 | 0.315 | 0.152 | 0.167 |
| BERT | **0.628** | **0.336** | 0.867 | **0.865** | 0.848 | 0.848 | 0.377 | 0.327 | **0.414** | 0.320 | 0.390 | 0.335 | 0.152 | 0.157 |

**Table 6**
Performance of the models trained on PHEME, Twitter15/16 and their combinations, evaluated on COVID-RV (Acc. – accuracy, MF – macro-averaged F1-score.). The majority class in the training data is *True*, so if the model predicts *True* all the time we will see accuracy of 0.398 and macro-averaged F1-score of 0.285. Bold: highest result in a column; underscore: highest result in a row.

| | PHEME4 | | PHEME3 | | Tw15 | | Tw16 | | Tw15+Tw16 | | Tw15+16+PHEME3 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. | MF | Acc. | MF | Acc. | MF | Acc. | MF | Acc. | MF | Acc. | MF |
| branchLSTM | 0.022 | 0.018 | 0.092 | 0.098 | 0.173 | 0.121 | 0.100 | 0.068 | 0.249 | 0.140 | **0.340** | **0.223** |
| TD-RvNN | 0.003 | 0.003 | 0.173 | 0.153 | 0.323 | 0.169 | 0.113 | 0.086 | **0.338** | **0.189** | 0.269 | 0.162 |
| BiGCN | 0.024 | 0.025 | 0.149 | 0.135 | 0.123 | 0.085 | 0.093 | 0.070 | 0.205 | 0.121 | 0.204 | 0.134 |
| SAVED | 0.010 | 0.010 | 0.273 | 0.231 | 0.186 | 0.112 | **0.461** | **0.199** | 0.178 | 0.107 | 0.254 | 0.161 |
| BERT | 0.005 | 0.006 | 0.063 | 0.081 | 0.088 | 0.059 | 0.021 | 0.016 | 0.135 | 0.082 | 0.033 | 0.029 |
| CT-SAVED | **0.238** | **0.125** | 0.253 | 0.202 | **0.386** | **0.215** | 0.081 | 0.062 | 0.215 | 0.149 | 0.270 | 0.176 |
| CT-BERT | 0.000 | 0.000 | **0.398** | **0.285** | 0.161 | 0.097 | 0.067 | 0.049 | 0.000 | 0.000 | 0.075 | 0.055 |

3. **Cross-dataset temporal robustness assessment** by evaluating all models on COVID-RV. This presents models with new rumours distant in time from the training data. When evaluating the models on COVID-RV we also compare results obtained using different training datasets and their combinations, to evaluate the effect of existing resources on model performance.
4. **Assessing few-shot learning capabilities** of the models on COVID-RV.

We use the following training combinations: 3-class classification with PHEME dataset (PHEME3, i.e. True(T)/ False(F)/ Unverified(U)), 4-class classification with PHEME dataset (PHEME4, i.e. True(T)/ False(F)/ Unverified(U)/ Non-Rumour(NR)), 4-class classification with Twitter15, Twitter16 datasets, and combinations of Twitter15 + Twitter16 and Twitter15 + Twitter16 + PHEME3 The new COVID-RV dataset is used exclusively for testing. While the PHEME dataset also has 4 classes, in previous work it is usually split into two separate tasks - binary rumour detection (Rumour vs Non-Rumour) and 3-way veracity classification (T/F/U). Here, to enable cross-dataset evaluation between PHEME and Twitter15/16, we use all available conversations from PHEME together for 4-way classification. While COVID-RV used for testing only has True and False classes, we chose to train the models using the original number of labels in the training datasets, including Unverified and Non-rumour classes. We aim to imitate a realistic scenario in which an existing pre-trained model is used for predictions.

In our experiments we truncate the largest conversations from COVID-RV to have a maximum of 1000 branches, and use the first 20 responses from each branch in order to make it computationally feasible.

We keep all the original hyper-parameter values as fixing hyper-parameters allows us to compare the models in different training settings (see the values in Appendix D). We report the result of a single run in our results tables.

We evaluate the performance of our models in terms of accuracy and macro-averaged F1-score. Macro-averaged F1-score (MF) is particularly suitable to evaluate performance on the PHEME dataset as it contains significant class imbalance. Evaluation on COVID-RV mainly focuses on per-class performance for the True and False classes.

## 5 Results

### *5.1 Generalisability across datasets from the same time period*

Table 5 shows in- and cross-dataset model performance for datasets from a similar time period (PHEME and Twitter15/16,[17] both from 2014–2016). In the in-dataset experiments we do not observe a consistent ranking of model performance. We acknowledge that this observation can be somewhat affected by tuning hyper-parameters of each model to each dataset individually. However, it is not the goal of this paper to reach the highest possible performance with each model, but to evaluate their generalisability in various settings. Thus we preserve the hyper-parameters across our experiments for fair comparison between setups. BERT and BiGCN models do consistently better than branchLSTM, TD-RvNN and SAVED in this setting. Performance on the PHEME dataset

[17] Here we only present results of training on Twitter15 when evaluating on PHEME as Twitter15 is the largest of the two very similar datasets.

**Table 7**
Per-class performance for *True* and *False* classes of models trained on PHEME, Twitter15/16 and their combinations, evaluated on COVID-RV. Bold: highest result in a column; underscore: highest result in a row. A majority baseline (always *True* class) would score 0.57 *True* class F-score.

| | | PHEME4 | | | PHEME3 | | | Tw 15 | | | Tw 16 | | | Tw 15+16 | | | PHEME3+Tw 15+16 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | P | R | F | P | R | F | P | R | F | P | R | F | P | R | F | P | R | F |
| branchLSTM | T | 0.08 | 0 | 0.01 | 0.40 | 0.23 | 0.29 | 0.34 | 0.26 | 0.30 | 0.14 | 0 | 0.01 | 0.31 | 0.06 | 0.11 | 0.34 | 0.61 | **0.43** |
| | F | 0.52 | 0.03 | 0.06 | 1 | 0 | 0.01 | 0.57 | 0.11 | 0.19 | 0.71 | 0.16 | 0.27 | 0.59 | 0.37 | 0.46 | 0.44 | 0.16 | 0.24 |
| TD-RvNN | T | 0.50 | 0.01 | 0.02 | 0.48 | 0.43 | 0.45 | 0.40 | 0.08 | 0.13 | 0.39 | 0.26 | **0.31** | 0.40 | 0.16 | 0.22 | 0.43 | 0.20 | 0.27 |
| | F | 0 | 0 | 0 | 0.20 | 0 | 0.01 | 0.62 | 0.49 | **0.55** | 0.35 | 0.02 | 0.04 | 0.64 | 0.46 | **0.53** | 0.63 | 0.32 | **0.43** |
| BiGCN | T | 0.34 | 0.05 | **0.08** | 0.37 | 0.26 | 0.28 | 0.48 | 0.05 | 0.08 | 0.38 | 0.05 | 0.09 | 0.34 | 0.07 | 0.11 | 0.44 | 0.12 | 0.19 |
| | F | 0.70 | 0.01 | 0.02 | 0.54 | 0.08 | 0.13 | 0.50 | 0.18 | 0.26 | 0.57 | 0.12 | 0.19 | 0.52 | 0.29 | 0.37 | 0.55 | 0.26 | 0.35 |
| SAVED | T | 0.83 | 0.01 | 0.03 | 0.64 | 0.28 | 0.39 | 0.38 | 0.02 | 0.04 | 0.47 | 0.08 | 0.14 | 0.32 | 0.02 | 0.04 | 0.30 | 0.19 | 0.23 |
| | F | 0.50 | 0.01 | 0.02 | 0.36 | 0.26 | **0.30** | 0.68 | 0.30 | 0.41 | 0.61 | 0.71 | **0.66** | 0.67 | 0.14 | 0.24 | 0.48 | 0.03 | 0.05 |
| CT-SAVED | T | 0 | 0 | 0 | 0.40 | 0.49 | 0.44 | 0.39 | 0.30 | **0.34** | 0.41 | 0.03 | 0.05 | 0.46 | 0.39 | **0.42** | 0.45 | 0.19 | 0.27 |
| | F | 0.68 | 0.40 | **0.50** | 0.53 | 0.10 | 0.17 | 0.64 | 0.44 | 0.52 | 0.66 | 0.12 | 0.20 | 0.60 | 0.10 | 0.17 | 0.66 | 0.32 | **0.43** |
| BERT | T | 0.75 | 0.01 | 0.02 | 0.58 | 0.24 | 0.34 | 1 | 0.01 | 0.02 | 0 | 0 | 0 | 0.57 | 0.02 | 0.03 | 0.75 | 0.02 | 0.05 |
| | F | 0 | 0 | 0 | 0.71 | 0.03 | 0.06 | 0.50 | 0.14 | 0.22 | 0.52 | 0.03 | 0.06 | 0.48 | 0.21 | 0.30 | 0.42 | 0.04 | 0.07 |
| CT-BERT | T | 0 | 0 | 0 | 0.40 | 1 | **0.57** | 0.20 | 0.01 | 0.02 | 0.67 | 0.01 | 0.02 | 0 | 0 | 0 | 0.25 | 0.03 | 0.06 |
| | F | 0 | 0 | 0 | 0 | 0 | 0 | 0.61 | 0.26 | 0.36 | 0.57 | 0.11 | 0.18 | 0 | 0 | 0 | 0.38 | 0.10 | 0.16 |

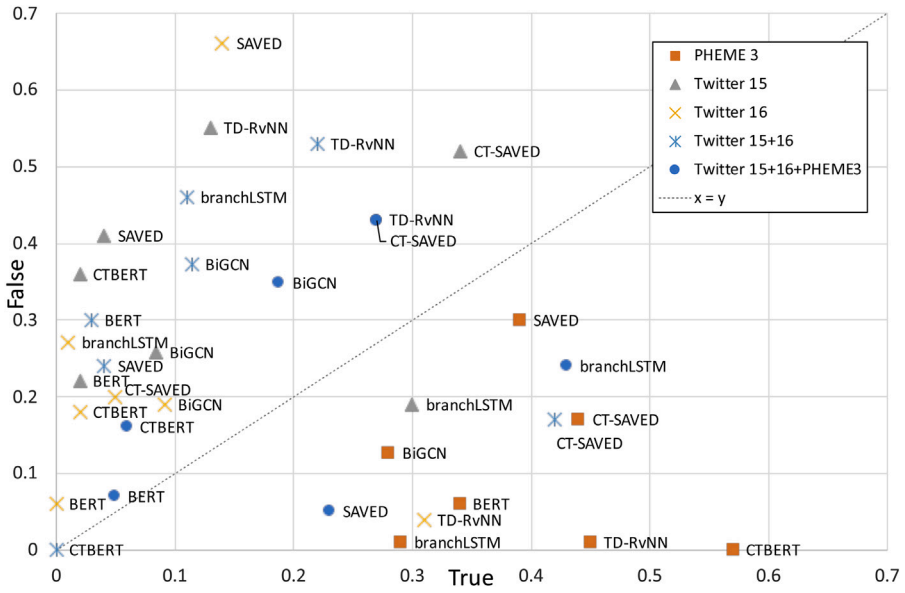


**Fig. 2.** The plot of F-scores for *True* and *False* classes of COVID-RV. The colour and shape of a marker identify a training dataset and each point is labelled with the model used to obtain the result. The best performing models are the ones closest to the dotted $x = y$ line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

is noticeably lower than on Twitter15/16. This could be due to several important differences in the datasets. Firstly, performance on the PHEME dataset is evaluated using leave-one-event-out cross-validation, while Twitter15/16 are split into 5 folds without separation between events. The folds in PHEME are of different size and different class proportions, while Twitter15/16 folds are balanced. Evaluation on the PHEME dataset is more challenging and closer to a real-world setting.

In the cross-dataset experiments between PHEME and Twitter15/16, we observe a sizeable performance drop (*RQ1*). This highlights that *none of the datasets realistically represent the distribution of the unseen data and the original performance was overestimated.* We notice that for models trained on PHEME4 and evaluated on Twitter15/16 the drop in macro-averaged F1-score is not as dramatic compared to the drop of the models trained on Twitter15. This indicates that *the performance on PHEME is not overestimated to the same degree because of its more realistic evaluation setup* (*RQ6*).

We find that for models trained on PHEME4 the performance ranking order of models remains similar, with BiGCN and BERT being the top performing models (*RQ3*). However, for models trained on Twitter15 and evaluated on PHEME, there is a dramatic drop in performance, the performance ranking is flipped, and the simpler branchLSTM model gives the best results (*RQ3*). This shows that previous performance scores and even *model ranking can be unreliable when tested on a dataset different to the one used for training* (*RQ2*). The performance of the models trained on PHEME4 dataset is also higher comparing to models trained on Twitter15. It shows the robustness of models trained on a dataset which contains cross-event variability, such as PHEME.

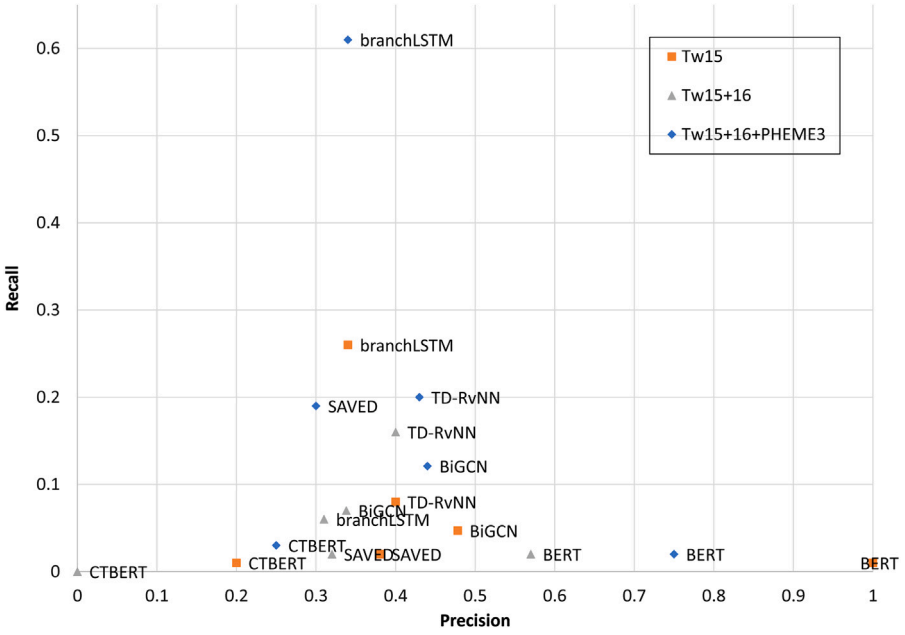


**Fig. 3.** Precision - Recall plot for *True* class.

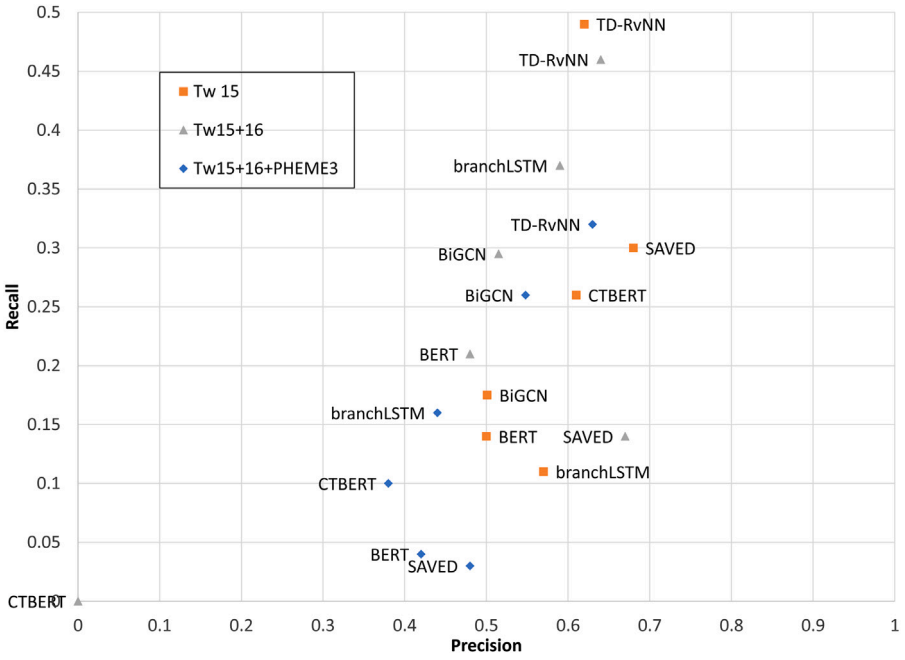


**Fig. 4.** Precision - Recall plot for *False* class.

### 5.2 Generalisability to COVID-RV

Table 6 shows the performance of rumour verification models on COVID-RV using different training data scenarios. In all of the scenarios, due to its size, COVID-RV is only used for testing and not for training. The training dataset used is shown in the column names. We observe very low performance scores in terms of accuracy and macro-averaged F1-score; all the models score lower than a majority class baseline in terms of macro-averaged F1-score, demonstrating the challenge of generalising across time (*RQ1*). These low performance scores can be somewhat explained by the fact that the models are trained to predict three or four classes rather

than two, and in experiments with 3-way classification we see higher performance than in those with 4-classes.[18] For example, zero performance occurs in several cases with the CT-BERT model when it predicts all of the instances as either 'Non-rumour' or 'Unverified'. However, we chose to preserve the amount of classes in the data as our task is to imitate the realistic scenario in which ready made models are facing new data with new vocabularies and class balance.

We find that model ranking differs from what we observed in previous experiments (*RQ2*). Each of the models outperforms the rest for some training dataset in Table 6, especially SAVED and CT-SAVED. This is no longer the case for BiGCN and BERT, which now have the lowest performance. When we calculate the mean average of scores per model across all training settings, SAVED (the model which exploits the difference in topics discussed in a conversation from they way they are discussed) has the highest performance in terms of both accuracy and macro-averaged F1-score (*RQ3*).

When we look at overall performance in Table 6 we can notice that not all of the models benefit from using all of the datasets for training. In particular, some models experience changes in performance when using combinations of the datasets (e.g., CT-BERT model trained on Twitter15 or Twitter16 individually has better performance than on their combination, and when training on Twitter15 it performs better than training on all three datasets combined). This is somewhat expected and can be explained by the combinations of datasets affecting per class performance differently, e.g., due to differences in class balance. As a result the models in some cases will start predicting the classes not present in the testing data more frequently. When we calculate the average performance for each of the datasets from Table 6 across all pre-COVID models (i.e. excluding CT-BERT and CT-SAVED), we find that indeed *on average the best performance is achieved by the combination of all three training datasets* (*RQ4*).

In our experiments, both across PHEME and Twitter15/16 datasets and testing on COVID-RV, unlike Gröndahl, et al. (2018), who found that the choice of training data is more significant than the choice of the model architecture for model generalisability (see Section 2.2), we do not observe that the change of model architecture has a significantly higher or lower impact on rumour verification performance than the change of the training dataset.

Given the low overall scores, we focus on per-class evaluation on the *True* and *False* classes to compare performance of models in various training setups. Fig. 2 shows per-class F-scores for *True* and *False* classes for each model and training set. These results in a table format, including per-class precision and recall, can be found in Table 7, along with Figs. 3 and 4 visualising per-class precision and recall. The best performing model needs to score high on both the *True* and *False* class, thus models that have balanced performance on both classes would lie along an $x = y$ line on Fig. 2. On the plot we can see that models trained on PHEME3 tend to perform better on the *True* class as it is a majority class in that dataset, while models trained on Twitter15 perform better on the *False* class. *Models trained on the combinations of all of the training data* are the closest to the $x = y$ line, i.e., *have the most balanced performance.* The CT-SAVED model trained on the Twitter15 dataset stands out in Fig. 2 as having the best and most balanced performance, followed by TD-RvNN and CT-SAVED trained on all of the training datasets. We also found per-class precision to be consistently higher than recall across the majority of the experiments, which is expected due to a high number of instances being classified as either *Non-Rumour* or *Unverified*.

We also test the benefits of using embeddings trained with data covering the topic and time period in the test set via CT-BERT and CT-SAVED. Results in Table 6 show that CT-BERT performs better than BERT in most cases. CT-SAVED also outperforms SAVED in most setups. This confirms that *updating embeddings is a promising method to improve performance of rumour verification models for new rumour events* in line with work in other fields (Alkhalifa et al., 2021) (*RQ7*).

### 5.3 Few-shot learning experiments

Few-shot learning (Wang, Yao, et al., 2020) enables testing the ability of models to learn effectively from a small number of instances. We evaluate the benefits of few-shot learning in experiments with COVID-RV, combining it with the use of up-to-date COVID tweet (CT) embeddings (*RQ7*). We consider two settings: (1) adding one conversation from each claim into the training data; and (2) adding three conversations from each claim. We have made this choice of a number of instances to add so that all of the claims would receive equal coverage among the few-shot learning examples added to the training. The COVID-RV dataset is not very large, and some claims only have 5 tweets associated with them. A few-shot learning approach implies only using a small amount of data, and therefore we chose to use 1 example per claim (lowest possible) and 3 examples per claim (towards the higher end, whist still being applicable to all of the claims). This setup somewhat changes our task as now the model is exposed to a set of annotated conversations around each claim during training and thus 'knows' the correct facts or has a chance of memorising them. Here we used the combination of all datasets (Twitter15+Twitter16+PHEME3) as our training data. Table 8 shows performance of the four-class classification models in a few-shot learning setup in terms of accuracy and macro-averaged F1-score. We compare these to the results in the last column of Table 6 (denoted as zero-shot in Table 8). We see improvement of performance for all of the models. Adding more instances is also beneficial in most of the cases. The observed performance is now on par with performance in evaluation across datasets from a similar time period. Therefore, *few-shot learning helps bridge the gap between the datasets distant in topic and time.* However, there is still need for improvement to reach the in-dataset performance.

We have also analysed these performance improvements from the per-class perspective (see Table 9). The combination of the few-shot approach with updated CT embeddings leads to further improvement in per-class performance. As COVID-RV is imbalanced towards the *False* class, our few-shot sample is also imbalanced towards *False*, therefore, for most of the models we see higher improvement in per-class performance for the *False* class.

[18] We have additionally performed binary classification experiments. We present and discuss them in Appendix E as the outcomes align with the conclusions we draw on the full datasets.

**Table 8**
Performance of the models on COVID-RV in a few-shot learning setup compared to zero-shot setup (first column). Best performance per row is highlighted in bold.

| | zero-shot | | 1 per claim | | 3 per claim | |
|---|---|---|---|---|---|---|
| | Acc. | MF | Acc. | MF | Acc. | MF |
| TD-RvNN | 0.27 | 0.16 | 0.49 | **0.19** | **0.50** | 0.18 |
| BiGCCN | 0.20 | 0.13 | 0.51 | **0.18** | **0.52** | **0.18** |
| SAVED | 0.25 | 0.16 | 0.29 | 0.17 | **0.40** | **0.19** |
| BERT | 0.03 | 0.03 | 0.58 | 0.20 | **0.60** | **0.50** |
| CT-SAVED | 0.27 | 0.18 | 0.40 | 0.21 | **0.47** | **0.21** |
| CT-BERT | 0.08 | 0.06 | 0.55 | 0.21 | **0.62** | **0.37** |

**Table 9**
Per-class performance of the models on COVID-RV in our few-shot learning setup. Each column shows the results of including the additional tweets for each claim into the training data.

| | | zero-shot | | | 1 per claim | | | 3 per claim | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | P | R | F | P | R | F | P | R | F |
| TD-RvNN | T | 0.43 | 0.20 | **0.27** | 0.58 | 0.03 | 0.06 | 0.43 | 0.03 | 0.05 |
| | F | 0.63 | 0.32 | **0.43** | 0.58 | 0.82 | 0.68 | 0.55 | 0.91 | 0.68 |
| BiGCN | T | 0.44 | 0.12 | 0.19 | 0.26 | 0.01 | 0.02 | 0.39 | 0.01 | 0.01 |
| | F | 0.55 | 0.26 | 0.35 | 0.57 | 0.88 | 0.70 | 0.54 | 0.95 | 0.69 |
| SAVED | T | 0.30 | 0.19 | 0.23 | 0.50 | 0.12 | 0.20 | 0.47 | 0.71 | 0.57 |
| | F | 0.48 | 0.03 | 0.05 | 0.60 | 0.43 | 0.50 | 0.47 | 0.14 | 0.22 |
| BERT | T | 0.75 | 0.02 | 0.05 | 0.75 | 0.03 | 0.05 | 0.87 | 0.16 | 0.27 |
| | F | 0.42 | 0.04 | 0.07 | 0.58 | 0.99 | **0.73** | 0.58 | 0.98 | 0.73 |
| CT-BERT | T | 0.25 | 0.03 | 0.06 | 0.86 | 0.05 | 0.10 | 0.88 | 0.23 | 0.36 |
| | F | 0.38 | 0.10 | 0.16 | 0.60 | 0.92 | **0.73** | 0.60 | 0.96 | **0.74** |
| CT-SAVED | T | 0.45 | 0.19 | **0.27** | 0.34 | 0.31 | **0.32** | 0.46 | 0.82 | **0.59** |
| | F | 0.66 | 0.32 | **0.43** | 0.58 | 0.47 | 0.52 | 0.59 | 0.17 | 0.27 |

**Table 10**
Kullback–Leibler Divergence, Intersection Over Union (IoU) and DICE scores for pairs of datasets and mean average accuracy across models for the corresponding dataset pairs.

| | KL | IoU | DICE | Avg. Acc |
|---|---|---|---|---|
| PHEME4-PHEME4 | 0.191 | 0.069 | 0.124 | 0.563 |
| Tw16-Tw16 | 0.060 | 0.168 | 0.287 | 0.771 |
| Tw15-Tw15 | 0.107 | 0.144 | 0.252 | 0.744 |
| PHEME4-Tw15 | 0.474 | 0.143 | 0.249 | 0.353 |
| PHEME4-Tw16 | 0.240 | 0.128 | 0.227 | 0.388 |
| Tw15-PHEME4 | 0.326 | 0.143 | 0.249 | 0.183 |
| Tw15-COVID-RV | 0.493 | 0.112 | 0.202 | 0.179 |
| Tw16-COVID-RV | 0.333 | 0.099 | 0.179 | 0.158 |
| PHEME4-COVID-RV | 0.560 | 0.124 | 0.221 | 0.013 |
| Tw15+16-COVID-RV | 0.462 | 0.121 | 0.216 | 0.221 |
| Tw15+16+PHEME3-COVID-RV | 0.512 | 0.125 | 0.222 | 0.220 |
| Tw15+16+PHEME3-COVID-RV (fewshot, 1 per claim) | 0.508 | 0.136 | 0.238 | 0.470 |
| Tw15+16+PHEME3-COVID-RV (fewshot, 3 per claim) | 0.499 | 0.287 | 0.445 | 0.505 |
| Pearson correlation coefficient | **−0.76** | 0.30 | 0.29 | |

We find that CT-SAVED and CT-BERT benefit the most from few-shot training and result in a relatively balanced performance on the *True* and *False* classes compared to TD-RvNN and BiGCN, which perform poorly and do not show improvement in the *True* class. This could be explained by the ability of these models to make better use of the COVID vocabulary, recognising post COVID words as meaningful rather than treating them as unknown tokens, which would make it harder to learn these from few-shot examples. The strengths of CT-SAVED and CT-BERT appear complementary to each other with CT-BERT performing best on the *False* class, and CT-SAVED performing best on the *True* class.

### *5.4 Effect of distance between datasets*

We hypothesise that a performance drop arises from differences between training and test data and that the performance gap decreases the closer the datasets are to each other (*RQ5*). To test this hypothesis we measure the difference between datasets using the Kullback–Leibler Divergence (KL) $D_{KL}(P\|Q)$, Jaccard Index (Intersection over Union - IoU) and DICE coefficient. We chose

**Table 11**
Pearson correlation coefficient between model performance and conversation length and depth.

| Model | Length | Depth |
|---|---|---|
| branchLSTM | 0.04 | 0.01 |
| TD-RvNN | −0.01 | 0.2 |
| BiGCN | 0.05 | 0.09 |
| SAVED | −0.07 | 0.02 |
| BERT | −0.04 | −0.04 |

these metrics because they are common metrics to measure distance between corpora (Lu, Henchion, & Mac Namee, 2021; Peinelt, Liakata, & Nguyen, 2019). We define them below.

A corpus can be regarded as a probability distribution across words in a vocabulary, and the KL divergence between two corpora P and Q can be calculated as $D_{KL}(P\|Q) = \sum_{i=1}^{n} p_i log_2 \frac{p_i}{q_i}$, where $n$ is the number of unique words in the two corpora, $p_i$ and $q_i$ are the probabilities of observing word $i$ in corpus $P$ and $Q$ respectively estimated through dividing the $i$th word occurrence frequency by the total number of words in the corpus.

The Jaccard Index (Intersection over Union - IoU) and DICE coefficient are calculated using the following equations:

$$IoU = \frac{|V^p \cap V^q|}{|V^p \cup V^q|},$$

$$DICE = \frac{2 \times |V^p \cap V^q|}{|V^p| + |V^q|},$$

where $V^p$ and $V^q$ - are the sets of unique words from two corpora, $|V^p|$ denotes the size of the set, thus $|V^p \cap V^q|$ - number of unique words present in the intersection of the corpora vocabularies, and $|V^p \cup V^q|$ - number of unique words present in the union of $V^p$ and $V^q$.

IoU and DICE equal to 1 when the datasets are identical, and zero for datasets with no vocabulary overlap. KL divergence is zero when the datasets are identical and has an unbounded upper value when the datasets are dissimilar. Further, we calculate the Pearson correlation coefficient between the distance scores and mean average of accuracy scores across models for the corresponding dataset pairs. Table 10 shows the distance metrics and average performance scores for the dataset pairs as well as Pearson correlation coefficients between the two. Note that IoU and DICE are symmetrical metrics (do not depend on the order of $P$ and $Q$), while KL divergence is not. Non-symmetrical metric in our case is beneficial to use in order to distinguish between setups in which training and testing datasets switch places. For example, cross-dataset evaluation between PHEME4 and Twitter 15 leads to different outcomes for the model performance depending which dataset was used for training. In Table 10 the dataset shown on the left is used for training $Q$, and the one of the right for testing $P$. Within dataset performance evaluations follow cross-validation procedure, so the distance metrics shown in Table 10 are mean average of distance scores calculated for each fold.

We find a negative correlation between the model performance and distance between datasets for all metrics, i.e. the higher the distance, the lower the performance scores. Adding few-shot examples into the training somewhat decreases the distance between datasets in line with improvement in experimental results when using few-shot learning. This highlights that, indeed, the *vocabulary difference is an important factor for performance drop*. The correlation is strong for KL divergence (coefficient of −0.76), and only moderate for IoU and DICE (coefficient values around 0.3), therefore KL divergence is a more suitable metric and holds some predictive power of an expected performance drop.

### 5.5 Effect of conversation length

Investigating which data properties may be related to performance drop, another aspect to be explored is the role of conversation length across rumours. For example, Zubiaga, et al. (2016) found that, in the PHEME dataset, the longer the conversations are, the more likely they are to diverge from the original topic. By contrast Ma et al. (2018) and Bian, et al. (2020) show that for Twitter15/16 performance increases in time as more information becomes available. These contradictory observations indicate a difference in the speed of topic shift within conversations across different datasets, which may affect model performance. Such topic shifts should therefore be exploited both in training and in evaluating robust models.

We consider whether the lengths of the conversations in COVID-RV affect the performance of the models. For each instance in COVID-RV and for each model we calculate in how many training settings the model made a correct prediction, as a proxy for the instance 'difficulty', as well as the length and the depth of the corresponding conversation. We then calculate Pearson correlation coefficients between the conversation 'difficulty' and its length and depth. Table 11 shows these Pearson correlation coefficients for each of the models. We find that the correlation coefficients are very low and thus we cannot establish either a positive or negative effect of the conversation length on model performance in this case.

## 6 Discussion

This section discusses the implications of this research for future work on automated social media rumour verification. The main goals were to analyse whether automated rumour verification models would encounter generalisability issues (*RQ1*) and, if so, where the challenges lie such that future research directions can focus on those aspects (*RQ2-7*). *RQ2,3,7* involve the role of model architecture and understanding what properties of models may affect generalisability, while *RQ4-6* discuss the role of training data and setup.

***RQ1:*** *How well can rumour verification models generalise to unseen rumours across datasets from similar time periods and those more distant in time?* For *RQ1* we found a drastic performance drop for both types of cross-dataset experiments. This lack of generalisability undermines the practical value of rumour verification models and highlights the need for further efforts to create generalisable methods. The fact that the original performance of each model was overestimated highlights that none of the datasets can realistically represent the distribution of the unseen data.

*Discussion on the role of models in generalisability of rumour verification systems*

***RQ2:*** *If the models experience performance drop on unseen datasets, will the ranking of model performance evaluated across datasets align with the ranking of these models when evaluated within one dataset?* Considering *RQ2*, we discovered variability of model rankings when tested outside the original dataset. This suggests that developing models and evaluating them within a single dataset is not enough to ensure creation of generalisable approaches to rumour verification for real-world applications. It is crucial to test models on events unseen during training (see also *RQ6*) and to include cross-dataset evaluation. Novel techniques that are able to better promote generalisability of rumour verification models is an important research direction.

***RQ3:*** *Which models or groups of models show better generalisability?* There doesn't seem to be a model or group of models that consistently generalise better across datasets. However, in the experiments with COVID-RV, SAVED (the model that obtains conversations around rumours by disentangling the topic and the manner of speech) has shown promising results. For future work this suggests that we need to develop models that use or find generalisable features indicative of rumour veracity (potential examples could be user stance or propagation patterns). We should draw on developments in domain adaptation and generalisability domains (Ramponi & Plank, 2020; Wang, et al., 2021), incorporating generalisability tools and approaches into rumour verification models.

***RQ7:*** *How receptive are different models to strategies such as few-shot learning and using embeddings updated with data from a new rumour?* Our investigation of *RQ7* found that updating word embeddings and providing models with a few training examples from a new event (few-shot learning) helps bridge the gap between datasets distant in topic and time with BERT and SAVED models benefiting the most. Thus, updating and/or temporally aligning (Alkhalifa et al., 2021) embeddings may be important in improving performance across time. Ni et al. (2021) show that BERT fine-tuned for rumour detection cannot identify common sense rumours with more than 50% accuracy. Incorporating commonsense knowledge and other inductive biases along with few-shot learning could be fruitful avenues for improving generalisability in future research. Bragg, Cohan, Lo, and Beltagy (2021) introduce a few-shot NLP benchmark and provide recommendations for reliable few-shot evaluation, which can aid future work on developing strong few-shot learners. However, these are only available when there is access to some labelled or unlabelled data from the domain of interest.

*Discussion on the role of training data in generalisability of rumour verification systems*

***RQ4:*** *Which training datasets lead to better model generalisability? Does increasing the size of the training data improve generalisability?* The quality of the training data is one of the key elements for training a generalisable model (Wang, et al., 2021). In this work we performed experiments with the widely used benchmark datasets for automated rumour verification and we investigated the impact of different combinations of datasets in training. While we found that the aggregate of all datasets does not always lead to performance improvement, on average the best performance is indeed achieved by combining all three training datasets. A considerable limitation in rumour verification research is the small size of existing datasets. A possible way to address this would be to develop novel strategies for effectively leveraging combinations of existing datasets with differences in annotations, e.g., through transfer and/or multitask learning. Alternatively, creating synthetic data instances or whole datasets can be beneficial and also lead to modelling innovation, as shown in Liu, Lee, Jia, and Liang (2021). Overall, we did not find strong effects that would suggest that we should weigh the contributions of training data over model architecture or vice versa, thus we recommend that future research should look into both of these directions.

***RQ5:*** *Which data properties may be related to performance drop?* We have shown that the difference between training and testing datasets is an important factor in performance drop. We have measured the distance between dataset vocabularies and argue that metrics such as KL divergence can be potentially used to estimate expected performance drop of a model. Other metrics defining the 'distance' between benchmarks can be also explored, e.g. benchmark concurrence as defined in Liu et al. (2021). However, we recommend that other data properties should also be covered in future work. These could be model-specific and depend on the produced embeddings, e.g., for fake news detection, Zhou, et al. (2021) show that similarity between RoBERTa embeddings of article titles in training and testing datasets are correlated with performance. The differences in stance and propagation patterns could also be addressed more explicitly in future work given extra annotations, rather than implicitly through models utilising conversations.

An approach to assess model behaviour and thus reveal potential deficiencies in the data used for training them, is to use a checklist (Ribeiro, Wu, Guestrin, & Singh, 2020). A checklist is a set of unit tests to assess different aspects of model functionality.

Task-specific checklists can be created, e.g. Röttger, et al. (2021) created one for hate speech, which includes low level functional tests such as 'leet speak' as well as higher level test instances containing 'hate expressed using slurs' vs 'hate expressed using profanity'. Ni et al. (2021) show that rumour detection models can learn shortcuts due to spurious correlation between words and veracity labels in training datasets. Checklists could be used to identify such data artefacts and further expand training data by creating artificial and, perhaps, adversarial examples. In the case of rumour verification, handling negations correctly would be a very important test. Human-created negative variations of claims used in COVID-RV are made available by Hossain, et al. (2020). Additional negations can be created using checklists. Furthermore, Zhou, et al. (2021) find that unreliable news detection datasets can be biased by the ways they are curated, annotated, and split. Steps should be taken to identify and mitigate these biases in rumour verification datasets.

***RQ6:*** *Does the evaluation strategy used in the original training dataset play a role in training more generalisable models or more realistically estimating their future performance?* We find that indeed the evaluation strategy used in the original training dataset does play a role in realistically estimating model performance. The performance on the PHEME dataset is not overestimated to the same degree as it allows evaluation through a leave-one-event-out cross-validation setting. This enables a more challenging and realistic evaluation scenario, leading to lower but more reliable scores. Thus, here we highlight again that releasing datasets that cover multiple events is a good practice that should be followed in future work.

There are important limitations that make rumour verification a challenging task. Rumour conversations may not always contain sufficient information to support a veracity verdict. Models may rely on the stances of users or their choice of words (learning rules like "formal language is more trustworthy", or "expression of doubt is indicative of unverified rumours"), rather than evidence. We believe that this may be addressed by combining social media signals with signals extracted from a range of trusted sources, such as peer-reviewed publications, trustworthy news organisations, independent fact-checking organisations, etc. Finally, annotating datasets in such a way that helps models learn to identify and provide explanations for their predictions (rationales) is also important if they are to be trusted and thus be effective in real-world settings (Jain, Kumar, & Shrivastava, 2022).

## 7 Conclusions

We have evaluated, quantified and characterised the generalisability of social media rumour verification models in two settings: across datasets from similar time periods and on a newly collected COVID-19 dataset, distant in time from the training data. We have demonstrated a significant performance drop in both settings, which is further pronounced when datasets are distant in time. The extent of the divergence between training and testing datasets is analogous to the drop in performance. We found that few-shot learning and updating unsupervised embeddings with posts from the new events reduces the drop in performance. However, significant scope for improvement remains and, based on our findings, we have outlined directions for future work.

## Ethical considerations

This work involves ethical considerations concerning the spread of rumours and misinformation on social networks such as Twitter and Facebook. Although the systems analysed in this work are intended to prevent such information from being disseminated, the data we collect for evaluating these systems could potentially be mis-purposed by bad actors to adversarially construct misinformation that avoids detection.

Pending publication, the COVID-RV dataset will be released in compliance with the Twitter Developer guidelines, which require further compliance of all downstream users of our data. These guidelines include provisions that users of our dataset do not disclose the identities of Twitter users who have protected or deleted their accounts or tweets during or after data collection.

Annotations were collected using a combination of paid crowd workers and student volunteers. Crowd workers from Amazon Mechanical Turk were paid $2.05 per HIT (Human Intelligence Task), which was calculated using the wage of $11.93 per hour and our estimation of average time to complete the HIT, where each HIT included annotation of 15 claim–tweet pairs.

## CRediT authorship contribution statement

**Elena Kochkina:** Conceptualization, Methodology, Software, Data curation, Visualization, Writing – original draft. **Tamanna Hossain:** Software, Data curation, Writing – review & editing. **Robert L. Logan IV:** Software, Data curation, Writing – review & editing. **Miguel Arana-Catania:** Writing – review & editing. **Rob Procter:** Supervision, Writing – review & editing. **Arkaitz Zubiaga:** Supervision, Writing – review & editing. **Sameer Singh:** Resources, Supervision, Writing – review & editing. **Yulan He:** Funding acquisition, Supervision, Writing – review & editing. **Maria Liakata:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

## Data availability

Data will be made available on request.

## Acknowledgements

This work was supported by a UKRI/EPSRC grant (EP/V048597/1) to Profs Yulan He, Rob Procter and Maria Liakata, as well as project funding from the Alan Turing Institute, UK, grant EP/N510129/1. ML and YH are supported by Turing AI Fellowships (EP/V030302/1, EP/V020579/1). This material is based upon work sponsored in part by NSF award #IIS-1817183 and in part by the DARPA MCS program under Contract No. N660011924033 with the United States Office Of Naval Research.

We thank Arjuna Ugarte, Staff Researcher III at the University of California Irvine (UCI) Emergency Medicine & Informatics department for leading the stance annotation team. We also thank our stance annotators: UCI undergraduate students Ali Al-Hakeem, Sharon Li, Abhi Madduri, Juhi Patel, and Anam Zahidi; and Evergreen Valley High School, San Jose student Nitya Golla.

## Appendix A. CovidLies claim changes

The following changes were made to the claims (misconceptions) from COVIDLies:

- **Removal:**
  **Political:** Claims pertaining to the actions of particular political parties, governments, religious groups, or ethnicities, were removed. Eg. *'Trump is fulfilling his promise to hit Iranian cultural sites, if Iranians took revenge for the US airstrike that killed of Quds Force Commander Qasem Soleimani.'*
  **Multi-modal:** Claims about non-textual modalities, such as, images and videos were removed. Eg. *'Coronavirus is a state-supported "a bioweapon that went rogue" and also fake videos alleging that Chinese authorities are killing citizens to prevent its spread.'*
  **Duplicates:** De-duplication of claims was performed. Eg. *'Holy communion cannot be the cause of the spread of coronavirus'* was removed while *'Coronavirus cannot be spread by practicing holy communion.'* was kept.
- **Compound to atomic:** Compound claims were split into atomic misconceptions. Eg. *Avocado and mint tea, hot whiskey and honey, essential oils, vitamins c and d, fennel tea and cocaine cure coronavirus.* → *'Avocado and mint tea cures coronavirus.'*, *'Essential oils cure coronavirus.'*, *'Vitamin C cures coronavirus.'*, *'Vitamin D cures coronavirus.'*, *'Fennel tea cures coronavirus.'*, and *'Cocaine cures coronavirus.'*
- **Corrections:** Eg. *'There were more than 50000 cremations in Wuhan for 4th Quarter, 2020.'* → *'There were more than 50000 cremations in Wuhan for 4th Quarter, 2019.'*
- **Edits:** Eg.*'Chloroquine was used to cure over 12,000 covid-19 patients.'* → *'Chloroquine can cure coronavirus.'*

## Appendix B. Relevance annotation instructions

As qualification criteria, we asked that workers must: be from an English-speaking country[19]; have completed more than 100 assignments; have assignment approval rate over 95%. Furthermore, workers had to pass a qualification test by annotating 7 claim–tweet pairs from the CovidLies dataset and get over 70% of correct answers. We recruited 3 annotators for each claim–tweet instance.

Fig. A1 shows the relevance annotation interface.

**Instructions**

**Goals.** You will be given a claim and a list of tweets. Your task is to tag whether the tweet is relevant to the given claim, i.e. whether the tweet and the claim are closely connected.

The relevance of each tweet should be judged based on its connection to the specific claim rather than the general topic of coronavirus COVID-19.

Relevant tweets do not have to agree with the claim, but they have to discuss the main topic of the claim. Tweets conveying sarcastic or joking sentiment should not be excluded as they could also be relevant.

Relevance should not be judged on the basis of the annotator's personal views towards the topic.

Choosing a "relevant" or "not relevant" option for each tweet is compulsory.

**External Links.** If a tweet contains a link, the contents of the link can be used to identify the relevance of the tweet to the claim. To access the link copy-paste it into your browser. If you opened a link, please tick the box under the tweet to indicate that. Please be advised that the Requester has not checked the content of links included in the tweets.

**Issues and Comments.** If there is an issue with annotating a claim–tweet pair, there is an option to flag it using the checkbox in the right bottom corner of the box bounding each tweet. This includes blank text, text in a different language and inappropriate content. If you have any comment regarding the task or a particular instance there is a free text comment form at the bottom of the page.

**Examples.**

(1)

Claim: Coronavirus is caused by 5G.

[19] ISO 3166 country codes: GB, US, AG, AT, BS, BB, BZ, CA, DM, GD, GY, IE, JM, MT, NZ, TT, LC, VC, KN.

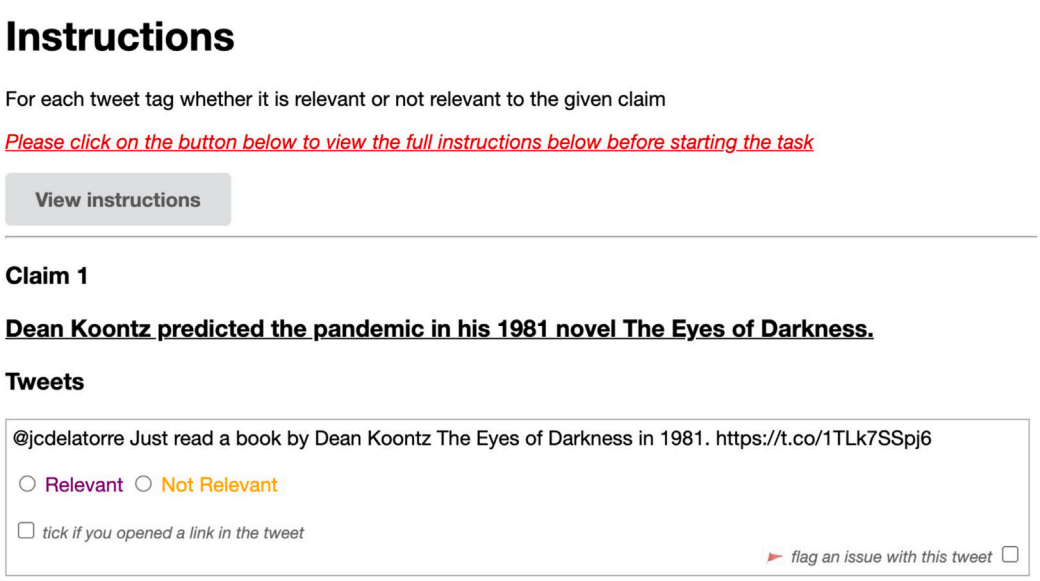


**Fig. A1.** Relevance annotation interface.

Tweet: Expand your thinking. Optogenetic tech is literal mind control via *light* combined with a GMO receptor cell which can be introduced into the body on the back of a *virus.* 5G transmit/receive. Coronavirus + 5G in Wuhan. ?? Experiment? Gone wrong?

Tag: Relevant

Explanation: The tweet supports the idea that 5G transmits coronavirus

(2)

Claim: Coronavirus is caused by 5G.

Tweet: 5G does not cause coronavirus, the theory is hilariously bad, but we'll explain why it's bad: Coronavirus spreads from human to human, that's why social distancing stopped the disease in South Korea. Also evil phone companies wouldn't implement something harmful for themselves.

Tag: Relevant

Explanation: The tweet explains that coronavirus is not caused by 5G

(3)

Claim: Coronavirus is caused by 5G.

Tweet: Coronavirus causes 'upheaval and uncertainty' for toy manufacturers

Tag: Not Relevant

Explanation: The tweet talks about toy manufacturers, not 5G being the cause of COVID-19

(4)

Claim: Coronavirus is caused by 5G.

Tweet: Vodafone to remove Huawei from core of its European network after UK decision to restrict Chinese company role in 5G Huawei is telecom coronavirus,you never know which part being infiltrated unless it is banned from all systems.

Tag: Not Relevant

Explanation: The tweet talks about Huawei and 5G, not the causes of COVID-19

## Appendix C. Stance annotation instructions

Fig. A2 shows the stance annotation interface.

**Goals.** You will be given a claim and a list of tweets. Your task is to tag the stance of the tweet towards the given claim, i.e. whether the tweet is agreeing with the topic of the given claim (Agrees), disagreeing with it (Disagrees), or expressing no stance towards it (No Stance).

Stance should not be judged on the basis of the annotator's personal views towards the topic.

Choosing a Agrees, Disagrees, or No Stance option for each tweet is compulsory.

When selecting a label of No Stance you must provide an explanation for why you think the tweet expresses no stance towards the claim.

**Questions.** Tweets that are questions can be tricky. Carefully consider your selected label in these cases.

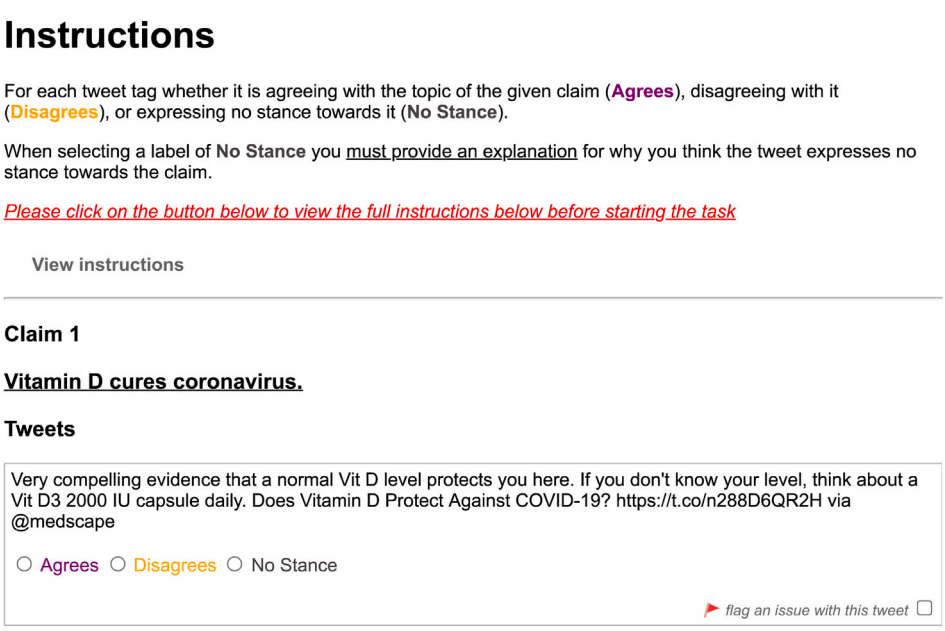


**Fig. A2.** Stance annotation interface.

- If the tweet seems like a genuine question for information on the claim, without any lean towards a particular stance (agreeing or disagreeing) then select No Stance and note that the tweet was a question in your Explanation. Eg.
  *Claim:*Hand sanitiser sold commercially does not destroy coronavirus.
  *Tweet:* Does hand sanitizer stop covid?
  *Tag:* No Stance
  *Explanation:* Question about the claim without expressing a stance.
- If the tweet is a question but has a lean towards agreeing with the given claim then select Agrees. Eg.
  *Claim:*Hand sanitiser sold commercially does not destroy coronavirus.
  *Tweet:* Are you sure hand sanitizer is killing covid and not your brain cells?
  *Tag:* Agrees
  *Explanation:* Question but doubting whether hand sanitizer destroys COVID-19, i.e., leaning towards agreeing with the claim.
- If the tweet is a question but has a lean towards disagreeing with the given claim then select Disagrees. Eg.
  *Claim:*Hand sanitiser sold commercially does not destroy coronavirus.
  *Tweet:* Which brand of hand sanitizer is most effective at killing covid?
  *Tag:* Disagrees
  *Explanation:* Question about the relative effectiveness of different brands implies belief in general effectiveness of hand sanitizers in destroying COVID-19.

**Sarcasm/Humour.** If you think the tweet is sarcastically or humorously agreeing with the given claim then select Disagrees. If you think the tweet is sarcastically or humorously disagreeing with the given claim then select Agrees.

**Issues and Comments.** If there is an issue with annotating a claim–tweet pair, there is an option to flag it using the checkbox in the right bottom corner of the box bounding each tweet. This includes blank text, text in a different language and inappropriate content. If you have any comment regarding the task or a particular instance there is a free text comment form at the bottom of the page.

**Examples.**

(1)

*Claim:*Hand sanitiser sold commercially does not destroy coronavirus.

*Tweet:* Hand washing is NOT the same as hand sanitizer. ONLY washing your hands with soap and water for 30 s will kill coronavirus

*Tag:* Agree

*Explanation:* The tweet supports the idea that hand sanitiser sold commercially does not destroy coronavirus.

(2)

*Claim:*COVID-19 is only as deadly as the seasonal flu.

*Tweet:* The SCIENCE behind this virus is all that matters. COVID-19 is estimated at 2–3

0.1% vs 1% = COVID-19 is 10x deadlier 0.1% vs 2% = 20x deadlier 0.1% vs 3% = 30x If COVID-19 is 2X as infectious as the flu, factors go to 20x, 40x, 60x etc.

*Tag:* Disagree

*Explanation:* The tweet refutes the idea that COVID-19 is only as deadly as the seasonal flu.

(3)

*Claim:*Dean Koontz predicted the pandemic in his 1981 novel The Eyes of Darkness.

*Tweet:* Just read a book by Dean Koontz The Eyes of Darkness in 1981.

*Tag:* No Stance

*Explanation:* The tweet does not connect Dean Koontz's The Eye of Darkness to the COVID-19 pandemic.

## Appendix D. Model hyper-parameters

*branchLSTM* (Kochkina & Liakata, 2020; Kochkina et al., 2018)

Number of LSTM layers: 2
Number of neurons in LSTM layer: 300
Number of Dense layers: 1
Number of neurons in Dense layer: 400
Dropout: 0.1
Optimiser: Adam
Batch size: 32
Number of epochs: 150
Learning rate: 0.001

*Top-Down Recursive Neural Networks (TD-RvNN)* (Ma et al., 2018)

Number of neurons in a layer: 100
Optimiser: Ada-grad
Batch size: 1
Max Number of epochs: 600 or until the loss value converges
Learning rate: 0.005

*Bidirectional Graph Convolutional Neural Network (BiGCN)* (Bian, et al., 2020)

Number of neurons in layer: 64
Optimiser: Adam
Batch size: 128
Number of epochs: 200
Learning rate: 0.0005
Weight decay: 1e−4
Patience: 10
Dropping rate in DropEdge: 0.2
Dropout: 0.5

*Stance-Augmented VAE Disentanglement framework (SAVED)* (Dougrez-Lewis et al., 2021)

Number of stance-dependent topics: 6
Number of stance-independent topics: 10
Number of neurons in layer: 400
Optimiser: Adam
Batch size: 12
Number of epochs: 200
Learning rate: 0.001
Output dropout rate: 0.2

*Large pre-trained language models BERT* (Devlin et al., 2019)

Optimiser: AdamW
Batch size: 8
Number of epochs: 10
Learning rate: 5e−5

We run our experiments using GPUs on the Microsoft Azure Cloud Computing service. We use Data Science Virtual Machines of size NC6, NC12 and NC24.[20]

[20] Azure Virtual Machine sizes are described in detail here: https://azure.microsoft.com/en-gb/pricing/details/virtual-machines/series/.

**Table A.1**
Binary classification results for within dataset experiments.

| In-dataset | PHEME | | Tw15 | | Tw16 | |
|---|---|---|---|---|---|---|
| | Acc. | MF | Acc. | MF | Acc. | MF |
| SAVED | 0.592 | **0.558** | 0.851 | 0.851 | 0.862 | 0.835 |
| BERT | **0.685** | 0.516 | 0.927 | 0.927 | **0.941** | **0.941** |
| CT-SAVED | 0.509 | 0.469 | 0.739 | 0.739 | 0.716 | 0.679 |
| CT-BERT | 0.672 | 0.478 | **0.949** | **0.949** | 0.933 | 0.933 |

**Table A.2**
Binary classification results on COVID-RV.

| | PHEME2 | | Tw15 | | Tw16 | | Tw15+Tw16 | | Tw15+Tw16+PHEME2 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. | MF | Acc. | MF | Acc. | MF | Acc. | MF | Acc. | MF |
| SAVED | **0.525** | **0.461** | 0.559 | 0.434 | 0.586 | 0.38 | 0.548 | 0.402 | 0.521 | **0.520** |
| BERT | 0.418 | 0.358 | 0.448 | 0.411 | 0.596 | 0.423 | 0.575 | 0.439 | **0.606** | 0.433 |
| CT-SAVED | 0.397 | 0.345 | 0.507 | **0.497** | **0.604** | 0.454 | **0.588** | **0.454** | 0.533 | 0.507 |
| CT-BERT | 0.394 | 0.34 | **0.601** | 0.375 | **0.604** | **0.471** | 0.567 | 0.408 | 0.398 | 0.285 |

## Appendix E. Binary classification experiments

While we consider the setup in which the existing, pre-trained models are applied to the new data regardless of the number of original labels, in the interest of approximating a realistic scenario, this setup may also lead to an underestimated performance. To address concerns about the mismatch between the number of labels in the training and testing sets we have performed binary classification experiments to compare performance between setups: (1) within dataset cross-validation and (2) training on the PHEME, Twitter 15 and Twitter 16 datasets and their combinations, then testing on COVID-RV. We have selected a subset of the models to perform these additional experiments with due to computational constraints, namely: SAVED, BERT and their COVID-aware variations CT-SAVED and CT-BERT. These models yielded the highest performance for four class classification on COVID-RV (see Table 6). The results are presented in Table A.1 for within dataset experiments and Table A.2 for results when testing the models across datasets on COVID-RV. We find that these results align with the conclusions from Tables 5 and 6. The PHEME dataset provides a more challenging scenario than Twitter 15 and Twitter 16 in the within dataset setting, thus the performance is not as strongly overestimated when evaluating on COVID-RV. CT-BERT and CT-SAVED show better performance than their COVID unaware versions in most cases on COVID-RV. Model performance benefits from larger training data size.

## References


Alkhalifa, R., Kochkina, E., & Zubiaga, A. (2021). Opinions are made to be changed: Temporally adaptive stance classification. In *Proceedings of the 2021 workshop on open challenges in online social networks* (pp. 27–32).

Arana-Catania, M., Kochkina, E., Zubiaga, A., Liakata, M., Procter, R., & He, Y. (2022). Natural language inference with self-attention for veracity assessment of pandemic claims. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 1496–1511). Seattle, United States: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.naacl-main.107, URL https://aclanthology.org/2022.naacl-main.107.

Bian, T., Xiao, X., Xu, T., Zhao, P., Huang, W., Rong, Y., et al. (2020). Rumor detection on social media with bi-directional graph convolutional networks. In *The thirty-fourth AAAI conference on artificial intelligence, AAAI 2020, the thirty-second innovative applications of artificial intelligence conference, IAAI 2020, the tenth AAAI symposium on educational advances in artificial intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020* (pp. 549–556). AAAI Press.

Bragg, J., Cohan, A., Lo, K., & Beltagy, I. (2021). Flex: Unifying evaluation for few-shot nlp. *Advances in Neural Information Processing Systems*, *34*.

Chen, E., Lerman, K., Ferrara, E., et al. (2020). Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, *6*(2), Article e19273.

Chen, X., Zhou, F., Zhang, F., & Bonsangue, M. (2021). Catch me if you can: A participant-level rumor detection framework via fine-grained user representation learning. *Information Processing & Management*, *58*(5), Article 102678.

Cheng, M., Wang, S., Yan, X., Yang, T., Wang, W., Huang, Z., et al. (2021). A COVID-19 rumor dataset. *Frontiers in Psychology*, *12*.

Cui, L., & Lee, D. (2020). Coaid: Covid-19 healthcare misinformation dataset. arXiv preprint arXiv:2006.00885.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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. Vol. 1* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1423.

Dharawat, A., Lourentzou, I., Morales, A., & Zhai, C. (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.

Dougrez-Lewis, J., Liakata, M., Kochkina, E., & He, Y. (2021). Learning disentangled latent topics for Twitter rumour veracity classification. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021* (pp. 3902–3908). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.findings-acl.341, Online.

Ettinger, A., Rao, S., Daumé, H., & Bender, E. M. (2017). Towards linguistically generalizable NLP systems: A workshop and shared task. In *Proceedings of the first workshop on building linguistically generalizable NLP systems* (pp. 1–10). Copenhagen, Denmark: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W17-5401.

Gao, J., Han, S., Song, X., & Ciravegna, F. (2020). RP-DNN: A tweet level propagation context based deep neural networks for early rumor detection in social media. In *Proceedings of the 12th language resources and evaluation conference* (pp. 6094–6105). Marseille, France: European Language Resources Association.

Glandt, K., Khanal, S., Li, Y., Caragea, D., & Caragea, C. (2021). Stance detection in COVID-19 tweets. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing* (pp. 1596–1611).

Gorrell, G., Kochkina, E., Liakata, M., Aker, A., Zubiaga, A., Bontcheva, K., et al. (2019). SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 845–854).

Graves, L., & Mantzarlis, A. (2020). Amid political spin and online misinformation, fact checking adapts. *The Political Quarterly*, *91*(3), 585–591.

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security* (pp. 2–12).

Hossain, T., Logan IV, R. L., Ugarte, A., Matsubara, Y., Young, S., & Singh, S. (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*. Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.nlpcovid19-2.11, Online.

Hou, Y., van der Putten, P., & Verberne, S. (2022). The COVMis-stance dataset: Stance detection on Twitter for COVID-19 misinformation. arXiv preprint arXiv:2204.02000.

Huang, Q., Yu, J., Wu, J., & Wang, B. (2020). Heterogeneous graph attention networks for early detection of rumors on twitter. In *2020 International joint conference on neural networks* (pp. 1–8). IEEE.

Huang, Q., Zhou, C., Wu, J., Wang, M., & Wang, B. (2019). Deep structure learning for rumor detection on twitter. In *2019 International joint conference on neural networks* (pp. 1–8). IEEE.

Islam, M. S., Sarkar, T., Khan, S. H., Kamal, A.-H. M., Hasan, S. M., Kabir, A., et al. (2020). COVID-19–related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene*, *103*(4), 1621.

Jain, D. K., Kumar, A., & Shrivastava, A. (2022). CanarDeep: a hybrid deep neural model with mixed fusion for rumour detection in social data streams. *Neural Computing and Applications*, 1–12.

Karafillakis, E., Van Damme, P., Hendrickx, G., & Larson, H. J. (2022). COVID-19 in Europe: New challenges for addressing vaccine hesitancy. *The Lancet*, *399*(10326), 699–701.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., et al. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 6769–6781). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.550, Online.

Khoo, L. M. S., Chieu, H. L., Qian, Z., & Jiang, J. (2020). Interpretable rumor detection in microblogs by attending to user interactions. In *The thirty-fourth AAAI conference on artificial intelligence, AAAI 2020, the thirty-second innovative applications of artificial intelligence conference, IAAI 2020, the tenth AAAI symposium on educational advances in artificial intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020* (pp. 8783–8790). AAAI Press.

Kochkina, E., & Liakata, M. (2020). Estimating predictive uncertainty for rumour verification models. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 6964–6981). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.623, Online.

Kochkina, E., Liakata, M., & Augenstein, I. (2017). Turing at SemEval-2017 task 8: Sequential approach to rumour stance classification with branch-LSTM. In *Proceedings of the 11th international workshop on semantic evaluation* (pp. 475–480). Vancouver, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/S17-2083.

Kochkina, E., Liakata, M., & Zubiaga, A. (2018). All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th international conference on computational linguistics* (pp. 3402–3413). Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Kumar, S., & Carley, K. (2019). Tree LSTMs with convolution units to predict stance and rumor veracity in social media conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 5047–5058). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-1498.

Kumari, R., Ashok, N., Ghosal, T., & Ekbal, A. (2022). What the fake? Probing misinformation detection standing on the shoulder of novelty and emotion. *Information Processing & Management*, *59*(1), Article 102740.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., et al. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, *7*, 452–466. http://dx.doi.org/10.1162/tacl_a_00276.

Lee, N., Li, B. Z., Wang, S., Fung, P., Ma, H., Yih, W.-t., et al. (2021). On unifying misinformation detection. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 5479–5485). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.naacl-main.432, Online.

Li, Y., Fan, Z., Yuan, X., & Zhang, X. (2022). Recognizing fake information through a developed feature scheme: A user study of health misinformation on social media in China. *Information Processing & Management*, *59*(1), Article 102769.

Li, Y., Jiang, B., Shu, K., & Liu, H. (2020). MM-COVID: A multilingual and multimodal data repository for combating COVID-19 disinformation. arXiv preprint arXiv:2011.04088.

Liu, N. F., Lee, T., Jia, R., & Liang, P. (2021). Can small and synthetic benchmarks drive modeling innovation? a retrospective study of question answering modeling approaches. arXiv preprint arXiv:2102.01065.

Liu, X., Nourbakhsh, A., Li, Q., Fang, R., & Shah, S. (2015). Real-time rumor debunking on Twitter. In J. Bailey, A. Moffat, C. C. Aggarwal, M. de Rijke, R. Kumar, V. Murdock, T. K. Sellis, & J. X. Yu (Eds.), *Proceedings of the 24th ACM international conference on information and knowledge management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015* (pp. 1867–1870). ACM, http://dx.doi.org/10.1145/2806416.2806651.

Lu, J., Henchion, M., & Mac Namee, B. (2021). Diverging divergences: Examining variants of Jensen Shannon divergence for corpus comparison tasks.

Lukasik, M., Bontcheva, K., Cohn, T., Zubiaga, A., Liakata, M., & Procter, R. (2019). Gaussian processes for rumour stance classification in social media. *ACM Transactions on Information Systems (TOIS)*, *37*(2), 1–24.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K., et al. (2016). Detecting rumors from microblogs with recurrent neural networks. In S. Kambhampati (Ed.), *Proceedings of the twenty-fifth international joint conference on artificial intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016* (pp. 3818–3824). IJCAI/AAAI Press.

Ma, J., Gao, W., & Wong, K.-F. (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* (pp. 708–717). Vancouver, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P17-1066.

Ma, J., Gao, W., & Wong, K.-F. (2018). Rumor detection on Twitter with tree-structured recursive neural networks. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (pp. 1980–1989). Melbourne, Australia: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P18-1184.

Marasović, A. (2018). NLP's generalization problem, and how researchers are tackling it. *The Gradient*.

Memon, S. A., & Carley, K. M. (2020). Characterizing COVID-19 misinformation communities using a novel twitter dataset. In *CEUR workshop proceedings. Vol. 2699*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Moore, A., & Rayson, P. (2018). Bringing replication and reproduction together with generalisability in NLP: Three reproduction studies for target dependent sentiment analysis. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1132–1144). Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Müller, M., Salathé, M., & Kummervold, P. E. (2020). Covid-twitter-bert: A natural language processing model to analyse COVID-19 content on twitter. arXiv preprint arXiv:2005.07503.

Ni, S., Li, J., & Kao, H.-Y. (2021). True or false: Does the deep learning model learn to detect rumors? In *2021 International conference on technologies and applications of artificial intelligence* (pp. 119–124). IEEE.

Nogueira, R., Lin, J., & Epistemic, A. (2019). From doc2query to docTTTTTquery. Online Preprint.

Peinelt, N., Liakata, M., & Nguyen, D. (2019). Aiming beyond the obvious: Identifying non-obvious cases in semantic similarity datasets. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2792–2798).

Pian, W., Chi, J., & Ma, F. (2021). The causes, impacts and countermeasures of COVID-19 "Infodemic": A systematic review using narrative synthesis. *Information Processing & Management*, *58*(6), Article 102713.

Ramponi, A., & Plank, B. (2020). Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th international conference on computational linguistics* (pp. 6838–6855). Barcelona, Spain: International Committee on Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.coling-main.603, (Online).

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4902–4912). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.442, Online.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M., et al. (1995). Okapi at TREC-3. In *Nist special publication Sp. Vol. 109* (p. 109). National Instiute of Standards & Technology.

Röttger, P., & Pierrehumbert, J. B. (2021). Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. arXiv preprint arXiv:2104.08116.

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2021). HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing* (pp. 41–58).

Roy, S., Bhanu, M., Saxena, S., Dandapat, S., & Chandra, J. (2022). gDART: Improving rumor verification in social media with discrete attention representations. *Information Processing & Management*, *59*(3), Article 102927.

Shaar, S., Nikolov, A., Babulkov, N., Alam, F., Barrón-Cedeno, A., Elsayed, T., et al. (2020). Overview of CheckThat! 2020 english: Automatic identification and verification of claims in social media. In *CLEF*.

Shahi, G. K., & Nandini, D. (2020). FakeCovid–A multilingual cross-domain fact check news dataset for COVID-19. In *Workshop on cyber social threats (CySoc 2020) at 14th international conference on web and social media 2020*.

Shu, K., Bhattacharjee, A., Alatawi, F., Nazer, T. H., Ding, K., Karami, M., et al. (2020). Combating disinformation in a social media age. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(6), Article e1385.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track*.

Tu, K., Chen, C., Hou, C., Yuan, J., Li, J., & Yuan, X. (2021). Rumor2vec: a rumor detection framework with joint text and propagation structure representation learning. *Information Sciences*, *560*, 137–151.

Wakamiya, T. M. S., & Aramaki, E. (2020). Fake news detection using temporal features extracted via point process. In *Proceedings of the workshop on cyber social threats*.

Wang, J., Lan, C., Liu, C., Ouyang, Y., Zeng, W., & Qin, T. (2021). Generalizing to unseen domains: A survey on domain generalization. arXiv preprint arXiv:2103.03097.

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., 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*. Association for Computational Linguistics, Online.

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, *53*(3), 1–34.

Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, *7*, Article e598.

Yuan, C., Ma, Q., Zhou, W., Han, J., & Hu, S. (2019). Jointly embedding the local and global relations of heterogeneous graph for rumor detection. In *2019 IEEE international conference on data mining* (pp. 796–805). IEEE.

Zhang, X., Cao, J., Li, X., Sheng, Q., Zhong, L., & Shu, K. (2021). Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021* (pp. 3465–3476).

Zhang, Q., Cook, J., & Yilmaz, E. (2021). Detecting and forecasting misinformation via temporal and geometric propagation patterns. In *ECIR* (2), (pp. 455–462).

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *8th International conference on learning representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhou, X., Elfardy, H., Christodoulopoulos, C., Butler, T., & Bansal, M. (2021). Hidden biases in unreliable news detection datasets. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume* (pp. 2482–2492).

Zhou, X., Mulay, A., Ferrara, E., & Zafarani, R. (2020). Recovery: A multimodal repository for COVID-19 news credibility research. In M. d'Aquin, S. Dietze, C. Hauff, E. Curry, & P. Cudré-Mauroux (Eds.), *CIKM '20: The 29th ACM international conference on information and knowledge management, virtual event, Ireland, October 19-23, 2020* (pp. 3205–3212). ACM, http://dx.doi.org/10.1145/3340531.3412880.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys*, *51*(2), 1–36.

Zubiaga, A., Kochkina, E., Liakata, M., Procter, R., Lukasik, M., Bontcheva, K., et al. (2018). Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing & Management*, *54*(2), 273–290.

Zubiaga, A., Liakata, M., Procter, R., Hoi, G. W. S., & Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS One*, *11*(3), Article e0150989.

**Dr. Elena Kochkina** Postdoctoral Researcher at the Queen Mary University of London and the Alan Turing Institute. Elena have completed a PhD in Computer Science at the Warwick Institute for the Science of Cities (WISC) CDT, funded by the Leverhulme Trust via the Bridges Programme. Her background is Applied Mathematics (BSc, MSc, Lobachevsky State University of Nizhny Novgorod) and Complexity Science (MSc, University of Warwick, Chalmers University). She has published in venues such as ACL, COLING, EACL and IP&M.

**Tamanna Hossain** Ph.D. student at the University of California, Irvine (UCI) Computer Science department. Her background is in Statistics (MSc, Oklahoma State University), Mathematics and Philosophy (BA, Lawrence University).

**Robert L. Logan IV** Ph.D. candidate at the University of California, Irvine, and fellow of the Irvine Initiative in AI, Law, and Society. His research focuses on addressing knowledge deficiencies in natural language processing models, and has been published in venues such as AAAI, ACL, EMNLP, JAMIA, and NAACL.

**Dr. Miguel Arana-Catania** Postdoctoral Researcher at the University of Warwick, working in NLP research projects in collaboration with the Alan Turing Institute. His background is in Theoretical Physics from the Universidad Autónoma de Madrid (UAM) and the Institute for Theoretical Physics (IFT) UAM-CSIC.

**Prof. Rob Procter** Professor of Social Informatics at the University of Warwick, co-chair of the Department's AI & Human-Centered Computing research theme and a fellow of the UK Alan Turing Institute for Data Science and AI, where he co-chairs the social data science interest group, which is dedicated to the development of robust data science methodologies and tools for social research. He has published over 300 papers (including 100+ journal articles) and was editor for the Health Informatics Journal from 2004-2020.

**Dr. Arkaitz Zubiaga** Senior Lecturer at the Queen Mary University of London, where he leads the Social Data Science lab. His research interests revolve around linking online data with events in the real world, among others for tackling problematic issues on the Web and social media that can have a damaging effect on individuals or society at large, such as hate speech, misinformation, inequality, biases and other forms of online harm. He has published over 130 research papers (including 40+ journal articles), and serves as academic editor for six journals.

**Prof. Sameer Singh** Associate Professor of Computer Science at the University of California, Irvine (UCI). He is working primarily on robustness and interpretability of machine learning algorithms, along with models that reason with text and structure for natural language processing. He has received the NSF CAREER award, selected as a DARPA Riser, UCI Distinguished Early Career Faculty award, and the Hellman Faculty Fellowship. Sameer has published extensively at machine learning and natural language processing venues and received conference paper awards at KDD 2016, ACL 2018, EMNLP 2019, AKBC 2020, and ACL 2020.

**Prof. Yulan He** Professor at the Department of Computer Science in the University of Warwick, UK. She is currently holding a prestigious UKRI Turing AI Fellowship (2021-2025). Yulan's research interests lie in the integration of machine learning and natural language processing for text analytics. She has published over 200 papers on topics including natural language understanding, sentiment analysis, topic and event extraction, question-answering, and fake news detection. She is an Action Editor for Transactions of the ACL and an Associate Editor for the Royal Society Open Science journal.

**Prof. Maria Liakata** is a Professor in Natural Language Processing (NLP) at the School of Electronic Engineering and Computer Science, Queen Mary University of London and Honorary Professor at the Department of Computer Science, University of Warwick. She holds a prestigious EPSRC/UKRI Turing AI fellowship (2020-2025) on Creating time sensitive sensors from user-generated language and heterogeneous content. At the Alan Turing Institute she founded and co-leads the NLP and data science for mental health special interest groups and supervises PhD students. She leads a team of 14 researchers. She has published over 140 papers on topics including sentiment analysis, semantics, summarisation, rumour verification, resources & evaluation and biomedical NLP. She is action editor for the ACL rolling review and has had numerous senior roles in conference and workshop organisation.