Early Detection of Rumours on Twitter via Stance Transfer Learning

Lin Tian[1], Xiuzhen Zhang[1(✉)], Yan Wang[2], and Huan Liu[3]

[1] RMIT University, Melbourne, Australia
{lin.tian,xiuzhen.zhang}@rmit.edu.au
[2] Macquarie University, Sydney, Australia
yan.wang@mq.edu.au
[3] Arizona State University, Tempe, USA
huan.liu@asu.edu

Abstract. Rumour detection on Twitter is an important problem. Existing studies mainly focus on high detection accuracy, which often requires large volumes of data on contents, source credibility or propagation. In this paper we focus on early detection of rumours when data for information sources or propagation is scarce. We observe that tweets attract immediate comments from the public who often express uncertain and questioning attitudes towards rumour tweets. We therefore propose to learn user attitude distribution for Twitter posts from their comments, and then combine it with content analysis for early detection of rumours. Specifically we propose convolutional neural network (CNN) CNN and BERT neural network language models to learn attitude representation for user comments without human annotation via transfer learning based on external data sources for stance classification. We further propose CNN-BiLSTM- and BERT-based deep neural models to combine attitude representation and content representation for early rumour detection. Experiments on real-world rumour datasets show that our BERT-based model can achieve effective early rumour detection and significantly outperform start-of-the-art rumour detection models.

Keywords: Twitter · Rumour detection · Stance detection · Transfer learning · CNN · BERT

1 Introduction

Nowadays, people tend to acquire more information from online social media platforms than traditional media channels. Especially Twitter allows users to freely publish short messages called "tweets" and has become a popular platform for spreading information. On the other hand, Twitter has also become an ideal place for rumor and misinformation propagation [25]. In 2013, the Associated Press (AP) Twitter account was hacked and published a tweet that two explosions rocked the White House and President was injured. The tweet led Dow Jones Industrial Average dropped 143.5 points and Standard & Poor's 500

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12035, pp. 575–588, 2020.
https://doi.org/10.1007/978-3-030-45439-5_38

Index lost more than $136 billion in a short time period after the event [6]. In this paper, rumours refer to any unconfirmed information, including misinformation, regardless of the intention of the information source.

To assess the truthfulness of rumours and combat misinformation, manual fact checking websites such as snopes.com and emergent.info heavily rely on human observers to report potential rumors and employ professional journalists to fact-check their truthfulness, which is costly and time consuming. Automatic rumour detection is thus desirable to reduce the time and human cost [11,28].

Automatic rumour detection has attracted significant research [28]. There are mainly three types of rumour detection approaches based on the type of data used. Content-based methods focus on rumour detection using the textual contents of tweets and their user comments [12,25,30]. Generally tweet contents have direct signals for misinformation and content analysis for rumour detection is desirable. Feature-based models exploit features other than tweet contents such as author profile information for rumour detection [3,9,10,13,23]. Propagation-based methods exploit patterns in tweet propagation for rumour detection [14, 16,18,27]. Most existing approaches rely on large volumes of training data that are only possible when users have shown sufficient usage or tweets have been propagated for a while, and therefore are not designed for early detection.

Early detection of rumours is most desirable, as it can trigger efforts for effective mitigation of rumours and misinformation at an early stage. But early rumour detection is a challenging task due to the lack of prominent signals in propagation and user metadata within the short period after tweet publication. It is shown by previous research [30] that users post comments to tweets early and they contain questioning or enquiring phrases (e.g. “Is this true?” or “Really?”) that can be exploited for early detection of rumours. But the reliance on fixed expressions implies low recall for the approach.

In this paper, we propose early rumour detection based on only tweet contents and their immediate user comments that are readily available at the early stage. Our main idea is to exploit the wisdom of the public crowd. As shown in previous studies [11,30], the crowd shows attitudes such as disagreeing and questioning toward rumours. We therefore hypothesize that attitudes of the crowd to a tweet contains signals for identifying rumour tweets. We propose to mine the user comments to predict crowd attitudes and detect rumours. But we face the challenge that there do not exist annotations of attitudes for tweet comments. We specifically address the following research questions:

- Can crowd attitudes be exploited for effective early rumour detection?
- How to learn attitude representation from tweet comments without costly human annotation?

Towards answering these research questions, we made several contributions. To address the issue of lack of attitude annotations for user comments, we propose CNN- and BERT-based deep neural models to learn attitude representation from user comments via transfer learning from resources for stance prediction [1,5,20,24,29]. We further propose CNN-BiLSTM and BERT neural models to integrate attitude representation and content representation for tweets

and their comments for rumour detection. Experiments on real-world Twitter rumour datasets show that our proposed models, especially the BERT-based model, outperform state-of-the-art rumour detection models.

2 Related Work

Rumour classification and rumour verification attract significant attention from the research community in shared tasks like RumourEval [8]. According to the type of data used, rumour detection approaches can be divided into three major categories, content-based, feature-based and propagation-based.

Content-based methods focus on rumour detection based on the textual contents of posts, including the original tweets, user comments and retweets. Generally textual contents have direct signals for misinformation and deep analysis of the Twitter messages is desirable for rumour detection. Zhao et al. [30] used a set of expressions (such as "is this true?", "what?") from user comments that express questioning and enquiring as signals for rumours. Limitations from the signal expressions lead to low recall for rumour detection. In [12] a RNN model is trained to automatically learn representations from tweets for rumour detection. In [25], linguistic features of different writing styles and sensational headlines from tweets are exploited to detect misinformation.

Feature-based methods use non-textual features such as user profile data for rumour and misinformation detection [3,9,10,13,23]. In [3] user registration age and number of followers are used for credibility assessment. In [11], features such as belief identification are used for rumour detection. Other studies [10,13,23] build time series model for information propagation and integrate other social and contextual features to detect rumours. Generally the feature-based approaches can be applied only when the original tweets have attracted significant attention on the social network after some time and therefore are not adequate for early detection of rumours or misinformation.

Propagation-based methods exploit tweet propagation information [18] to build classification models such as kernel-based methods [14,27] for rumour classification. Recently a neural network model [16] is proposed, where an extended tree-structured recursive neural network (RvNN) is constructed to model information propagation. Propagation-based approaches require large amounts of metadata and intensive pre-processing to model the propagation process.

Research shows that the public respond differently to rumours than non-rumours [11,16,18,22]. However most existing research treats rumour detection and stance detection as separate tasks. In one exception [7], crowd stance is examined as a feature to classify true and false rumours. In another exception [15] a multi-task learning problem for rumour and stance detection is formulated. It is found that the proposed multi-task model is inferior to models designed specifically for rumour detection.

Stance detection [1,5,20,24,29] aims to automatically detect user attitudes towards given posts, whether the user is in favour of, against or neutral toward the target post. Some deep neural models are proposed for the task and achieve reasonable performance [1,5,29].

More generally transfer learning is widely applied to NLP tasks. As one transfer learning strategy, feature transfer can utilise the feature representation from the source to target domains in order to reduce the target task error rate. In [26], multiple shared layers are created to capture cross-domain features and domain-specific features. To minimise the feature differences between the source and target domains, Cao et al. [2] fine-tuned a shared embedding layer to automatically transfer features from the source to the target domain.

3 Problem Formulation

The task of rumour detection can be formulated as a supervised classification problem. Consider a set of n source tweets $S = \{s_1, s_2, s_3, ..., s_n\}$. Each source tweet s_i, $s_i \in S$, is associated with a label l indicating its rumour class label and a set of comments $C_i = \{c_{i1}, c_{i2}, c_{i3}, ..., c_{im}\}$. Based on the observation that users respond to rumours and non-rumours differently, comments C_i reflect the attitudes of users towards source tweet s_i; significant variation in user attitudes to s_i indicates the uncertainty from the public towards the truthfulness of s_i. Conversely unamimous attributes towards a source tweet likely indicates that truthfulness of the source tweet is clear. The problem of rumour detection for tweet s_i can thus be decomposed to two sub-problems, stance detection from user comments C_i and rumour detection for tweet s_i.

We propose to formulate the task of rumour detection as a transfer learning problem. To achieve rumour classification for a tweet message s_i where user comments C_i do not have attitude annotation, we propose to learn representation for attitudes for user comments via transfer learning based on the readily available annotated resources for stance prediction in the literature [1,5,20,24,29]. The idea is to pretrain a model on the stance data source to learn stance representation and then transfer and integrate this knowledge to the neural model for tweet and comment contents for rumour detection.

The SemEval [19] dataset with stance annotation is employed in our study but generally other stance resources can also be used. The SemEval dataset is a public Twitter dataset where each tweet is annotated with one of three stance labels "Favor", "Against" and "Neither". The tweets are about six target topics, including "Atheism","Climate Change is a Real Concern","Feminist Movement","Hillary Clinton","Donald Trump" and "Legalization of Abortion".

4 Methodology

We propose two approaches to learn vector representation for different stance classes based on the SemEval dataset and then transfer the knowledge to the model for rumour classification, as detailed next.

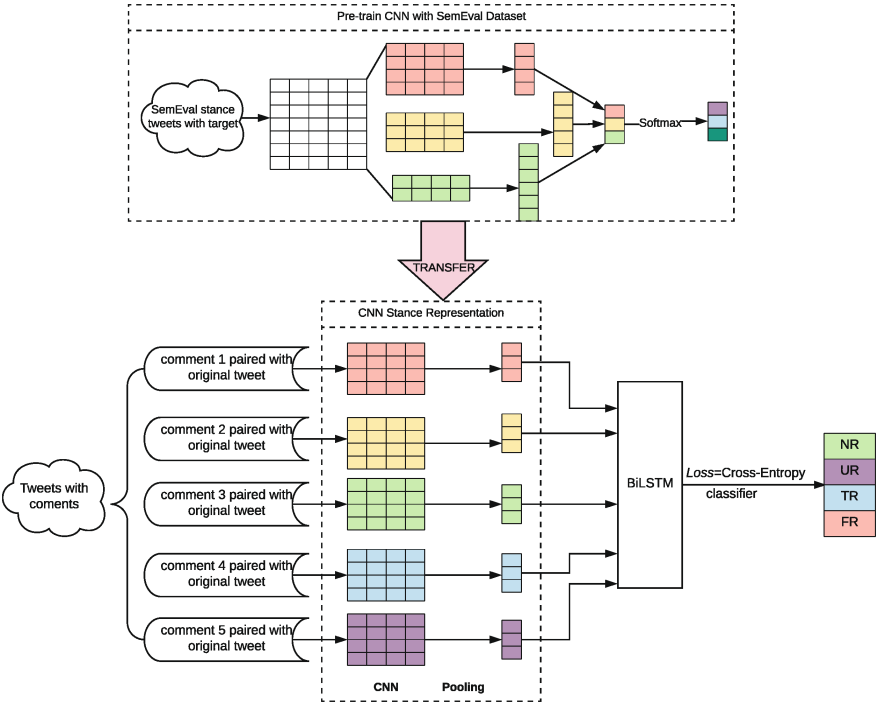


Fig. 1. Our Stance-CNN+BiLSTM model

4.1 Stance-CNN+BiLSTM

Our first model, namely Stance-CNN+BiLSTM, models crowd stances in each comment for a tweet. Specifically we pre-train a CNN model on the SemEval dataset based on the stance labels and then transfer the knowledge to learn attitude representation for each tweet comment. The CNN architecture has the ability to learn high-level feature representation for the interaction between low-level input based on annotated labels. The attitude representation for comments are then integrated into a CNN-biLSTM (bi-directional Long Short Term Memory) model for rumour prediction for tweets with comments.

The model architecture is shown in Fig. 1. The CNN model has convolutional layers and max pooling layers to capture high level features for each comment. Vectors generated from the CNN model become the input for BiLSTM for rumour detection for tweets, where the chronological order of comments and their stance variations from content representations are captured and employed to classify tweets into rumours and non-rumours. In addition, BiLSTM has the ability of ignoring unnecessary features using the delete gate. The entire model was trained to minimise the categorical cross-entropy error: $Loss = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$, where M is number of rumors labels, y is the binary indicator and p stands for the predicted probability.

4.2 Stance-BERT

Our second model, namely Stance-BERT, models the stance distribution for a tweet and its comments via transfer learning from tweet pairs generated from the SemEval dataset. BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained transformer language model to generate deep bidirectional context representations by jointly conditioning on both left and right context in all layers [4]. The main idea of Stance-BERT is to leverage the structure of BERT to capture the complex stance distribution for tweets based on their comments, and to further integrate with a second BERT architecture modelling the language patterns for tweets and comments for rumour classification.

The architecture of our proposed stance-BERT model is shown in Fig. 2. As shown in Fig. 2, the first BERT model is to learn stance distribution for a tweet and its comments. Input are tweet pairs constructed from the SemEval dataset. If tweet A holds the "Favour" stance for topic A, and tweet B also holds the "Favour" stance for topic A, then it can be inferred that tweet A and tweet B has the same Agree stance for topic A; in other words, the new instance, the (tweet A, tweet B) pair, has the label "Favour-Favour". Similarly if tweet C has "Favour" stance for topic A and tweet D has "Against" stance for topic A, then we generate an instance (tweet C, tweet D) with the label "Favour-Against". As there are three stance labels in the original SemEval dataset, there are six combinations for labels, which are "Favour-Favour" (FF), "Against-Against" (AA) and Neither-Neither (NN), "Against-Favour" (AF), "Against-Neither" (AN), and "Favour-Neither" (FN). The six label combinations are used to label tweet pairs. Using the tweet pairs with combined stance labels as input the first BERT model is trained, which is then transferred to learn representation for (tweet, comment) pairs. This formulation of tweet pairs is aimed to capture the different language patterns of (source-tweet, comment) pair for different stance combinations.

To transfer the stance knowledge from the first BERT model for rumour prediction, the stance language patterns flow from the first BERT model to the second BERT model; the feature vector for stance representation is transferred to Twitter comments. Based on the degree of consistency among comments, the second BERT structure is trained for rumour classification.

At the first stage, the uncased BERT base model is fine-tuned with tweet pairs generated based on the SemEval dataset. The generated representation vector for [CLS] are then concatenated and input to the second BERT model to further fine-tune the BERT model for rumour classification based on the original tweet and comments. The second stage of the model has one additional output layer with softmax function for rumour classification, namely $\hat{y} = Softmax(Wh + b)$, where h is the linear vector, W and b are the weights and bias in the output layer.

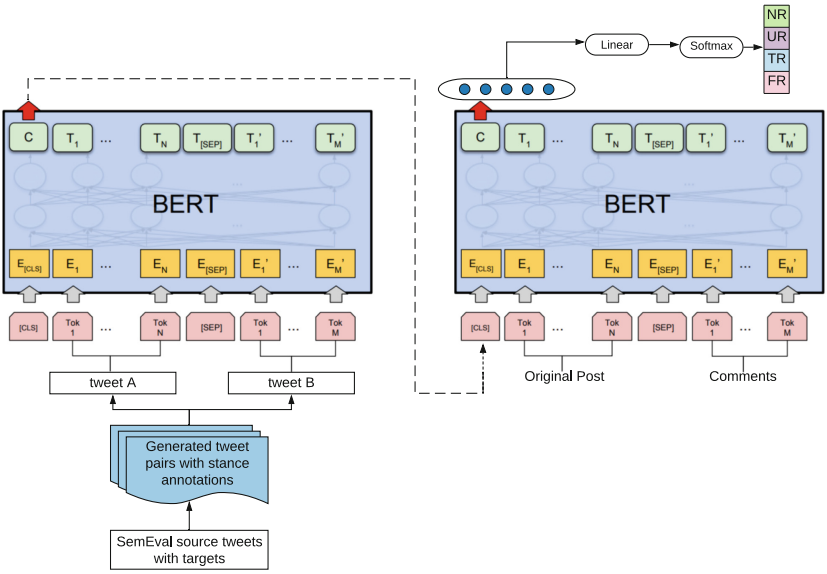


Fig. 2. The architecture of our Stance-BERT model

5 Experiments

We first describe the datasets and then the performance for early rumour detection by our models compared with other baseline models. We further evaluate our stance-transfer models against their counterparts without stance transfer.

Table 1. The Twitter15 and Twitter16 datasets

	Twitter15				Twitter16			
	NR	FR	TR	UR	NR	FR	TR	UR
#tweets	374	370	372	374	205	205	207	201
#comments	25867	21059	14948	15105	17006	7876	5397	9970
Min delay (mins)	1.08	1.50	1.48	1.96	1.02	3.45	1.76	2.25
Max delay (mins)	2714.26	1731.08	1248.85	1161.39	2690.72	2075.73	216.65	1748.02

5.1 Datasets and Experiment Setup

We use two public Twitter datasets [14], namely Twitter15 and Twitter16 (Table 1), for our experiments. In each dataset, tweets and their associated retweets and user response comments are included. Twitter15 and Twitter16 contain 1490 and 818 source tweet posts respectively. Four different rumour labels are applied with these two datasets, including True Rumour (TR), Non-Rumour

(NR), False Rumour (FR) and Unverified Rumour (UR). We removed retweets from the original datasets since retweets are not providing any new information in terms of contents. The comments and retweet contents are not included in the original dataset, only tweet ids are provided. We therefore crawled all the comments through Twitter API according to the tweets ids and user ids.

We compare our models against state-of-the-art rumour detection models:

- Stance-BERT: our BERT-based stance transfer learning models.
- Stance-CNN+LSTM: our CNN+LSTM-based stance transfer learning model.
- SVM [30]: SVM with linguistic features from tweets and comments.
- MT-ES [15]: Multi-task learning model for stance and rumour classification.
- GRU-RNN [12]: RNN model with GRU units for capturing rumour representations with sequential structure of relevant posts.
- TD-RvNN [16]: Propagation tree-based recursive neural network model.

We implemented the SVM model using scikit-learn package in Python and TD-RvNN model with Theano. The SVM model is implemented with radial basis function kernel where $C = 1.0$. All other neutral network models are based on Tensorflow v1.14. We use overall macro F_1 and F_1 scores for each class as model performance evaluation metrics. Five-fold cross-validation experiments are applied for evaluation of models.

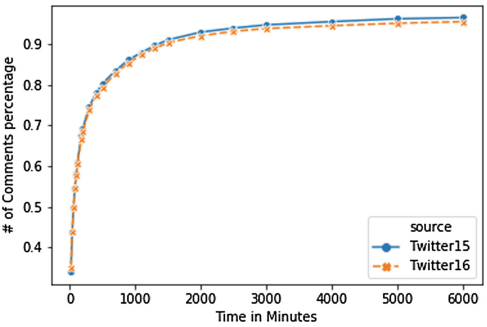
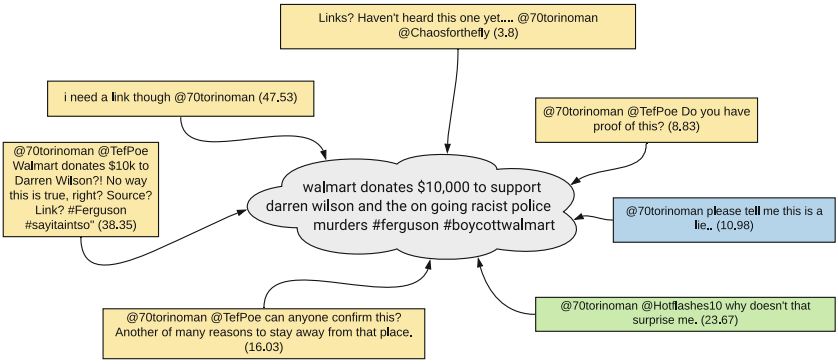


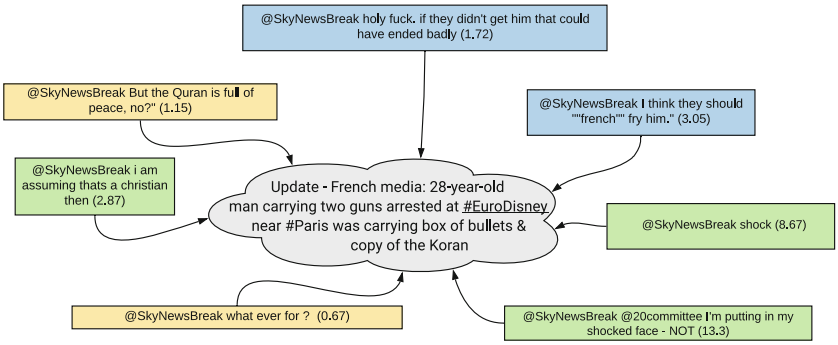
Fig. 3. Number of comments over time delay in minutes

5.2 Analysis of Early Comments for Tweets

We first evaluate the feasibility of using comments for early rumour detection. Figure 3 plots the number of comments with increasing time delay from when the original source tweet was published in the Twitter15 and Twitter16 datasets. It can be seen that over 50% comments appear within the first 60 min since the original tweet was published. Over 80% of comments appear within the first 100 min since publication of the original tweet. The number of comments plateaus at 1000 min since publication of the original source tweet. Our analysis confirms that it is feasible to use comments for early rumour detection [30]. Our default setting for early rumour detection is 60 min.



(a) An example false-rumour tweet with comments



(b) An example non-rumour tweet with comments

Fig. 4. Different types of tweets and their comments. The green, blue and yellow boxes indicate the Favour, Against and Neutral user stances for comments, and numbers in brackets indicate time delay in minutes. (Color figure online)

We next analyse the user stances expressed in comments for different types of rumours in our datasets. Figure 4 shows examples of different types of tweets. Figure 4(a) shows an example false rumour (misinformation) tweet and its comments. It can be seen that most comments contain questioning phrases such as "No way this is tru, right?" and "Source?" [30]. On the other hand Fig. 4(b) shows an example non-rumour (truthful information) tweet and its comments. It can be seen that there are more presence of Favour stance in the comments. Note also that the first user comment appeared at only 0.67 min after publication of the original tweet.

5.3 Our Stance-Tranfer Models Versus Baseline Models

As shown in Table 2, our stance-based models Stance-CNN+BiLSTM and Stance-BERT yield significantly better performance than all other methods over-

Table 2. Rumour detection results (F1 score) based on the 60-min window. Bold indicates the best result for each column. Stars (*) indicate statistical significance against four baselines with Bonferroni correction under the corrected t-test [21] in 5-fold cross validation experiments.

	Twitter15					Twitter16				
	MacroF1	NR	FR	TR	UR	MacroF1	NR	FR	TR	UR
SVM [30]	0.345	0.380	0.330	0.320	0.350	0.338	0.420	0.190	0.330	0.410
MT-ES [15]	0.460	0.350	0.480	0.600	0.410	0.470	0.390	0.480	0.600	0.410
GRU-RNN [12]	0.644	0.684	0.634	0.688	0.571	0.609	0.617	0.715	0.577	0.527
TD-RvNN [16]	0.700	0.630	0.710	0.800	0.660	0.695	0.580	0.670	0.840	0.690
Stance-CNN+LSTM	0.735*	0.680	0.735	0.785	0.740	0.740*	0.690	0.680	0.780	0.810
Stance-BERT	**0.823**	**0.850**	**0.796**	**0.852**	**0.794**	**0.825***	**0.826**	**0.766**	**0.856**	**0.850**

all. Especially Stance-BERT performs consistently the best for each class. Only for the True Rumour class, it seems that stance-CNN+LSTM performs slightly worse than TD-RvNN, the propagation tree-based model. This can be explained by that for the true rumours, the stance information is harder to capture. It appears that the tree-structure neural network TD-RvNN model performs worse than our models in general. It confirms that the structural information can contribute the rumor detection to some extent, but for early detection, the average length of tree nodes can only get up to 5, and can not capture sufficient propagation signals for effective rumour detection.

It can be observed that the SVM and MT-ES models performance badly compared with other baselines. Even though the SVM model uses some expression to capture the stance information from user comments, but only 19.6% and 22.2% tweets contains these keywords. It fails due to very low recall across all classes and results in the low F_1 scores across each class. The unsatisfactory performance of MT-ES shows that the multi-task formulation of stance and rumour detection is far less effective than our transfer learning formulation for the rumour detection task.

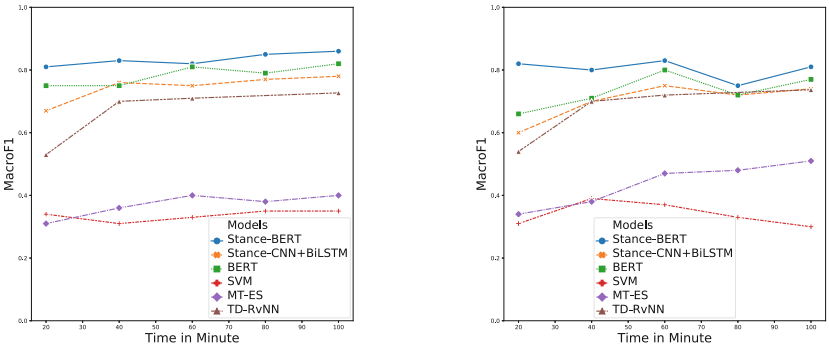
Figure 5 plots the performance of different models in terms of the size of time windows, from 20 min to 100 min, after publication of the source tweet. It can be seen that our stance-BERT model can achieve better performance at the very early stage. The stable performance of Stance-BERT confirms the strong language signals for stance in the early user comments.

5.4 Stance-Based Models Versus Non-stance Models

To evaluate the utility of stance features for rumour detection, we compare our models Stance-CNN+BiLSTM and Stance-BERT against their non-stance variants. As shown in Table 3, Stance-CNN+biLSTM outperforms its non-stance counterpart CNN+biLSTM for the overall MacroF1, and generally outperforms CNN-LSTM for each class. Stance-BERT always outperforms its variants by big margins. Note that Stance-BERT based on tweet-comment pairs outperforms Stance-BERT (comment) based on comments. Moreover, Stance-BERT always

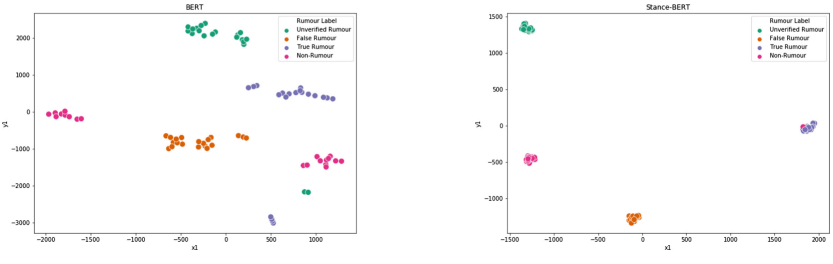
outperforms the other non-stance models BERT(comment), BERT(tweet) and BERT(tweet-comment). These results confirm our hypothesis that the stance feature extracted from user comments data can effectively contribute to rumour detection at the early stage. Moreover our approach of modelling stance for tweet-comment pairs is especially effective.

By transfer learning using the language model BERT, it better captures the language features. In more specific terms, BERT can adjust the weights associated with the model to better represent text originating from comments. This means that during classifier fine-tuning, the starting points of the weights are closer to values that correctly model Twitter data. Closer values mean that the model has a better chance of finding good representations, even with very limited amount of training data.



(a) Twitter15 (b) Twitter16

Fig. 5. Early rumour detection accuracy at different time windows



(a) BERT (b) Stance-BERT

Fig. 6. t-NSE of [CLS] hidden state

To evaluate the utility of stance transfer, we randomly selected 80 samples within the 40-min window from the Twitter15 dataset and use t-SNE [17] to visualize the embeddings of [CLS] for BERT (without stance transfer) and Stance-BERT, which shows the hidden state for sequence embedding. As shown in Fig. 6, Stance-BERT clearly performs better than BERT by grouping the same type of rumours into clusters. It confirms that transferred stance knowledge work effectively with rumour data. In addition, the clear boundaries among different types of rumours shows that strong stance signals exist in the user comments, which confirms our hypothesis that stance can help directly on rumour detection at the early stage.

Table 3. Results (F1 score) for comparing stance models against non-stance models. Best results for each column are in bold. Stars (*) indicate statistical significance with Bonferroni correction under corrected t-test [21] in five-fold cross validation experiments.

	Twitter15					Twitter16				
	MacroF1	NR	FR	TR	UR	MacroF1	NR	FR	TR	UR
Stance-CNN+BiLSTM	0.735*	0.735	0.680	0.735	0.785	0.740*	0.690	0.680	0.780	0.810
CNN-LSTM	0.682	0.590	0.794	0.644	0.700	0.664	0.560	0.602	0.708	0.784
Stance-BERT	**0.823***	**0.850**	**0.796**	0.852	0.794	**0.825***	0.826	**0.766**	0.856	**0.850**
Stance-BERT(comment)	0.747	0.712	0.747	0.810	0.717	0.677	0.683	0.580	0.767	0.677
BERT(comment)	0.708	0.744	0.670	0.676	0.740	0.660	0.728	0.456	0.740	0.722
BERT(tweet)	0.762	0.784	0.710	0.824	0.730	0.781	0.802	0.656	**0.862**	0.804
BERT(tweet-comment)	0.814	0.836	0.774	**0.858**	0.786	0.797	**0.828**	0.718	0.846	0.796

6 Conclusion

We proposed stance transfer learning models based on user comments for early detection of rumours on Twitter. To address the lack of stance annotation for user comments on Twitter, we proposed to design deep CNN model and fine-fune BERT model to learn stance representation for user comments via transfer learning from public resources. We further propose CNN-BiLSTM and BERT-based models to integrate stance representation into the representation for tweets for rumour detection. Experiments on two public Twitter datasets showed that user comments contain early signals for detection rumour tweets. Especially our model based on BERT achieves consistently good performance for early rumour detection and significantly outperforms state-of-the-art baselines. For future work, we will investigate making use of non-content information to further improve the performance of early rumour detection.

Acknowledgement. This research is supported in part by the Australian Research Council Discovery Project DP200101441.

References

1. Augenstein, I., Rocktäschel, T., Vlachos, A., Bontcheva, K.: Stance detection with bidirectional conditional encoding (2016). arXiv preprint arXiv:1606.05464
2. Cao, Z., Li, W., Li, S., Wei, F.: Improving multi-document summarization via text classification. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
3. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on Twitter. In: Proceedings of the 20th International Conference on World Wide Web, pp. 675–684. ACM (2011)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018). arXiv preprint arXiv:1810.04805
5. Dey, K., Shrivastava, R., Kaushik, S.: Topical stance detection for twitter: a two-phase LSTM model using attention. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) ECIR 2018. LNCS, vol. 10772, pp. 529–536. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76941-7_40
6. Domm, P.: False rumor of explosion at white house causes stocks to briefly plunge; Ap confirms its twitter feed was hacked. CNBC. COM, vol. 23 (2013)
7. Dungs, S., Aker, A., Fuhr, N., Bontcheva, K.: Can rumour stance alone predict veracity? In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 3360–3370 (2018)
8. Gorrell, G., et al.: 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. Association for Computational Linguistics, Minneapolis, Minnesota, USA, June 2019. https://doi.org/10.18653/v1/S19-2147https://www.aclweb.org/anthology/S19-2147
9. Gupta, A., Kumaraguru, P., Castillo, C., Meier, P.: TweetCred: real-time credibility assessment of content on twitter. In: Aiello, L.M., McFarland, D. (eds.) SocInfo 2014. LNCS, vol. 8851, pp. 228–243. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13734-6_16
10. Kwon, S., Cha, M., Jung, K., Chen, W., Wang, Y.: Prominent features of rumor propagation in online social media. In: 2013 IEEE 13th International Conference on Data Mining, pp. 1103–1108. IEEE (2013)
11. Liu, X., Nourbakhsh, A., Li, Q., Fang, R., Shah, S.: Real-time rumor debunking on twitter. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 1867–1870. ACM (2015)
12. Ma, J., et al.: Detecting rumors from microblogs with recurrent neural networks. In: Ijcai, pp. 3818–3824 (2016)
13. Ma, J., Gao, W., Wei, Z., Lu, Y., Wong, K.F.: Detect rumors using time series of social context information on microblogging websites. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 1751–1754. ACM (2015)
14. Ma, J., Gao, W., Wong, K.F.: Detect rumors in microblog posts using propagation structure via kernel learning. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 708–717 (2017)
15. Ma, J., Gao, W., Wong, K.F.: Detect rumor and stance jointly by neural multi-task learning. In: Companion of the the Web Conference 2018 on the Web Conference 2018, International World Wide Web Conferences Steering Committee, pp. 585–593 (2018)

16. Ma, J., Gao, W., Wong, K.F.: 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), vol. 1, pp. 1980–1989 (2018)
17. Maaten, L.V.D., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(Nov), 2579–2605 (2008)
18. Mendoza, M., Poblete, B., Castillo, C.: Twitter under crisis: can we trust what we RT? In: Proceedings of the First Workshop on Social Media Analytics, pp. 71–79. ACM (2010)
19. Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: Semeval-2016 task 6: detecting stance in tweets. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 31–41 (2016)
20. Mohtarami, M., Baly, R., Glass, J., Nakov, P., Màrquez, L., Moschitti, A.: Automatic stance detection using end-to-end memory networks (2018). arXiv preprint arXiv:1804.07581
21. Nadeau, C., Bengio, Y.: Inference for the generalization error. In: Advances in Neural Information Processing Systems, pp. 307–313 (2000)
22. Qazvinian, V., Rosengren, E., Radev, D.R., Mei, Q.: Rumor has it: identifying misinformation in microblogs. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1589–1599. Association for Computational Linguistics (2011)
23. Rath, B., Gao, W., Ma, J., Srivastava, J.: From retweet to believability: utilizing trust to identify rumor spreaders on twitter. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pp. 179–186. ACM (2017)
24. Riedel, B., Augenstein, I., Spithourakis, G.P., Riedel, S.: A simple but tough-to-beat baseline for the fake news challenge stance detection task (2017). arXiv preprint arXiv:1707.03264
25. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor. Newslett. **19**(1), 22–36 (2017)
26. Shu, X., Qi, G.J., Tang, J., Wang, J.: Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 35–44. ACM (2015)
27. Wu, K., Yang, S., Zhu, K.Q.: False rumors detection on Sina Weibo by propagation structures. In: 2015 IEEE 31st International Conference on Data Engineering, pp. 651–662. IEEE (2015)
28. Yang, F., Liu, Y., Yu, X., Yang, M.: Automatic detection of rumor on Sina Weibo. In: Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, p. 13. ACM (2012)
29. Zarrella, G., Marsh, A.: Mitre at semeval-2016 task 6: Transfer learning for stance detection (2016). arXiv preprint arXiv:1606.03784
30. Zhao, Z., Resnick, P., Mei, Q.: Enquiring minds: early detection of rumors in social media from enquiry posts. In: Proceedings of the 24th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, pp. 1395–1405 (2015)