UWaterloo at SemEval-2017 Task 8: Detecting Stance towards Rumours with Topic Independent Features

Hareesh Bahuleyan and **Olga Vechtomova**
University of Waterloo, ON, Canada
{hpallika, ovechtomova}@uwaterloo.ca

Abstract

This paper describes our system for subtask-A: SDQC for RumourEval, task-8 of SemEval 2017. Identifying rumours, especially for breaking news events as they unfold, is a challenging task due to the absence of sufficient information about the exact rumour stories circulating on social media. Determining the stance of Twitter users towards rumourous messages could provide an indirect way of identifying potential rumours. The proposed approach makes use of topic independent features from two categories, namely cue features and message specific features to fit a gradient boosting classifier. With an accuracy of 0.78, our system achieved the second best performance on subtask-A of RumourEval.

1 Introduction

In the recent years, with the increasing popularity of smartphones, social media has become one of the top sources of news. However, because all the content is user-generated, the truth behind such news stories may become difficult to verify. Spread of misinformation during the event of an emergency can potentially have negative impacts. Although a few studies in the literature have developed rumour classification algorithms for Twitter (Qazvinian et al., 2011), these studies assume that the circulating stories about a topic or an event are known *a priori* (Eg: *Is Barack Obama muslim ?*). On the other hand, identifying rumour stories for breaking news events, as they unfold, is even more challenging (Zubiaga et al., 2016b). This is because during these early stages, the exact rumour stories propagating about the event are still unknown.

In such a scenario, studying the conversation between users discussing the event on Twitter can possibly give insights about the veracity of a circulating rumour story (Zubiaga et al., 2016c). By making use of the so-called 'wisdom of the crowd', the idea here is to understand how other users respond to rumourous tweets. It would be useful to identify if users may reply with an intent to support the story, deny the rumour by providing counter evidence or pose questions about the information stated (Zubiaga et al., 2016a). Collating the stance of other users could indirectly help in resolving the veracity of a rumour.

The rest of this paper is organized as follows. Section 2 is a brief overview of the task. The features used and the modeling technique are described in section 3. The results are analyzed in section 4 and the conclusions from the study are provided in section 5.

2 Task Description

The objective of subtask-A of RumourEval was to identify the stance of Twitter users towards rumour tweets. Given a rumourous tweet (source) and its conversation thread, the participants were required to classify the stance of each tweet (including the source tweet) with respect to the underlying rumour (Derczynski et al., 2017). The type of interaction could be one of the following:

1. Support(S): responding user supports the veracity of the rumour
2. Deny(D): responding user denies the veracity of the rumour
3. Query(Q): responding user demands additional evidence
4. Comment(C): responding user's tweet is not useful in determining the veracity of the rumour

The training dataset consisted of 4519 tweets from

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 452–455,
Vancouver, Canada, August 3 - 4, 2017.

eight breaking news stories. The test set had 1049 tweets corresponding to a mix of topics from different events.

3 System Description

Breaking news events, as they unfold on social media, may not have sufficient topic-specific information that could assist in rumour identification. For this reason, we chose to design topic independent features for the task of rumour stance classification. Our hypothesis was that the presence of specific words in the reply tweets could potentially be indicative of reply type.

Prior to feature extraction, the following data pre-processing steps were carried out: (1) removal of quoted text (reply tweets at times quote the source tweet), (2) discarding URLs, unicode characters, HTML tags, and (3) stripping out the extra whitespaces and carriage returns in the text.

We began by manually inspecting tweet messages in the training dataset to come up with an initial hand-curated list of word features. On further analyzing these features, it was found that these words could be categorized into meaningful groups. Such 'cue words' have previously been reported to be useful in identifying an author's certainty in journalism (Soni et al., 2014), determining veracity of rumours (Reichel and Lendvai, 2016) and detecting disagreement in online dialogue (Misra and Walker, 2013). As listed in Table 1, the first five categories of the cue features are **Belief**, **Report**, **Doubt**, **Knowledge**, **Denial**. The presence of belief or knowledge words could be indicative of a reply where the author expresses his support. As for doubt or denial word cues, they are more likely to be used when the replying author wishes to convey his disagreement. On the other hand, report cue words could be present in either a supporting tweet or a denying tweet. Table 2 provides example tweet messages containing different cue words and their corresponding true class-labels from the original dataset .

Internet slang and curse words are more likely to be present in reply tweets which are of type 'comment'. While negation words were useful in identifying denying replies, the occurrence of question words in the text were very informative in capturing query type replies. We have a list of certain other cue words, which could not be fit into any particular category, but were useful in this 4-class classification problem. The cue word feature categories along with examples are shown in Table 1. In total, there were 153 such features.[1]

Feature	Example Words
Belief	*assume, believe, apparent, perhaps, suspect, think, thought, consider*
Report	*evidence, source, official, footage, capture, assert, told, claim, according*
Doubt	*wonder, allege, unsure, guess, speculate, doubt*
Knowledge	*confirm, definitely, admit*
Denial	*refuse, reject, rebuff, dismiss, contradict, oppose*
Curse Words & Internet Slang	*lol, rofl, lmfao, yeah, stfu, aha, wtf, shit*
Negation Words	*no, not, no one, nothing, never, don't, can't, hardly*
Question Words	*when, which, what, who, how, whom, why whose*
Others	*irresponsible, careless, liar, false, witness, untrue, neglect, integrity, murder, fake*

Table 1: Set of cue features and examples

Example Tweet	Cue Word Type	Reply Type
*@TroyBramston Source from Ray Hadley shows **confirmed** same report of gunman claiming there are four packages around Sydney*	Knowledge/ Report	Support
*@PhilSerrin Me thinks you like to emote in suppositions. Truth is, you don't know what happened, but want to **speculate**.*	Doubt	Deny
*@DaveBeninger @SheilaGunnReid Canadian news **contradicts** this*	Denial	Deny
*@Manning_Eli_1 @TheAnonMessage2 I **thought** the same thing*	Belief	Support

Table 2: Example tweets with cue words

Apart from the cue word features discussed earlier, certain other tweet specific features were also used as part of our model. These message level features provide information about the writing style, such as the presence of punctuation marks, Twitter-specific characters (such as #, @) and number of words/characters in the message. The entire list of features under this category have been summarized in Table 3. For calculating the sentiment polarity score, the lexicon based social media sentiment calculator, VADER, developed by Hutto and Gilbert (2014), was used.

[1]The cue word feature list used in this study is available at https://github.com/HareeshBahuleyan/rumour-eval/blob/master/cue_word_list.txt

It is to be noted that all of the features discussed in this section (except for *similarity*) were extracted from the reply tweets in the dataset. The task also required the source tweet to be classified as one of the four reply types. Since there wasn't enough data to train a separate model for predicting the label of source tweets, we made an assumption that all source tweets were 'supporting' the rumour, which was the majority class in the training set.

Feature	Description
Word Count	Number of words in the tweet
Capital Words	Count of words in ALL CAPS
Punctuation	Number of '?', '!' and '.'
Character Count	Number of characters
Sentiment	VADER sentiment score
Similarity	Cosine similarity between source and reply tweets
Hashtag	Count of hashtags
@user	Count of *@user* mentions
Part-of-Speech	A vector of POS tag counts

Table 3: Set of tweet features and description

The numeric features, most of which were counts of specific characters or words, were used for training a supervised classification algorithm, specifically Gradient Boosting. Boosting is an additive and iterative tree-based supervised machine learning approach where a strong classifier is sequentially constructed from multiple weak learners. The XGBoost implementation of the gradient boosting algorithm was utilized in this study (Chen and Guestrin, 2016). The hyperparameters were tuned and set to be as follows:

1. **n_estimators** = 100: Refers to the number of trees to be grown to fit the model.
2. **max_depth** = 9: Number of splits for each of the weak learner trees.
3. **sub_sample** = 0.8; Each tree uses a random subset of size 80% of the original training set size.

Baseline: We also construct a unigram model as a baseline, which is compared against the proposed model that uses topic independent features. Because the unigram terms are unfiltered, the baseline model uses topic specific features as well.

4 Results

In this section, we discuss the performance of the model with topic independent features. We also compare it against the unigram baseline. Classification accuracy was the evaluation metric for this RumourEval subtask. However, since a majority of the tweets (about 70%) in the dataset belonged to the class label 'comment', we also report the macro-averaged F-score here.

The development set provided by the organizers was the set of tweets corresponding to the topic *germanwings-crash*. This was used for validating the model and determining the best combination of features from among the ones listed in the previous section. The results of the proposed model, in terms of accuracy and F-score, on the development set are shown in Table 4. The model with the proposed set of features is observed to have a reasonable accuracy and F-score for all class labels, except for the 'deny' label, which it found difficult to identify.

The results on the actual test set, which was a mix of all topics, are summarized in Table 5. All models performed similarly in terms of accuracy because a large proportion of the predicted labels belong to 'comment' class. However, the models with the topic independent features outperformed the baseline unigram model in terms of F-score. While the baseline model had an F-score of 0.31, the best combination of the proposed features resulted in an F-score of 0.45. The features were chosen by running validations with different feature combinations on the development dataset. The highest accuracy and F-score was obtained when the following features were discarded from the model: @user, hashtag, similarity, sentiment, characters. The submission with this model made our system the one with the second best performance for subtask A of RumourEval.

We also tried out models with features only from one category. When the cue features alone were used, the F-score was 0.34. On the other hand, the model with only the message specific features provided a higher F-score of 0.42. When all the proposed features were used for the classification task, it resulted in an accuracy of 0.77 with an F-score of 0.44, suggesting that, when used in tandem, the features yield a better result than using only a single category of features.

5 Conclusions

This paper provides a description of our submission for subtask A of RumourEval in which the participants were required to classify the stance of tweets towards rumours. The proposed model used topic independent features from two categories: cue features and message specific features.

Features	Accuracy	F-score	Comment	Deny	Query	Support
Unigrams (Baseline)	0.690	0.32	0.799	0.000	0.000	0.489
Only Cue Features	0.697	0.38	0.804	**0.153**	0.067	0.489
Only Message Specific Features	0.718	0.46	0.802	0.000	0.428	**0.621**
All Proposed Features	**0.729**	**0.51**	**0.813**	**0.153**	**0.450**	0.617
All Features -{@user, hashtag, similarity, sentiment, characters}	0.718	**0.51**	0.803	**0.153**	0.465	0.608

Table 4: Results for different feature combinations on the Development Set - Accuracy and F-score (macro-averaged and per class)

Features	Accuracy	F-score	Comment	Deny	Query	Support
Unigrams (Baseline)	0.750	0.31	0.856	0.000	0.000	0.386
Only Cue Features	0.757	0.34	0.860	0.000	0.085	**0.406**
Only Message Specific Features	0.763	0.42	0.858	0.000	0.432	0.400
All Proposed Features	0.770	0.44	0.867	0.027	0.473	0.388
All Features -{@user, hashtag, similarity, sentiment, characters}	**0.780**	**0.45**	**0.869**	**0.052**	**0.494**	0.397

Table 5: Results for different feature combinations on the Test Set - Accuracy and F-score (macro-averaged and per class)

A gradient boosting classifier was implemented for this 4-class classification problem. Our system ranked second in terms of accuracy. For future work, we plan to investigate if the tree structure of the conversation could provide insights about the reply type.

References

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. *arXiv preprint arXiv:1603.02754* .

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 SemEval*. ACL.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Amita Misra and Marilyn A Walker. 2013. Topic independent identification of agreement and disagreement in social media dialogue. In *Conference of the Special Interest Group on Discourse and Dialogue*. page 920.

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*. Association for Computational Linguistics, pages 1589–1599.

Uwe D Reichel and Piroska Lendvai. 2016. Veracity computing from lexical cues and perceived certainty trends. *arXiv preprint arXiv:1611.02590* .

Sandeep Soni, Tanushree Mitra, Eric Gilbert, and Jacob Eisenstein. 2014. Modeling factuality judgments in social media text. In *ACL (2)*. pages 415–420.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, and Michal Lukasik. 2016a. Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. *arXiv preprint arXiv:1609.09028* .

Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2016b. Learning reporting dynamics during breaking news for rumour detection in social media. *arXiv preprint arXiv:1610.07363* .

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