GWU NLP at SemEval-2019 Task 7: Hybrid Pipeline for Rumour Veracity and Stance Classification on Social Media

Sardar Hamidian and Mona Diab
Department of Computer Science
The George Washington University
Washington DC, USA
{sardar, mtdiab} @gwu.edu

Abstract

Social media plays a crucial role as the main resource news for information seekers online. However, the unmoderated feature of social media platforms lead to the emergence and spread of untrustworthy contents which harm individuals or even societies. Most of the current automated approaches for automatically determining the veracity of a rumor are not generalizable for novel emerging topics. This paper describes our hybrid system comprising rules and a machine learning model which makes use of replied tweets to identify the veracity of the source tweet. The proposed system in this paper achieved 0.435 F-Macro in stance classification, and 0.262 F-macro and 0.801 RMSE in rumor verification tasks in Task7 of SemEval 2019.

1 Introduction

The number of users who rely on social media to seek daily news and information rises daily, but not all information online is trustworthy. The unmoderated feature of social media makes the emergence and diffusion of misinformation even more intense. Consequently, the propagation of misinformation online could harm an individual or even a society. Most of the current approaches on verifying credibility perform well for the unfold topics which are already verified by a trustworthy resource (Qazvinian et al., 2011; Hamidian and Diab, 2015). However, the performance suffers when it comes to real life application for dealing with the emerging rumors which are priorly unknown. Identifying the emerging rumor and veracity of the rumor by relying on previous observations is a challenging task as the new emerging rumor could be entirely new regarding the event, propagation pattern, and also the provenance. Despite these challenges, many researchers have been studying the generalizable metrics that could be aggregated from the source, replied posts, or network information (Vosoughi et al., 2018b; Kochkina et al., 2018; Grinberg et al., 2019). Our first mission in this paper is to automatically determine the veracity of rumors as part of the SemEval task. SemEval is an ongoing shared task for evaluations of computational sentiment analysis systems. Task 7 (RumourEval19) (Gorrell et al., 2019) is one of the twelve tasks, consisting of two subtasks. Task A is about stance orientation of people as supporting, denying, querying or commenting (SDQC) in a rumor discourse and Task B is about the verification of a given rumor. We propose a hybrid model with rules and a neural network machine learning scheme for both tasks. For task A we rely on the text content of the post, and its parent. In Task B not only do we aggregate contextual information of the source of the rumor but also using the veracity orientation of the others in the same conversation. We devise some rules to improve the performance of the model on query, deny, and support cases which are relatively essential classes in the verification tasks. Integrating the rule-based component we could reach a better performance in both tasks in comparison with a model which only relied on a machine learning approach.

2 Related work

There are several studies about the behavior of misinformation on social media, how it is distinguished and how social media users react to it. Most of these studies use data from Twitter since it has an infrastructure which allows researchers to access network information and meta information of all the users through Twitter APIs. In this section, we mainly focus on machine learning approaches in the study of rumor credibility and stance on Twitter.

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 1115–1119
Minneapolis, Minnesota, USA, June 6–7, 2019.

One of the earliest work and the most relevant work to this task is that reported in Qazvinian et al. (2011), which addresses rumor detection (rumor/Not-rumor/undetermined) and opinion classification (deny/support/question/neutral) on Twitter using content-based as well as microblog-specific meme features. According to this work, content-based features performed better than meta information and network features for rumor identification and opinion classification tasks. In another study (Castillo et al., 2013), leveraged both information cascade and content features of the tweets by applying a supervised mechanism to identify credible and newsworthy content. According to Castillo's work "confirmed truth," or the rumors which are verified as true, are less likely to be questioned than false rumors regarding their validity. In mor recent study, (Vosoughi, 2015) proposes his two-step rumor detection and verification model on the Boston Marathon bombing tweets. The Hierarchical-clustering model is applied for rumor detection, and after the feature engineering process, which contains linguistic, user identity, and pragmatic features, the Hidden Markov model is applied to find the veracity of each rumor. Vosoughi (2015) also analyses the sentiment classification of tweets using the contextual Information, which shows how tweets in different spatial, temporal, and authorial contexts have, on average, different sentiments. In his recent work (Vosoughi et al., 2018a) he analyzed the spread of false and true news on Twitter on a large dataset. According to his research, fake news is more likely to diffuse deeper and longer in the information network than sound news. Moreover, his research suggests that false news are more novel and likely to be shared in comparison to the true news. Vosoughi et al. (2018a) also studied the false and true stories from an emotional perspective. According to his work, false stories inspired fear, disgust, and surprise in replies, while true stories inspired anticipation, sadness, joy, and trust.

3 Dataset

The dataset provided for this task contains Twitter and Reddit conversation threads associated with rumors about nine different topics on Twitter and thirty different topics on Reddit. The Ottawa shootings, Charlie Hebdo, the Ferguson unrest, Germanwings crash, and Putin missing are some of the rumors in this dataset. The overall size of the data including the development and evaluation set is 65 rumors on Reddit and 37 rumors with 381 conversations on Twitter. Table 1 illustrates all the information of underlying replies and source rumor in both social media platforms.

	Reddit		**Twitter**	
	#Src	*#Rep*	*#Src*	*#Rep*
Training	30	667	297	4222
Development	10	426	28	1021
Evaluation	25	736	56	1010
Total	65	1829	381	6253

Table 1: Number of source (Src) conversations and replies (Rep) on Reddit and Twitter in the training, development and Evaluation sets.

3.1 Data insight

Figure 1 shows the distribution of the tags for both tasks across different platforms. According to the table, the stance orientation of the rumor conversations varies between Twitter and Reddit. In general, Reddit users leave more comments than Twitter users and this is regardless of the rumor veracity. In false rumors Twitter conversations are more oriented toward denial than Reddit's conversations; however, Twitter users support and deny false rumors to relatively the same extent. Twitter users are more supportive and ask more questions in regards to true rumors than the Reddit users, but they both deny true rumors to almost the same amount.

Interestingly, in both platforms, people question unverified rumors more than true and false rumors. For the source of conversation, Reddit and Twitter are significantly different. Regardless of the veracity, the source in Reddit conversations is more skewed to the query than the other stance tags, while Twitter is more toward the support. Despite some common characteristics Reddit and Twitter users behave differently when it comes to rumors. Reddit users do not deny the TRUE or UNVERIFIED rumors and question more when the rumor is false, yet Twitter users support more without any inquiries. It is worth noting that the conclusions mentioned in this section could only be valid for the data provided and in other conditions the same correlations might not be present.

4 System Description

For both tasks, we mainly rely on the content to determine the stance and verification of the

Source/Replies	Task B	Task A	Socialmedia Reddit	Socialmedia Twitter
Replies	False_Source	comment	93.58%	68.29%
		deny	3.95%	12.15%
		query	1.73%	6.61%
		support	0.74%	12.96%
	True_Source	comment	88.89%	68.67%
		deny	5.56%	6.29%
		query	1.85%	8.57%
		support	3.70%	16.47%
	Unverified_Source	comment	88.31%	68.65%
		deny	6.49%	7.84%
		query	3.90%	9.20%
		support	1.30%	14.31%
Source	False	comment		3.23%
		deny	11.76%	6.45%
		query	64.71%	
		support	23.53%	90.32%
	True	comment		0.73%
		deny		2.19%
		query	85.71%	
		support	14.29%	97.08%
	Unverified	comment		8.16%
		deny		2.04%
		query	83.33%	
		support	16.67%	89.80%

Figure 1: The distribution percentage of stance and verification tags on Twitter and Reddit dataset. "TaskB_Source" (exp. False_Source) indicate the verification tag of the source of conversation.

sources in the conversation. Our primary analysis in the insight section showed that there is a significant correlation between the two tasks. For the unverified rumors, the stance orientation is more toward queries rather than concluding support or denial; on the other hand, for true rumors, people are more likely to support or comment on the conversation than question or deny. Therefore, stance is key information to determine the veracity of the source rumor in the conversation. Task A is a four-way classification experiment in which we propose a hybrid model including a neural network-based (NN) model to encode the contextual representation of the post and its parent and then a rule-based model which is mainly designed to improve the performance on the minority classes including "support," "deny," and also "query." Task-B is a three-way classification task (True, False, and Unverified) in which we rely on both source and conversation content. We expand the veracity tags for the source of the conversation to the underlying posts and create a new set of veracity tags including Source_True, Source_False, and Source_Unverified (Six-way classification). We first apply a sequential neural network-based approach to identify the veracity tag of the source and also replied posts. From the sources with a low confidence value a voting mechanism is applied among all the posts in associated conversation, i.e. if the majority of the tweets in the conversation classified as Parent_true then the source of the conversation will be labeled as True.

4.1 Neural Network Approach

Given the success of recurrent neural networks (RNN) on language problems, we build a standard Bi-LSTM network for both tasks as illustrated in Figure 2. We also investigated the effectiveness of multitask learning in this experiment by sharing the information of two tasks in the same pipeline, but it does not lead to noticeable improvement in the performance.

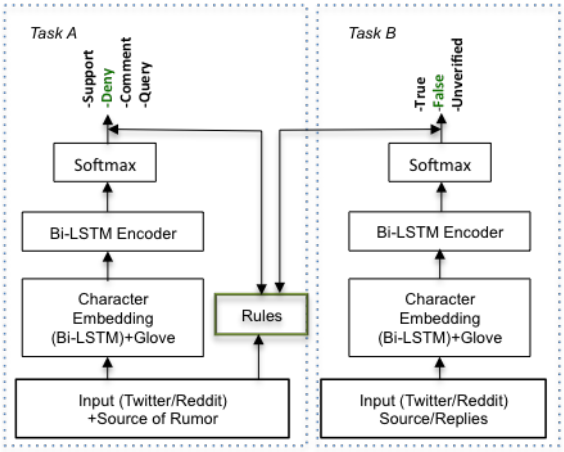

Figure 2: Illustration of the hybrid network comprising the rule-based and Bi-LSTM-Softmax network on Task A and Task B.

4.1.1 Input Representations

Recent studies on NLP applications are reported to have good performance applying the pre-trained word embedding (Socher et al., 2013). We adopted two widely-used methods including the character embedding and pre-trained word vectors, i.e., GloVe (Pennington et al., 2014). We use a Bi-LSTM network to encode the morphology and word embeddings from characters. Intuitively the concatenated fixed size vectors $W_{Character}$ capture word morphology. $Word_{Character}$ is concatenated with a pre-trained word embedding from GloVe $W_{pre-trained-Glove}$ to get the final word representation. For Contextual Encoding, once the word embedding is created we use another Bi-LSTM layer to encode the contextual meaning from the sequence of word vectors $W_1, W_2, ..., W_t$

	Task A						Task B				
	Accuracy	Macro-F	Support	Query	Deny	Comment	Accuracy	Macro-F	True	False	Unverified
Dev	0.802	0.487	0.420	0.586	0.058	0.885	0.315	0.187	0.418	0.0	0.142
Test	0.796	**0.435**	0.446	0.408	0.0	0.886	0.382	**0.262**	0.525	0.0	0.260

Table 2: Accuracy and F score (macro-averaged) results on the development and test sets of Task A and B.

and consequently obtain a vector representation of a sentence from the final hidden state of the LSTM layer. The input representation would capture the word level syntax, semantics and contextual information. For Twitter data, we only rely on the tweet content for both source and replies, but for the Reddit rumor we use the "title" and "selftext" and only "body" for the replies.

4.2 Rule-based components

The first analysis of the data showed that stance knowledge could significantly help the determination of the source rumor in the conversation. However, due to imbalanced data, identifying the minority classes including deny, query, and support is challenging. We devise a new set of rules to improve the performance of Task A. Using the confidence values of the NN model we only selected the cases with low confidence for the rule-based experiments. We relied on simple rules for each stance class. For Query, a new set of rules was designed to identify the query cases using question marks and syntactic information of the sentence. For Deny, we calculated the cosine similarity of the source and response in addition to sentiment differences of the source and replied post. For the support cases, we mainly relied on the URL and picture existence in the content. The domain of the URL checked for being a fact-checking or news source. We also checked the existence of the picture in the post and consider that as one of the conditions for the supporting tweets.

5 Experimental Setup

The shared task dataset is split into training, development and test sets by the SemEval-2019 task organizers. We conducted and tuned the optimal set of hyperparameters by testing the performance on the development set and the output of the final model on the test set evaluated by the organizers. The statistics of the dataset are shown in Table 1.

5.1 Preprocessing

We applied various degrees of preprocessing on the content, we first removed the very short, deleted, and also the removed cases (Those that are labeled [deleted] or [removed] by the task organizers) from the dataset. We replaced the URLs from news sources with the token NURL and all the fact-checking URLs with FURLs. For compound words and hashtags, we used a simple heuristic. If the hashtag or a word contained an uppercase character in the middle of the word, then we split it before the uppercase letter. For instance, #PutinMissing are separating into two words Putin Missing.

5.2 Training

For all of the pipelines, the network is trained with backpropagation using Adam (Kingma and Ba, 2014), Root Mean Square Propagation (RmsProp), and Stochastic Gradient Descent (SGD) optimization algorithms. The parameters get updated in every training epoch. The character and Glove pre-trained embedding size [100, 200, 300] are examined with batch size 20 with 100 epochs. The training is stopped after no improvements in five consecutive epochs to ensure the convergence of the models. The highest performance on the development set was achieved under the following parameters: hidden size of Bi-LSTM (100); optimization (RMSprop); initial learning rate (0.003); L2 ($Lambda = 0.1$); character and word embedding size (300, 100); dropout size (0.3).

6 Result and Evaluation

In this section, we discuss the experimental results in both tasks. Table 2 shows overall and per category results for Task A and B. The proposed model achieved 0.435 F-Macro in stance classification, and 0.262 F-macro and 0.801 RMSE in rumor verification tasks. In overall evaluation, we ranked as the third group in Task B and tenth in Task A out of twenty-five teams.

7 Conclusion

Identifying rumor veracity is an important and challenging task. Our first mission in this paper is to automatically determine the veracity of rumors as part of the SemEval task. We proposed a hybrid model comprising the rules and NN machine learning approach to identify the stance in the rumor conversation and the veracity of the source in Twitter and Reddit datasets. The proposed system achieved the third best performance for RumourEval, Task7 of Semeval 2019.

—

References

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2013. Predicting information credibility in time-sensitive social media. *Internet Research*, 23(5):560–588.

Genevieve Gorrell, Kalina Bontcheva, Leon Derczynski, Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2019. SemEval-2019 Task 7: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of SemEval*. ACL.

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378.

Sardar Hamidian and Mona Diab. 2015. Rumor Detection and Classification for Twitter Data. *SOTICS 2015: The Fifth International Conference on Social Media Technologies, Communication, and Informatics*, (c):71–77.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Qazvinian et al. - 2011 - Rumor has it Identifying Misinformation in Microblogs(2). *Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

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

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018a. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018b. The spread of true and false news online. *Science*, 359(6380):1146–1151.

—