RumourEval 2019: Determining Rumour Veracity and Support for Rumours

Genevieve Gorrell[1] and **Elena Kochkina**[3,4] and **Maria Liakata**[3,4] and **Ahmet Aker**[1] and **Arkaitz Zubiaga**[5] and **Kalina Bontcheva**[1] and **Leon Derczynski**[2]

[1]University of Sheffield, UK
(g.gorrell,a.aker,k.bontcheva)@sheffield.ac.uk

[2]IT University of Copenhagen, Denmark
ld@itu.dk

[3]University of Warwick, UK
(e.kochkina,m.liakata)@warwick.ac.uk
[4]Alan Turing Institute, UK

[5]Queen Mary University of London, UK
a.zubiaga@qmul.ac.uk

Abstract

Since the first RumourEval shared task in 2017, interest in automated claim validation has greatly increased, as the danger of "fake news" has become a mainstream concern. However automated support for rumour verification remains in its infancy. It is therefore important that a shared task in this area continues to provide a focus for effort, which is likely to increase. Rumour verification is characterised by the need to consider evolving conversations and news updates to reach a verdict on a rumour's veracity. As in RumourEval 2017 we provided a dataset of dubious posts and ensuing conversations in social media, annotated both for stance and veracity. The social media rumours stem from a variety of breaking news stories and the dataset is expanded to include Reddit as well as new Twitter posts. There were two concrete tasks; rumour stance prediction and rumour verification, which we present in detail along with results achieved by participants. We received 22 system submissions (a 70% increase from RumourEval 2017) many of which used state-of-the-art methodology to tackle the challenges involved.

1 Introduction

1.1 Background

Since the first RumourEval shared task in 2017 (Derczynski et al., 2017), interest in automated verification of rumours has deepened, as research has demonstrated the potential impact of false claims on important political outcomes (Allcott and Gentzkow, 2017). Living in a "post-truth world", in which perceived truth can matter more than actual truth (Dale, 2017), the dangers posed by unchecked market forces and cheap platforms, as well as poor ability by many readers to discern credible information, are evident. As a result the importance of educating young people about critical thinking is increasingly emphasised.[1]. Moreover the European Commission's High Level Expert Group on Fake News provides tools to empower users and journalists to tackle disinformation as one of the five pillars of their recommended approach.[2] Platforms are increasingly motivated to engage with the problem of damaging content that appears on them, as society moves toward a consensus regarding their level of responsibility. Independent fact checking efforts, such as Snopes[3], Full Fact[4], Chequeado[5], are also becoming valued resources (Konstantinovskiy et al., 2018). Zubiaga et al. (2018) present an extensive list of projects. Effort so far is often manual, and struggles to keep up with the large volume of online material.

Within NLP research the tasks of stance classification of news articles and social media posts and the creation of systems to automatically identify false content are gaining momentum. Work in credibility assessment has been around since 2011 (Castillo et al., 2011), making use initially

[1]http://www.bbc.co.uk/mediacentre/latestnews/2017/fake-news
[2]http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50271
[3]https://www.snopes.com/
[4]https://fullfact.org/
[5]http://chequeado.com/

Veracity prediction. Example 1:

u1: Hostage-taker in supermarket siege killed, reports say. #ParisAttacks LINK [true]

Veracity prediction. Example 2:

u1: OMG. #Prince rumoured to be performing in Toronto today. Exciting! [false]

Table 1: Examples of source tweets with veracity value

of local features. Fact checking is a broad complex task, challenging the resourcefulness of even a human expert. Claims such as "we send the EU 350 million a week" which is partially true would need to be decomposed into statements to be checked against knowledge bases and multiple sources. Ways of automating fact checking has inspired researchers (Vlachos and Riedel, 2015) and has resulted in a new shared task FEVER.[6] Other research has focused on stylistic tells of untrustworthiness in the source itself (Conroy et al., 2015; Singhania et al., 2017). Rumour verification is a particular case of fact checking. Rumours are *"circulating stories of questionable veracity, which are apparently credible but hard to verify, and produce sufficient skepticism and/or anxiety so as to motivate finding out the actual truth"* (Zubiaga et al., 2016). One can distinguish several component to a rumour resolution pipeline such as rumour detection, rumour tracking and stance classification, leading to the final outcome of determining the veracity of a rumour (Zubiaga et al., 2018). Thus what characterises rumour verification compared to other types of fact checking is time sensitivity and the importance of dynamic interactions between users, their stance and information propagation. Initial work on rumour detection and stance classification (Qazvinian et al., 2011) was succeeded by more elaborate systems and annotation schemas (Kumar and Geethakumari, 2014; Zhang et al., 2015; Shao et al., 2016; Zubiaga et al., 2016). Vosoughi (2015) demonstrated the value of making use of propagation information, i.e. the ensuing discussion, in rumour verification. Stance detection is the task of classifying a text according to the position it takes with respect to a statement. Research supports the importance of this subtask as a first step to veracity identification. (Ferreira and Vlachos, 2016; Enayet and El-Beltagy, 2017). Crowd response, stance and the details of rumour propagation feature in the work by Chen et al. (2016) as well as the most successful system in RumourEval 2017 (Enayet and El-Beltagy, 2017), and the highest performing systems in RumourEval 2019.

1.2 Datasets for rumour verification

The UK fact-checking charity Full Fact provides a roadmap[7] for development of automated fact checking. They cite open and shared evaluation as one of their five principles for international collaboration, demonstrating the continuing relevance of shared tasks in this area. Shared datasets are a crucial part of the joint endeavour. Datasets for rumour resolution are still relatively few, and likely to be in increasing demand. In addition to the data from RumourEval 2017, the dataset released by Kwon et al. (2017) is also suitable for veracity classification. It includes 51 true rumours and 60 false rumours, where each rumour includes a stream of tweets associated with it. Twitter 15 and 16 datasets (Ma et al., 2018) contain claim propagation trees and combine tasks of rumour detection and verification in one four-way classification task (Non-rumour, True, False, Unverified). A Sina Weibo corpus is also available (Wu et al., 2015), in which 5000 posts are classified for veracity, but responses are not available. Partially generated statistical claim checking data is now becoming available in the context of the FEVER shared task, mentioned above, but is not suitable for this type of work. Twitter continues to be a highly relevant platform for rumour verification, being popular with the public as well as politicians. RumourEval 2019 also includes Reddit

[6] https://sheffieldnlp.github.io/fever/

[7] https://fullfact.org/media/uploads/full_fact-the_state_of_automated_factchecking_aug_2016.pdf

data, thus providing more diversity in the types of users, more focussed discussions and longer texts.

1.3 RumourEval 2017 vs 2019

RumourEval 2019 furthers progress on stance detection and rumour verification, both still unbested NLP tasks. They are currently moderately well performed for English short texts (tweets), with data existing in a few other languages (notably as part of IberEval). In 2019, many more teams took part, demonstrating the rising relevance of the tasks. Specifically, as in 2017, RumourEval 2019 comprises two subtasks:

- In subtask A, given a source tweet, tweets in a conversation thread discussing the claim are classified as either supporting, denying, querying or commenting on the rumour mentioned by the source tweet
- In subtask B, the rumour introduced by the source tweet that spawned the discussion is classified as true, false or unverified.

In 2017 we had two variants of the task, a closed and an open one.

- In the open variant, a system could consider the source tweet itself, the discussion as well as additional background information.
- In the closed variant, only the source tweet and the ensuing discussion were used by systems.

Eight teams entered subtask A, achieving accuracies ranging from 0.635 to 0.784. In the open variant of subtask B, only one team participated, gaining an accuracy of 0.393 and demonstrating that the addition of a feature for the presence of the rumour in the supplied additional materials does improve their score. Five teams entered the closed variant of task B, scoring between 0.286 and 0.536. Only one of these made use of the discussion material, specifically the percentage of responses querying, denying and supporting the rumour but scored joint highest on accuracy and achieved the lowest RMSE. A variety of machine learning algorithms were employed. Among traditional approaches, a gradient boosting classifier achieved the second best score in task A, and a support vector machine achieved a fair score in task A and first place in task B. However, deep learning approaches also fared well; an LSTM-based approach took first place in task A and an approach using CNN took second place in task B, though performing less well in task A. Other teams used different kinds of ensembles and cascades of traditional and deep learning supervised approaches.

For 2019 we wanted to encourage participants to be more innovative in the information they make use of, particularly in exploiting the output of task A in their task B approaches.

We extended the challenges through the addition of new data and by including Reddit posts.

In order to encourage more information-rich approaches, we combined variants of subtask B into a single task, allowing participants to use additional material. This was selected to provide a range of options whilst being temporally appropriate to the rumours in order to mimic the conditions of a real world rumour checking scenario.

1.4 Subtask A - SDQC support classification

Related to the objective of predicting a rumour's veracity, and as a first step in a rumour verification pipeline, Subtask A deals with the complementary objective of tracking how other sources orient to the accuracy of the rumourous story. A key step in the analysis of the surrounding discourse is to determine how other users in social media regard the rumour (Procter et al., 2013). Given a source post containing a rumourous claim and a conversation thread discussing the rumour as input, the objective is to label each of the posts in the conversation thread with respect to their stance towards the rumour.

Success on this task supports success on task B by providing additional context and information; for example, where the discussion ends in a number of agreements, it could be inferred that human respondents have verified the rumour. In this way, task A provides an intermediate challenge in which a larger number of data points can be provided. See Table 2 for an example conversation thread and refer to Derczynski et al. (2017) for more details about the task definition.

1.5 Subtask B - Veracity prediction

As in RumourEval 2017 (Derczynski et al., 2017), the goal of subtask B is to predict the veracity of a given rumour, where the latter is presented in the form of a post reporting an update associated with a newsworthy event. Given such a claim as input,

SDQC support classification. Example 1:

u1: We understand that there are two gunmen and up to a dozen hostages inside the cafe under siege at Sydney.. ISIS flags remain on display #7News **[support]**

u2: @u1 not ISIS flags **[deny]**

u3: @u1 sorry - how do you know its an ISIS flag? Can you actually confirm that? **[query]**

u4: @u3 no she cant cos its actually not **[deny]**

u5: @u1 More on situation at Martin Place in Sydney, AU LINK **[comment]**

u6: @u1 Have you actually confirmed its an ISIS flag or are you talking shit **[query]**

SDQC support classification. Example 2:

u1: These are not timid colours; soldiers back guarding Tomb of Unknown Soldier after today's shooting #StandforCanada PICTURE **[support]**

u2: @u1 Apparently a hoax. Best to take Tweet down. [deny]

u3: @u1 This photo was taken this morning, before the shooting. [deny]

u4: @u1 I dont believe there are soldiers guarding this area right now. [deny]

u5: @u4 wondered as well. Ive reached out to someone who would know just to confirm that. Hopefully get response soon. [comment]

u4: @u5 ok, thanks. [comment]

Table 2: Examples of tree-structured threads discussing the veracity of a rumour, where the label associated with each tweet is the target of the SDQC support classification task.

plus additional data such as stance data classified in task A and any other information teams chose to use from the selection provided, systems return a label describing the anticipated veracity of the rumour. Examples are given in Table 1. In addition to returning a classification of true, or false, a confidence score was also required, allowing for a finer grained evaluation. A confidence score of 0 should be returned if the rumour is unverified.

2 Data & Resources- RumourEval 2019

The data are structured as follows. Source posts introduce a rumour, and may be true, false or unverified. These are accompanied by an ensuing discussion (tree-shaped) in which users support, deny, comment or query (SDCQ) the rumour in the source text. This is illustrated in figure 1 with an example rumour about Putin. Note that source posts also need to be annotated for stance, as the way a post presents a rumour usually gives stance information also. For example, when introducing a rumour, an implicit "support" stance may be present, in that the rumour is assumed to convey valid information. In the Reddit data, rumours were often introduced with an implicit "query", as they were presented for discussion/debunking.

The RumourEval 2017 corpus contains 297 source tweets grouped into eight breaking news events, and a total of 7100 discussion tweets. This became training data in 2019, and was augmented with new Twitter test data and new Reddit material. The Reddit material was split into training and test sets. Each are discussed in turn below.

In RumourEval 2017 along with the tweet threads, we also provided additional context that participants could make use of (Derczynski et al., 2017). However, only one system had made use of this additional context. Due to lack of time such context data was not provided in RumourEval 2019 but we would look into re-introducing this in future editions of the task.

2.1 English Twitter data about natural disasters

The additional English Twitter testing data is about natural disasters. In such events, where chaos dominates the situation, rumours are spread on various issues and false rumours have the potential to increase the chaos. Detecting such false rumours are important to plan actions that will

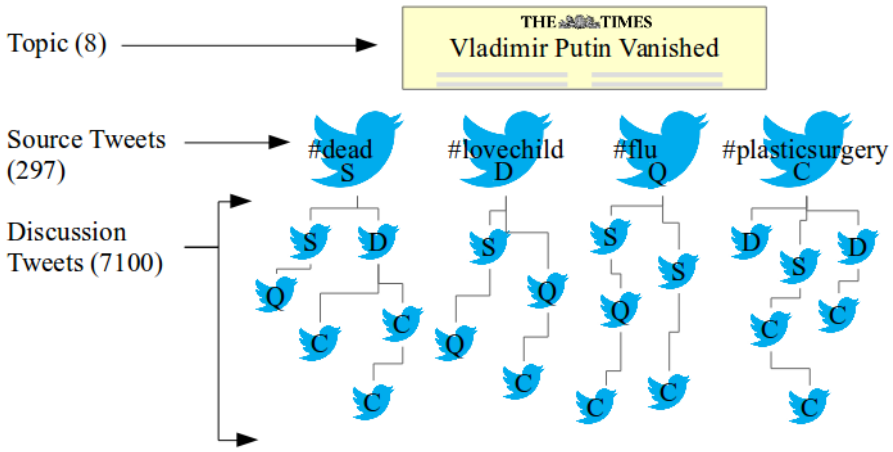

Figure 1: Structure of the first rumours corpus

eliminate the additional negative impact on the already existing chaotic situation. Therefore, for this year we decided to introduce such a dataset as test data. To collect this dataset rumours about natural disasters were chosen manually through Snopes.com and Politifact.com: we searched manually for rumours about known natural disasters such as hurricanes, floods, etc. If the search returned some results, we quickly scanned this result list for social media posts (specifically tweets) that people had created about the disaster and which had been verified by the debunking web-site.

Once we collected the rumour introducing tweets (the source tweets) we aimed to collect also the cascades, i.e. the reactions/replies to the source tweet. The replies encode the reactions (stance information) of other users to the rumour and can be of importance when verifying the rumour. To collect the replies we used an existing scraper (Zubiaga et al., 2016). The number of source tweets of different veracities and replies of different stances are given in Tables 3 and 4.[8]

2.1.1 Annotation of new English Twitter data

As noted above a rumour consists of a source tweet and a thread of tweets that respond to the source one, where the source tweet contains the rumour. The veracity of each source tweet is already known a priori. However, the dataset is missing stance labels for the replies. To get also the stance labels we performed annotation through crowd sourcing. Zubiaga et al. (2016) distinguish between the following stance labels for each replying tweet: supporting, denying, questioning and commenting.

Following the same strategies and design reported by Zubiaga et al. (2016) we posted our datasets for stance annotation to FigureEight (F8)[9]. We applied a restriction so that annotation could be performed only by people from the USA and UK. We also made sure that each annotation was performed maximum by 10 annotators and that an annotator agreement of min. 70% was met. Note if the agreement of 70% was met with fewer annotators then the system would not force an annotation to be done by 10 annotators but would finish earlier. The system requires 10 annotators if the minimum agreement requirement is not met. Each annotator saw five source tweets on a page. The source tweets were accompanied by replying tweets followed by the stance labels to choose from. Each page showed also instructions and definitions about the stance labels. We paid for each tweet annotation 3 US Dollar Cents.

The agreement among the annotators is directly taken from F8s aggregated scores and is computed based on percentage agreement. On the entire dataset we have 76.2% agreement.

We also computed the distribution of stances provided for the replying tweets (see Tables 3 and 4). As we see from the tables, overall the distribution of stances is skewed towards the comment category. This is also the case with the PHEME dataset reported by Zubiaga et al. (2016).

[8]The labels were taken from the debunking web-sites. As in the RumourEval2017 test data the false rumours dominate. However, unlike the previous dataset the number of unverified rumours is proportionally smaller compared to the other two classes. In the 2017 dataset the test data included 12 false rumours, 8 true and 8 unverified ones.

[9]www.figure-eight.com

2.2 Reddit Data

Rumours were identified on Reddit by manually searching debunking forums and current affairs forums to identify suitable threads. Reddit discussions are deeper than Twitter discussions, with often a complex conversational structure exploring the topic. They are usually introduced by a post implicitly querying the rumour, unlike Twitter rumours which are more often presented as valid information and therefore the source tweets usually support the rumour. The Reddit material is less time-sensitive than the Twitter material, and may discuss long-standing conspiracy theories, for example. Threads were downloaded using a bespoke script.

2.2.1 Annotation of Reddit discussions

Since the Reddit discussions are complex, there is more of a danger that careless annotators won't distinguish between posts that disagree with the immediately preceding comment and posts that disagree with the rumour. A response such as "absolutely!" might therefore get a high agreement from annotators who all made the mistake of annotating it as "support", even if it was in response to a preceding comment which denied the rumour. To avoid this, an extensive quiz of 51 test questions was used to ensure that annotators understood the task properly. Reddit threads tend to be longer and more diverse, leading to a more challenging task as discussion may be only loosely related to the main topic, leading to a preponderance of "comments" (88% overall compared with 67% in the Twitter data). Tables 3 and 4 give totals in training and test data for both tasks alongside the figures for Twitter data.

Up to five judgements were collected, or an agreement of 0.7, whichever came first. Since Reddit annotators were highly trained by the time they were accepted on the task, this was found sufficient. Four US dollar cents per post was offered, which is higher than usual for a Figure Eight task, in order to attract annotators to this relatively hard task. The final macro-agreement for the entire Reddit set is 78%, and an average of 3.84 annotations annotated each item. For "support" items, more annotations were required, at 4.22 on average, and a lower macro agreement was achieved of 67%. Similarly for deny items, 4.04 judgements were obtained on average and a macro agreement of 63% was achieved. For query items, 4.36 judgements on average were obtained and a macro agreement of 64% was achieved.

	Supp.	**Deny**	**Query**	**Com.**	**Total**
Twitter Train	1004	415	464	3685	5568
Reddit Train	23	45	51	1015	1134
Total Train	1027	460	515	4700	6702
Twitter Test	141	92	62	771	1066
Reddit Test	16	54	31	705	806
Total Test	157	146	93	1476	1872
Total Task A	1184	606	608	6176	8574

Table 3: Task A corpus

	True	**False**	**Unver.**	**Total**
Twitter Train	145	74	106	325
Reddit Train	9	24	7	40
Total Train	154	98	113	365
Twitter Test	22	30	4	56
Reddit Test	9	10	6	25
Total Test	31	40	10	81
Total Task B	185	138	123	446

Table 4: Task B corpus

For Task B, rumours were annotated for veracity with the aid of Snopes and similar sites. This is a change from RumourEval 2017, where manually-annotated veracity was assigned. Instead, we used community experts working professionally in a range of organisations to construct the Task B veracity judgments. The volume of data was also significantly extended beyond e.g. the 21 stories in the test set of RumourEval 2017 Task B.

3 Evaluation

In task A, stance classification, care must be taken to accommodate the skew towards the "comment" class, which dominates, as well as being the least helpful type of data in establishing rumour veracity. Therefore we used macro-averaged F1 to evaluate performance on task A.

In task B participants supply a true/false classification for each rumour, as well as a confidence score. Macro-averaged F1 was again the score of choice to evaluate the overall classification. For the confidence score, a root mean squared error (RMSE, a popular metric that differs only from the Brier score in being its square root) was calculated relative to a reference confidence of 1. Unverified rumours were considered correctly annotated if they received a confidence score of zero regardless of true/false classification.

The previous RumourEval task used accuracy as the evaluation metric, but that approach allowed higher scores to be obtained through less sensitivity to minority classes. For the stance task, 80% of test items were comments, and this is the least interesting class. For the verification task, class imbalance is not so extreme, with 50%"false" in the dataset and close to 40% "true" (the remainder are "unverified").

Whilst participants weren't evaluated on accuracy for task A, we note that generally speaking, teams that obtained higher macro F1 scores also obtained higher accuracies, and that around 50% of the teams obtained accuracies higher than might be obtained simply by assigning all items to the comment class (majority baseline). However, the correlation between accuracy and macro F1 was only 0.47, and use of macro F1 revealed that three teams surged ahead. For task B, where class imbalance was less pronounced, the relationship between accuracy and macro F1 was much closer, with a correlation of 0.87, though again, F1 was the better differentiator. Interestingly, RMSE showed a stronger relationship with macro F1 than with accuracy (correlations -0.92 vs -0.77).

4 Baselines

We provided participants with our implementation of several baseline systems[10], described below.

4.1 Stance classification baseline

For subtask A we released a Keras (Chollet et al., 2015) implementation of branchLSTM, the winning system of RumourEval 2017 Task A (Kochkina et al., 2017). This system uses the conversation structure by splitting it into linear branches. It is a neural network architecture that uses LSTM layer(s) to process sequences of tweets, outputting a stance label at each time step. Each tweet is represented by the average of its word vectors [11] concatenated with a number of extra features. This baseline was outperformed by 3 submitted systems (BLCU NLP, BUT-FIT, eventAI).

4.2 Veracity classification baselines

For subtask B we provided two baselines.

1. A model which is an extension of branchLSTM (Kochkina et al., 2018) uses the same features as the stance classification system but produces a single output per branch. The veracity prediction for the thread is then decided using majority voting over per-branch outcomes.

2. The NileTMRG baseline (Enayet and El-Beltagy, 2017) is a linear SVM that uses a bag-of-words representation of the source tweet, concatenated features defined by the presence of URL, presence of hashtag and proportion of supporting, denying and querying tweets in the thread. In our implementation of NileTMRG e use the branchLSTM model to obtain stance labels for the tweets in the testing set rather than the model originally used in (Enayet and El-Beltagy, 2017).

Baseline systems in subtask B were outperformed by the winning system eventAI (outperforms both baselines) and a late submission by FINKI NLP (outperforms NileTMRG and reaches similar result to branchLSTM, see Table 5). If participants made their own run of the baseline sys-

User or Team name	Subtask B, MacroF	Subtask B, RMSE	Subtask A, Macro F
eventAI	**0.5765 (1)**	**0.6078 (1)**	0.5776 (3)
WeST (CLEARumor)	0.2856 (2)	0.7642 (2)	0.3740 (11)
GWU NLP LAB	0.2620 (3)	0.8012 (3)	0.4352 (7)
BLCU NLP	0.2525 (4)	0.8179 (5)	**0.6187 (1)**
shaheyu	0.2284 (5)	0.8081 (4)	0.3053 (17)
Columbia	0.2244 (6)	0.8623 (7)	0.3625 (13)
mukundyr	0.2244 (6)	0.8623 (7)	0.3404 (15)
Xinthl	0.2238 (7)	0.8623 (7)	0.2297 (18)
lzr	0.2238 (7)	0.8678 (8)	0.3404 (15)
UPV-28-UNITO	0.1996 (8)	0.8264 (6)	0.4895 (4)
NimbusTwoThousand	0.0950 (9)	0.9148 (9)	0.1272 (19)
nx1 (deanjjones)	-	-	0.3267 (16)
jurebb	-	-	0.3537 (14)
UI-AI	-	-	0.3875 (10)
LECS	-	-	0.4384 (6)
magc	-	-	0.3927 (9)
BUT-FIT	-	-	0.6067 (2)
HLT(HITSZ)	-	-	0.4792 (5)
wshuyi	-	-	0.3699 (12)
SINAI-DL	-	-	0.4298 (8)
FINKI NLP 2018/2019 (late)	0.3326	0.6846	0.2165
IASBS (late)	0.1845	0.7857	0.2530
baseline branchLSTM	0.3364	0.7806	0.4929
baseline NileTMRG	0.3089	0.7698	-
baseline Majority class	0.2241	0.7115	0.2234

Table 5: Results table. Ranking is in brackets.

[10] https://github.com/kochkinaelena/RumourEval2019

[11] We are using word2vec (Mikolov et al., 2013) model pretrained on the GoogleNews dataset (300d)

tems, their outcome might differ from ours due to variation in random seeds, package versions and hardware used.

5 Participant Systems and Results

We have had 22 system submissions at RumourEval 2019 (70% up from RumourEval 2017), confirming the significant increase in interest in this area. All submissions tackled subtask A (Rumour SDQC) and 13 systems attempted both tasks (more than a 100% increase). The participating systems and the results achieved can be found in Table 5. Note that system ranking is presented according to macro-F1 score in subtask B, which is considered the core task and the more challenging of the two. As in RumourEval 2017 subtask A was the more popular task of the two and whilst participation in both tasks has significantly increased, it is still the case that systems seem to focus and do better in one of the two tasks. Specifically, the best performing system in substask B (eventAI) ranked third in subtask A and the best performing system in subtask A (BLCU NLP) ranked fourth in subtask B. Three systems outperformed the branchLSTM subtask A baseline (BLCU NLP, BUT-FIT, eventAI), whereas almost all systems outperformed the majority baseline macro-F1 in this task. In subtask B, over 60% of systems outperformed the majority baseline in macro-F1, two systems outperformed the NILETMRG baseline (eventAI,FINKI-NLP–late) and one system (eventAI) beat both the NILETMRG and branchLSTM baselines.

The trend for neural approaches has demonstrably increased with almost all systems adopting a neural network (NN) architecture for their models, with the exception of the best performing system in subtask B (eventAI), which implemented an ensemble of classifiers (SVM,RF,LR), including a NN with three connected layers, where individual post representations are created using an LSTM with attention. This also considered a range of other features and postprocessing module to find similarities between source tweets. A similar ensemble model also considering sophisticated features and feature selection using RF would have ranked second in this task (FINKI-NLP, submitted late) as it outperformed the NILETMRG baseline. The second best performing system in subtask A (BUT-FIT) uses an ensemble of BERT (Devlin et al., 2018) models, which allows the pre-training of bidirectional representations to provide additional context. They experiment with different parameter settings and if the model increased overall performance it was added to the classifier. Interestingly the best performing system in task A (BLCU-NLP) and the third best (CLEARumor) also use pre-trained contextual embedding representations with BLCU-NLP using OpenAI GPT (Radford et al., 2018) and ClEARumor using ELMo (Peters et al., 2018). While most systems use single tweets or pairs of tweets (source-response) as their underlying structure to operate on, BLCU-NLP employ an inference chain-based system for this paper. Thus they consider the conversation thread starting with a source tweet, followed by replies, in which each one responds to an earlier one in time sequence. They take each conversation thread as an inference chain and concentrate on utilizing it to solve the problem of class imbalance in subtask A and training data scarcity in subtask B. They also have augmented the training data with external public datasets. Other popular neural models among participants include BiLSTM and LSTM. Judging from the approaches of two best performing systems in each of subtask A and B (BLCU-NLP and eventAI respectively) one could infer that: (1) for subtask A considering the sequence of earlier posts is important to identifying correctly the stance of a post towards the rumour (2) for rumour verification it is more important to consider a variety of different features.

6 Conclusion

We evaluated multiple teams in the tasks of rumour stance detection and rumour veracity evaluation. Interest in these tasks continues to increase, driving performance of systems higher and pushing the sophistication of systems, which are now often using state-of-the-art neural network methods and beyond. Further challenges include use of the rich context available, in terms of both time, conversation, and broader discourse during the evolution of rumours. Additionally, we need to work better with other languages. While we tried to make more available in this task, framing the task and annotating the data proved challenging and demanding. On the other hand, leaving stance detection just to English leaves the majority of the world without this important technology.

Acknowledgements

This work is supported by the European Commissions Horizon 2020 research and innovation programme under grant agreements No. 654024 SoBigData and No. 687847 COMRADES. This work was also partially supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. Cloud computing resources were kindly provided through a Microsoft Azure for Research Award. Work by Elena Kochkina was partially supported by the Leverhulme Trust through the Bridges Programme and Warwick CDT for Urban Science & Progress under the EPSRC Grant Number EP/L016400/1.

References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM.

Weiling Chen, Chai Kiat Yeo, Chiew Tong Lau, and Bu Sung Lee. 2016. Behavior deviation: An anomaly detection view of rumour preemption. In *Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2016 IEEE 7th Annual*, pages 1–7. IEEE.

François Chollet et al. 2015. Keras. https://keras.io.

Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Robert Dale. 2017. Nlp in a post-truth world. *Natural Language Engineering*, 23(2):319–324.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76.

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

Omar Enayet and Samhaa R El-Beltagy. 2017. Niletmrg at semeval-2017 task 8: Determining rumour and veracity support for rumours on twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 470–474.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1168.

Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. Turing at semeval-2017 task 8: Sequential approach to rumour stance classification with branch-lstm. *In Proceedings of SemEval.ACL*.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. *arXiv preprint arXiv:1806.03713*.

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2018. Towards automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *arXiv preprint arXiv:1809.08193*.

KP Krishna Kumar and G Geethakumari. 2014. Detecting misinformation in online social networks using cognitive psychology. *Human-centric Computing and Information Sciences*, 4(1):14.

Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. 2017. Rumor detection over varying time windows. *PloS one*, 12(1):e0168344.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Rob Procter, Farida Vis, and Alex Voss. 2013. Reading the riots on twitter: methodological innovation for the analysis of big data. *International journal of social research methodology*, 16(3):197–214.

Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, Technical report, OpenAI.

Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web*, pages 745–750. International World Wide Web Conferences Steering Committee.

Sneha Singhania, Nigel Fernandez, and Shrisha Rao. 2017. 3han: A deep neural network for fake news detection. In *International Conference on Neural Information Processing*, pages 572–581. Springer.

Andreas Vlachos and Sebastian Riedel. 2015. Identification and verification of simple claims about statistical properties. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2596–2601. Association for Computational Linguistics.

Soroush Vosoughi. 2015. *Automatic detection and verification of rumours on Twitter*. Ph.D. thesis, Massachusetts Institute of Technology.

Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*, pages 651–662. IEEE.

Qiao Zhang, Shuiyuan Zhang, Jian Dong, Jinhua Xiong, and Xueqi Cheng. 2015. Automatic detection of rumor on social network. In *Natural Language Processing and Chinese Computing*, pages 113–122. Springer.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):32.

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