Rumor Detection By Exploiting User Credibility Information, Attention and Multi-task Learning

Quanzhi Li, Qiong Zhang, Luo Si

Alibaba Group, US

Bellevue, WA, USA

Abstract

In this study, we propose a new multi-task learning approach for rumor detection and stance classification tasks. This neural network model has a shared layer and two task specific layers. We incorporate the user credibility information into the rumor detection layer, and we also apply attention mechanism in the rumor detection process. The attended information include not only the hidden states in the rumor detection layer, but also the hidden states from the stance detection layer. The experiments on two datasets show that our proposed model outperforms the state-of-the-art rumor detection approaches.

1 Introduction

Social media platforms, such as Twitter, Reddit and Facebook, do not always pose authentic information. Rumors sometimes may spread quickly over these platforms, and they usually spread fear or hate. Therefore, rumor detection and verification has gained great interest recently. Social media platforms and government authorities are also taking great efforts to defeat the negative impacts of rumors.

Rumor Detection: Rumor definition varies over different publications. The lack of consistency makes it difficult to do a head-to-head comparison between existing methods. In this paper, a rumor is defined as a statement whose truth value is *true*, *unverified* or *false* (Qazvinian et al., 2011). When a rumor's veracity value is *false*, some studies call it "*false rumor*" or *"fake news"*. However, many previous studies give "*fake news*" a stricter definition: fake news is a news article published by a news outlet that is intentionally and verifiably false (Shu et al., 2017; Zubiaga et al., 2018). The focus of this study is rumor on social media, not fake news. There are also different definitions for *rumor detection*. In some studies, rumor detection is defined as determining if a story or online post is a rumor or non-rumor (i.e. a real story, a news article), and the task of determining the veracity of a rumor (*true*, *false* or *unverified*) is defined as rumor verification (Zubiaga et al., 2016; Kochkina et al., 2018). But in this paper, as well as in many previous studies (Ma et al., 2016; Shu et al, 2017), *rumor detection* is defined as determining the veracity value of a rumor. This means it is the same as *rumor verification* defined in some other studies. *Rumor detection* and *rumor verification* will be used interchangeably in this paper.

Zubiaga et al. (2018a) consider the rumor resolution process as a pipeline involving four sub-tasks: (1) rumor identification, determining whether a claim is worth verifying rather than the expression of an opinion, i.e. checking a claim is rumor or non-rumor; (2) rumor tracking, collecting opinions on a rumor as it unfolds; (3) stance classification, determining the attitude of users towards the truthfulness of the rumor, and (4) rumor verification, the ultimate step where the veracity value of the rumor is predicted. This study involves the last two tasks: stance classification (detection) and rumor verification (i.e. rumor detection). And this paper mainly focuses on the final step, rumor detection.

Problem Statement: Now we formally define the rumor detection problem: A story *x* is defined as a set of *n* pieces of related messages $M = \{m_1, m_{2, \ldots,} m_n\}$. m_1 is the source message (post) that initiated the message chain, which could be a tree-structure having multiple branches. For each message m_i, it has attributes representing its content, such as text and image. Each message is also associated with a user who posted it. The user also has a set of attributes, including name, description, avatar image, past posts, etc. The rumor detection task is then defined as follow: Given a story *x* with its message set *M* and user set *U*, the rumor

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1173–1179
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

detection task aims to determine whether this story is *true, false* or *unverified* (or just *true* or *false* for datasets having just two labels). This definition formulates the rumor detection task as a veracity classification task. The definition is the same as the definition used in many previous studies (Shu et al, 2017; Ma et al., 2016).

There are four stance categories: *supporting*(S), *denying*(D), *querying*(Q) and *commenting*(C), i.e. SDQC. The veracity of a rumor has three values: *true, false,* or *unverified*. For both stance detection and rumor detection, traditional approaches used supervised learning algorithms incorporating a variety of features generated from post content, user profiles, and diffusion patterns (Castillo et al., 2011; Kwon et al., 2013; Liu et al., 2015; Ma et al., 2015; Zhao et al., 2015). Recent studies have shown that the sequential time-sensitive approach has benefited both rumor detection and stance detection tasks (Ma et al., 2016; Kwon et al., 2017; Ma et al., 2017; Ma et al., 2018a; Kochkina et al., 2018). In this study, we also use the sequential classification approach on these two tasks. A rumor consists of a source post that makes a claim, and a set of replies, directly or indirectly towards the source post. This set of posts may have multiple conversation branches. Our model exploits the structural information of these conversations.

Multi-task learning (Caruana, 1998; Liu et al., 2016) has been applied in many NLP tasks. In this study, we use a shared Long-Short Term Memory (LSTM) layer to learn a set of common features relevant to both tasks, while each task can also learn their task-specific features via their specific layer. Compared to previous studies (Ma et al., 2018; Kochkina et al., 2018) that also use multi-task learning for stance detection and rumor verification, the main differences between ours and them are: 1. We incorporate features that describe user credibility information into the rumor detection layer. User credibility information, which is derived from user profile in this study, is critical in rumor detection task, as already proven in Liu et al. (2015) and Castillo et al. (2011). But recent studies using sequential classification have not made use of it. To our knowledge, this is the first study that incorporates user credibility/profile information in neural network for sequential classification. 2. We apply attention mechanism in the rumor detection process. And the attention includes not only the hidden states in the rumor detection layer, but also the hidden states of the stance detection layer. In a conversation branch, some posts, especially the ones with strong stance, will be more important than others in determining the rumor veracity. No previous study has exploited this on rumor detection.

Although stance detection is included in the multi-task learning network, in this study, we focus on the main task, rumor detection, so the experiments are conducted for evaluating the performance of rumor detection. Our experiments show that our approach outperforms the state-of-the-art methods.

2 Related Studies

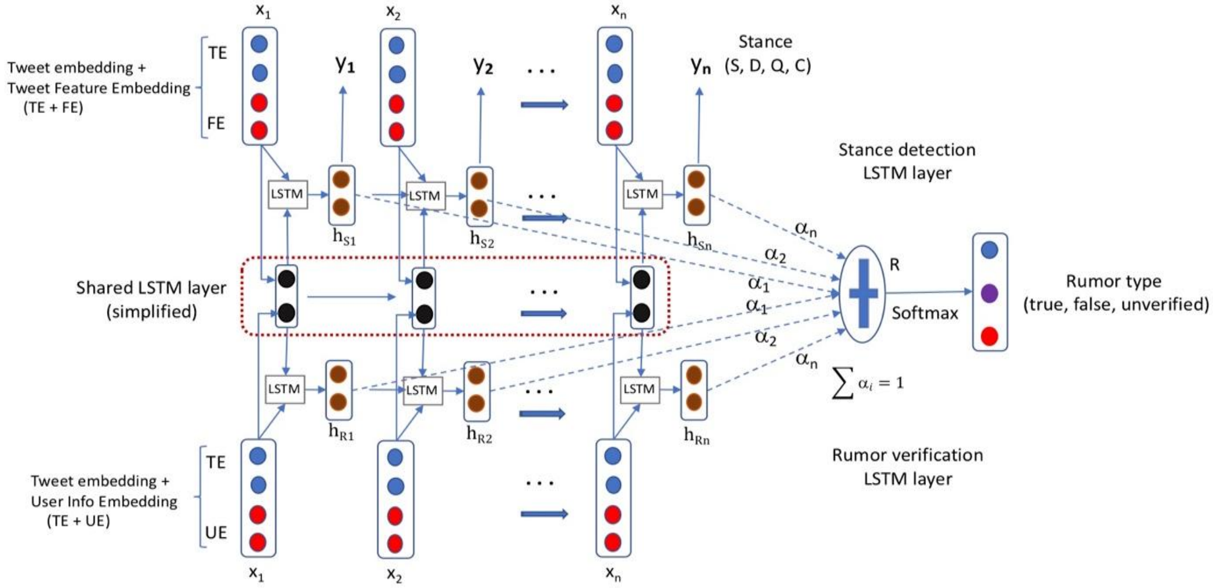
Many existing algorithms (Liu et al., 2015; Wu et al., 2015; Yang et al., 2012) for debunking rumors followed the work of Castillo et al. (2011). They studied information credibility and various features. Stance classification is also an active research area that has been studied in previous work (Ranade et al., 2013; Chuang and Hsieh, 2015; Lukasik et al., 2016; Zubiaga et al., 2016; Kochkina et al., 2017).

Several studies have employed neural networks on rumor verification (Ma et al., 2016; Kochkina et al., 2017; Ma et al., 2017), and they mainly focus on analyzing the information propagation structure. Multi-task learning has been used in various NLP tasks, including rumor verification (Collobert et al., 2011; Aguilar et al., 2017; Lan et al., 2017; Ma et al., 2018a; Kochkina et al., 2018). Kochkina et al. (2018) proposed a multi-task method without task specific layer for rumor verification. MT-ES is a multi-task approach using Gated Recurrent Unit (GRU) (Cho et al., 2014) with a task specific layer for each task (Ma et al., 2018a). MT-ES has no attention mechanism, and it does not use user information. Ma et al. (2018b) proposed a model based on tree-structured recursive neural networks.

3 The Proposed Model

3.1 The Multi-task Network Structure

Figure 1 presents the high-level structure of our proposed multi-task learning approach. The middle layer is a shared layer, shared by the two tasks. This layer is to extract the common patterns between these two tasks, via the shared parameters. The upper layer is for stance detection, and the lower layer is for rumor detection. These two layers will capture task specific features. In this figure, we assume the posts are tweets, and will use tweets as examples in the following sections. The input to the two task specific layers is a claim (rumor, thread) branch. Take the rumor propagation path in Figure 2 as an example, this rumor has four branches, and each branch has an input sequence [x_1, x_2, …, x_n], fed into the two task specific layers. x_1 is the source tweet (post), and x_n is the last tweet in a branch.



x1
x2
xn
Tweet embedding + Tweet Feature Embedding (TE + FE)
TE
FE
y1
y2
yn
Stance (S, D, Q, C)
LSTM
hS1
hS2
hSn
Stance detection LSTM layer
Shared LSTM layer (simplified)
α1
α2
αn
R
Softmax
Rumor type (true, false, unverified)
∑αi = 1
hR1
hR2
hRn
Rumor verification LSTM layer
Tweet embedding + User Info Embedding (TE + UE)
UE

Figure 1. The high-level structure of our proposed approach. The shared LSTM layer is in the middle (in the red dot-line rectangle). The upper layer is the stance detection specific layer, and the lower layer is for rumor verification task.

Tweet Embedding (TE): We generate the tweet embedding through an attention-based LSTM network. The word embeddings were built from 200 million tweets using the word2vec model (Mikolov et al., 2013; Li et al., 2017).

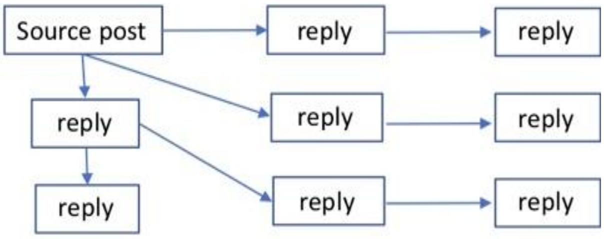


Figure 2: A rumor propagation example. There are four branches in this rumor.

3.2 The Stance Detection Layer

As shown in Figure 1, the stance detection layer uses a standard LSTM model. The input x_i is a concatenation of two types of features: the tweet embedding (TE) and a tweet feature embedding (FE). FE is generated using the same list of features described in (Kochkina et al., 2017). Some FE feature examples are content length, presence of a URL, and if it is a source tweet or not.

At each time step i, the hidden state h_{si} is fed to a fully connected hidden layer, and a softmax layer is used to predict the stance type (e.g. S, D, Q, C). These hidden states are also used in the attention step of the rumor verification task.

3.3 The Rumor Verification Layer

The lower layer of Figure 1 shows the structure of the rumor verification process. At each step, the input x_i is represented by two vectors, tweet embedding (TE) and user information embedding (UE). UE is to represent user credibility information.

User Credibility Information: Many previous studies have shown that user credibility information is very important in rumor verification (Li et al., 2016; Liu et al., 2015). This is especially true when a rumor is debunked or supported by a credible user, such as a verified user, news agent, government agent, or a professional in the area of the rumor topic. But recent studies using sequential classification and neural network have not made use of this information. We hypothesize that this information will improve rumor verification performance. In this study, we derive the credibility information from user profile. We use the features described in (Liu et al., 2015) to derive this information. Some feature examples are: is verified account, if profile includes location, if profile has description, etc. These information are processed and concatenated together as the UE embedding, and then UE is concatenated with TE as input.

Attention-based LSTM: In a conversation branch, different posts will have different impacts on the rumor veracity. For example, the tweets with strong *support* or *deny* stance should have more impact for predicting rumor veracity. In order to better exploit the stance information, we explicitly include the hidden states from the stance layer in the attention calculation. Besides the tweets with strong stance, we should also pay more attention to the credible users. This can be done through attention in the rumor-specific layer, since it has already encoded the user credibility information through UE embedding. Therefore, we use an attention-based LSTM to give more attention to the important tweets. At each step *i*, the hidden state from the upper layer and the state from the lower layer are actually concatenated and attended together. In other words, they use the same attention weight, α_i. Vectors in sequence h_{Ri} and h_{Si} are fed into a learnable function $a(h_{Ri}, h_{Si})$ to generate a probability vector a_i . The vector R is then computed as a weighted average of (h_{Ri}, h_{Si}), with weighting given by a_i:

$$R = \sum_{i=1}^{n} \alpha_{\boldsymbol{i}} \, (h_{\boldsymbol{Ri}}, h_{\boldsymbol{Si}}) \qquad (1)$$

The hidden state R is fed into a fully connected layer, and softmax is used for veracity prediction.

4 Experiments and Results

Datasets: Two publicly available rumor datasets are used: RumorEval (Derczynski et al., 2017) and PHEME (Zubiaga et al., 2016; Zubiaga et al., 2017). RumorEval was released as part of the SemEval-2017 Task 8 competition (Derczynski et al., 2017). It contains 325 rumors (4017 branches) from Twitter. Each tweet is also labeled with a stance. The PHEME dataset has 1,972 rumors. But its tweets have no stance label. To get their stance labels for the multi-task learning, following (Kochkina et al., 2018), we also used the stance detection algorithm described in (Kochkina et al., 2017) to automatically annotate these tweets. The RumorEval dataset was provided with a training/development/testing split. For PHEME dataset, we use cross validation, same as (Kochkina et al., 2018). Accuracy and Macro F1 are used as the evaluation metrics.

Regarding the stance annotation of the RumorEval data set (Derczynski et al., 2017), as the task description paper already pointed out: the overall inter-annotator agreement rate of 63.7% showed the task to be challenging, and easier for source tweets (81.1%) than for replying tweets (62.2%). This means that there are many conflicting or inconsistent stance labels. When we analyzed the training data set, we found many such examples. To make the labels more consistent, we run an analysis to find the posts that are basically the same or highly similar, but their labels are different. We then mark these posts, and use the same label, the one labeled on the majority of these posts, on them during training. The similarity between two posts is calculated by cosine similarity measure. The similarity threshold for being considered as similar posts is empirically set as 0.75.

Compared Methods: We compare our proposed model with the following approaches, including the state-of-the-art algorithms:

Majority vote: this is a strong baseline which results in high accuracy due to the class imbalance in the veracity classification task.

NileTMRG: this is the best veracity prediction system from SemEval-2017 Task 8 (Enayet and El-Beltagy, 2017). It is based on a linear SVM using a bag-of-words representation of the tweet concatenated with selected features.

BranchLSTM: a method based on an LSTM layer followed by several dense ReLU layers and a softmax layer (Zubiaga et al., 2018b).

MTL2: a multi-task method without task specific layers (Kochkina et al., 2018).

Method	**Accuracy**	**Macro F1**
Majority(False)	0.438	0.304
NileTMRG	0.57	0.539
BranchLSTM	0.5	0.491
MTL2	0.571	0.558
Proposed model	0.638	0.606

Table 1: Rumor verification result on RumorEval

Ma et al. (2018a) proposed a multi-task approach using GRU, with a task specific layer for each task. It has no attention mechanism, and does not use user information. Our implementation of their approach did not achieve the performance reported in their paper using the data sets they used, so we do not compare our method to theirs here. Ma et al. (2018b) proposed a model based on tree-structured recursive neural networks . We did not include this model in our experiments, because it uses recursive network and it performs not well on datasets without long propagation path, which is the case for our datasets.

Experimental Settings: Our model is trained to minimize the squared error between the probability distributions of the predictions and the ground truth, same as (Ma et al., 2018a). Stochastic gradient descent, shuffled mini-batch, AdaDelta update, back-propagation and dropout are used in the training process. The TE size is 300. During training, for each branch, the stance task is first executed, followed by the rumor verification task, in order for the verification task to utilize the hidden states of the stance detection layer in its attention step. Zero-padding and masks are used for handling the varying lengths of the input branches; they are also used in (Kochkina et al., 2017; Ma et al., 2018a). A rumor's final veracity is based on the voting result of all its branches.

Method	**Accuracy**	**Macro F1**
Majority (True)	0.511	0.226
NileTMRG	0.438	0.339
BranchLSTM	0.454	0.336
MTL2	0.441	0.376
Proposed model	0.483	0.418

Table 2: Rumor verification result on PHEME dataset

Results: Table 1 shows the result on RumorEval dataset, and Table 2 is for the PHEME dataset. We can see that our proposed method outperforms other approaches on both datasets. In both cases, the performance improvement is statistically significant at the level of *p=0.01* for both accuracy and F1, using *t-test* (Rice, 2006).

Compared to other multi-task models, our model has three main features: 1. it incorporates user credibility information in the rumor verification task, 2. it uses attention mechanism to pay more attention to the important tweets, and 3. it integrates the stance information into the attention computation.

5 Conclusion

We proposed a multi-task learning approach for rumor detection and stance classification tasks. This model incorporates the user credibility information into the rumor detection layer, and uses attention mechanism in the rumor detection process. The experiments on two datasets show that our proposed model outperforms the state-of-the-art rumor detection approaches.

References

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor L´opez Monroy, and Thamar Solorio. 2017. A multi-task approach for named entity recognition in social media data. In Proceedings of the 3rd Workshop on Noisy User-generated Text, pages 148–153.

Carlos Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. WWW 2011.

Rich Caruana. 1998. Multitask learning. In Learning to learn. Springer, 95–133.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014).

Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. J. Mach. Learn. Res., 12:2493–2537, 2011.

Ju-han Chuang and Shukai Hsieh. Stance classification on post comments. In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation. 2015

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.

Omar Enayet and Samhaa R El-Beltagy. 2017. Niletmrg at semeval-2017 task 8: Determining rumour and veracity support for rumours on twitter. SemEval-2017.

Genevieve Gorrell, Kalina Bontcheva, Leon Derczynski, Elena Kochkina, Maria Liakata, and

Arkaitz Zubiaga, RumourEval 2019: Determining Rumour Veracity and Support for Rumours. SemEval 2019

Gupta, H. Lamba, P. Kumaraguru, and A. Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. WWW 2013

Elena Kochkina, Maria Liakata, Isabelle Augenstein, 2017, Turing at SemEval-2017 Task 8: Sequential Approach to Rumour Stance Classification with Branch-LSTM, SemEval 2017

Elena Kochkina, Maria Liakata, Arkaitz Zubiaga, All-in-one: Multi-task Learning for Rumour Verification, COLING 2018

Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. ICDM.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. 2017. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In Proceedings of the 2017

Conference on Empirical Methods in Natural Language Processing, pages 1299–1308.

Quanzhi Li, Xiaomo Liu, Rui Fang, Armineh Nourbakhsh, Sameena Shah, 2016, User Behaviors in Newsworthy Rumors: A Case Study of Twitter. The 10th International AAAI Conference on Web and Social Media (ICWSM 2016)

Quanzhi Li, Sameena Shah, Xiaomo Liu, Armineh Nourbakhsh, 2017, Data Set: Word Embeddings Learned from Tweets and General Data, The 11th International AAAI Conference on Web and Social Media (ICWSM 2017).

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. IJCAI 2016

Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, Sameena Shah, 2015, Real-time Rumor Debunking on Twitter, CIKM 2015.

Michal Lukasik, P. K. Srijith, Duy Vu, Kalina Bontcheva, Arkaitz Zubiaga, and Trevor Cohn. 2016. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. ACL 2016

Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect Rumors Using Time Series of Social Context Information on Microblogging Websites. In Proceedings of CIKM.

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

Jing Ma, Wei Gao, Kam-Fai Wong, 2017, Detect rumors in microblog posts using propagation structure via kernel learning, ACL 2017

Jing Ma, Wei Gao, Kam-Fai Wong, Detect Rumor and Stance Jointly by Neural Multi-task Learning, WWW 2018

Jing Ma, Wei Gao, Kam-Fai Wong, Rumor Detection on Twitter with Tree-structured Recursive Neural Networks, ACL 2018

M. Mendoza, B. Poblete, and C. Castillo. Twitter under crisis: Can we trust what we rt? In Proc. First Workshop on Social Media Analytics, 2010.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.

Saif M Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. SemEval 2016.

K. Popat, S. Mukherjee, A. Yates, G. Weikum: DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning, in Proceedings of EMNLP, 2018.

V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. EMNLP 2011.

Sarvesh Ranade, Rajeev Sangal, and Radhika Mamidi. 2013. Stance classification in online debates by recognizing users' intentions. SIGDIAL 2013.

John A. Rice. 2006. *Mathematical Statistics and Data Analysis*, Third Edition, Duxbury Advanced

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. SIGKDD Explorations Newsletter

K. Wu, S. Yang, and K. Q. Zhu. False rumors detection on sina weibo by propagation structures. IEEE ICDE 2015.

Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. 2012. Automatic detection of rumor on sina weibo. ACM SIGKDD Workshop on Mining Data Semantics.

Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts. WWW 2015

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.

Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2017. Exploiting context for rumour detection in social media. In International Conference on Social Informatics, pages 109–123. Springer.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018a. Detection and resolution of rumours in social media: A survey. ACM Comput. Survey.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018b. Discourse-aware rumour stance classification in social media using sequential classifiers. IPM