AndrejJan at SemEval-2019 Task 7: A Fusion Approach for Exploring the Key Factors pertaining to Rumour Analysis

Andrej Janchevski, Sonja Gievska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Rugjer Boshkovikj 16, Skopje, Republic of North Macedonia
andrej.jancevski@students.finki.ukim.mk
sonja.gievska@finki.ukim.mk

Abstract

The viral spread of false, unverified and misleading information on the Internet has attracted a heightened attention of an interdisciplinary research community on the phenomenon. This paper contributes to the research efforts of automatically determining the veracity of rumourous tweets and classifying their replies according to stance. Our research objective was to investigate the interplay between a number of phenomenological and contextual features of rumours, in particular, we explore the extent to which network structural characteristics, metadata and user profiles could complement the linguistic analysis of the written content for the task at hand. The current findings strongly demonstrate that supplementary sources of information play significant role in classifying the veracity and the stance of Twitter interactions deemed to be rumourous.

1 Introduction

Social networks continue to contribute to the way people connect, stay informed and contribute above and beyond what concerns their lives. However, these platforms currently play a crucial role in viral spreading of false information, and no doubt will become even more instrumental in the future. Consequently, it will be increasingly important to devise systems and establish practices of automatically identifying, filtering and labeling false information in order to help people make sense of information dispersed through social channels.

The damaging consequences of malicious intents to misinform, confuse and provoke through social media platforms remain persistently high in the discussions of researchers and practitioners. Perhaps even more so than the benefits of information dissemination. While the veracity of some false information can be determined unambiguously from external sources, it is a major challenge to verify the truthfulness of rumourous postings. Indeed, predictive analysis remains the primary manner in which social platforms could face the challenge of identifying rumours and taking appropriate actions. Over the past several years, research has emerged and at the same time, many challenges remain.

Decoding elusive social phenomenon such as spreading rumours is challenging not only due to their complexity, but also due to the diversity of the underlying causes and heterogeneity in their manifestations. Rumours represent intertwined threads of sensemaking Bordia and Difonzo (2004) that are initiated and spread by people trying to explain, solve and remove uncertainty relating to events and persons that attract public interest Peterson and Gist (1951). We argue that solutions that address these multi-dimensional issues cannot solely rely on natural language processing (NLP). Combining natural language processing with social analytics extends beyond the traditional realms of either technology to a variety of emerging applications, including rumour analysis.

The analysis of rumours can take on different meanings depending on the application domain - this research focuses on identifying the veracity of a rumourous tweet (Task B) and the stance of its replies i.e., classifying responses to a rumourous post as supporting, denying, querying or commenting (SDQC) the claim (Task A), both part of the RumourEval 2019 Gorrell et al. (2019). While this work was developed on the foundations laid out by previous research in the field Derczynski et al. (2017), it is among the rare solutions that explore complementary types of information that could augment the linguistic analysis when analyzing rumourous tweets. After a brief discussion

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 1083–1089
Minneapolis, Minnesota, USA, June 6–7, 2019.

of relevant research that is closely related to our objective, we highlight the primary findings of our research.

2 Related work

Our research follows the line of work of the research groups that have contributed to the discourse on the rumour analysis with several experimental studies. A review of past SemEval related tasks points out that much of the research of the problem of examining the veracity and support for rumours focused more on language analysis and less on utilizing the external information made available by the Twitter dataset. The findings that extend across several studies is how rather simplistic NLP techniques and analysis are capable of obtaining satisfactory results when classifying the support of rumourous posting. The similarities and differences between the work presented in this paper and previous research is discussed.

A model for automatic identification of rumourous tweets and classification of their responses into two denying and supporting classes, presented in Qazvinian et al. (2011), is based on linguistic features, such as, unigrams, bigrams, POS tags, URLs, and hashtags. While, we built upon the experiences of this research, especially in the approach of replacing a multi-class classification problem with a hierarchical pipeline of binary classifiers, our approach differs in a number of aspects. In particular, we further enhance the representation of rumourous posts with a number of features, from word embeddings, sentiment and stylistic features to structural properties of interaction threads. In addition, we argue that a suitable preprocessing of tweets is essential for successive NLP steps, which is in contrast to their decision not to perform any preprocessing of the text.

The research study by Lukasik et al. (2015) extends on the work of Qazvinian et al. (2011), especially by incorporating preprocessing steps, bag-of-words (BoW) and word clustering. The approach presented in this paper follows some of these ideas, although it differs in the scope and the way features are operationalized. For instance, TF-IDF n-gram counts and clusters of Word2Vec embeddings were used in our model.

Based on the premise that psychological and sociological information could play a key role in determining the truthfulness of rumours and inspired by the research of Mihaylova et al. (2018) on fact checking of questions posted on online forums, we have further explored the effect of various metadata and information from users profiles on the performance of the task.

3 Methods

3.1 Dataset

The dataset used for evaluating the model proposed in this research was made available by Zubiaga et al. (2016b). The rumourous tweets and their corresponding interactions threads were harnessed in 2016, verified and labeled by journalists and sociologists using crowdsourcing platforms Zubiaga et al. (2016a). The dataset contains rumourous tweets written in English, associated with nine events, five being news stories and four concerning specific events. Out of a total of 4560 tweets, 297 represent rumourous posts annotated for veracity, while the rest are responses to the original rumour, annotated according to their stance i.e., supporting, denying, querying or commenting the initial claim.

3.2 Preprocessing

Tokenization, part-of-speech tagging, lemmatization, and substitutions were used as preprocessing techniques; the set and order of preprocessing steps varies between features. In addition, a number of context-appropriate corrections of language variations of English e-dialect were performed:

- Characters are converted to lower-case letters; for sentiment feature extraction the original letter case was preserved;
- URLs, numerical sequences, email addresses were replaced with special tokens (e.g., *URL*, *NUM*, *EMAIL*). Special tokens, *QUOTE* and *USER*, were used whenever the original rumour tweet was quoted or a user is mentioned in a response tweet;
- For each hashtag, the # sign is replaced with a special token *HASHTAG*, while the text of the hashtag was kept for further analysis;
- Emojis are identified and each was represented as a different token;
- Consecutive repetitions of a character in a word were contracted to 2 instances of the character;

- All special characters were removed, with the exception of `\s\n\r.,?!:-+` that were treated as separate tokens

3.3 Feature Extraction

Automatic extraction of information related to language, discourse and context is a difficult task. Among different combinations of features, the model that yielded the best result is described. Linguistic analysis for extraction of seven types of features was performed on the preprocessed text of each tweet. In addition, we have analyzed the content of user profiles, including profiles of the initiators of rumours and those replying to initial posts. Our assumption was that while tweet analysis captures indicators pertaining to a particular rumour thread, the language analysis of a user profile could provide insights into personality, attitudes and online behavior of a user.

Language style features - In examining previous research on rumour analysis, we found that stylistic features are frequently presented as simple statistical features which had an effect on the performance of the task. We have considered the following: number of words and sentences, average number of words per sentence, ratio of word vs. non-word tokens, percentage of present dictionary words, mean and variance of word length, and percentage of unique words. The same set of stylistic features were calculated for the content of each tweet as well the text content found in the profile of the user who has posted the tweet.

Language model n-grams - Unigrams, bigrams and trigrams were extracted from the tweets and users profiles, keeping only 1/8, 1/16 and 1/20 of the most frequent unique unigrams, bigrams and trigrams respectively. Six vocabularies were created; three n-gram vocabularies for the tweets and three vocabularies of n-grams found in user profiles. Term frequency-inverse document frequency (TF-IDF) values were calculated and utilized as the final language model features.

POS tags - In accordance with the well-established practice to complement the language model n-grams with their part-of-speech tags, TF-IDF values were calculated for two POS tag vocabularies, one relating to tweets, the other to self-descriptions left by users in their profiles.

Word2Vec embeddings - Stop-words were removed and lemmatization was performed on the tweets as well as user profile text. Two Word2Vec models were trained on the SemEval 2019 dataset, one was trained on the sentences of the tweets, the other on the content of user profiles. The dimensionality of the vectors was set to 500. A context window of size 5 was used, while words with frequencies above 0.001 were subsampled. The two vocabularies of embedding vectors were clustered using the K-Means algorithm; the parameter K was set automatically to ensure that each word cluster will contain an average of 10 items. At last, for each token list, a Bag-of-Centroids feature vector was calculated by counting the word clusters the tokens belong to.

Sentiment features - Consistent with related studies, which suggest the predictive power of affective words on the task relating to detecting deception in online text, we perform a polarity sentiment analysis on the tweets, calculating three polarity scores, positive, negative and neutral for each sentence. The NLTK, Vader Sentiment Intensity Analyzer, was used because of its reported robustness to the style of online e-language (e.g., capitalization, punctuation, slang) Hutto and Gilbert (2014). A sentiment feature vector for each tweet was generated from the mean and variance for each polarity score.

Network structure - Social network communication exchanges (e.g., tree-structured threads of tweets) are naturally represented as graphs (DAG) - nodes represent rumour posts and their responses, while directed edges of the graph associate responses (e.g., reply, comments) with the target of their response. Based on the premise that variations in the structural properties of the underlying rumour threads could play an important role in identifying and classifying the veracity of rumours tweets and their responses, the following network characteristics were used: DFS and BFS priority, degree centrality, betweenness centrality, closeness centrality, HITS hub score and PageRank.

Twitter metadata - Several Twitter metadata were retrieved and included in our model. The following list of information were considered to be relevant and were added as separate features: the number of characters in a tweet, the number of favorites and retweets and the number of days since initial posting. For each user, the information whether the user account is verified, the number of user's followers, the number of statuses, the number of friends, the number of favorites, number of times listed and how long the user has had

a Twitter account were also included. Some additional information on the graphical design choices a user made were also considered. For example, whether there is a background image or a default image was used and the colors selected for the text, border, sidebar, background and hyperlinks.

Similarity measure - Transmission of context shared by the initial rumour tweet and its responses was postulated to be important for the task of classifying the support of the responses (task A). Consequently, it was decided to calculate the differences between the feature vectors of each response tweet and the rumour tweet it replies to.

3.4 Feature Selection

The feature extraction process resulted in a set of approximately 14000 features for each tweet, in addition to the ones resulting from textual analysis of the user profile. Appropriate feature selection is essential for achieving good performance and avoiding overfitting by removing uninformative, redundant or noisy information. First, the features were normalized to the interval [-1, 1] using generalized Min-Max normalization. Then, a Random Forest classifier was used as a basis for calculating the information gain of each feature. The features with an above-average information gain score were selected, reducing the number of features by a factor of 10. It was decided not to eliminate any correlated features as the number of samples is quite low and any detected correlation might be due to statistical falsehood.

3.5 Model training

At the outset of our explorations in order to circumvent the difficulties imposed by a multi-class classification problem, presented as task A, a hierarchical one-vs-rest approach was adopted. To this end, a binary classifier was used to classify each response tweet into two classes, comment or non-comment; the next step classifies each non-comment tweet as a query or a non-query type. In the end, each non-query response was classified as being in support or denial of the original rumourous post.

Balancing of the datasets was a necessary step during the training of the models for task A, as the distribution of the original four SDQC classes was not uniform. The balancing process was performed by repeated random sampling. Thirty candidate datasets with uniform class distribution were formed by random sampling the more prevalent class and the best candidate was chosen using 3-fold cross validation with a Naive Bayes classifier.

The models used for Task A and B are ensembles of six different classifiers, including: Naive-Bayes, K-Nearest-Neighbours, Logistic Regression, Support Vector Machine, Neural Network and Random Forest. The ensemble classifier for task A was operationalized as majority voting, while for task B, probability-weighted voting was used (probability weights correspond to the confidence level of the veracity scores).

In order to train and evaluate the ensemble, each dataset was split into a training and a validation set using 3-fold cross-validation with stratified sampling. This process was repeated 10 times and the final evaluation scores were calculated by averaging the scores from each iteration. For some training runs, the parameters of the classifiers had been optimized by 3-fold cross validation using grid search. However, no notable improvements in evaluation results were witnessed.

The system[1] was implemented using Python 3.6, with the ScikitLearn, NLTK, gensim and NetworkX packages Pedregosa et al. (2011); Bird et al. (2009); Řehůřek and Sojka (2010); Hagberg et al. (2008).

4 Discussion of results

The performance results obtained by the model that included all categories of features, discussed in the previous section, when tested on our validation set are presented in Table 1. The same model was used to create our final submission for SemEval Task 7. It is worth noting that the best results were obtained for identifying queries in task A, obtaining an accuracy of 0.784, demonstrating that the model accuracy rivals the performance of domain experts. We would like to highlight the consideration that no external sources were used for determining the veracity of a rumour posting, including the resources deemed as appropriate by the task organizers, which might explain the lower performances on the second task B.

Although the aforementioned results provide compelling evidence suggesting that fusing linguistic analysis, metadata, user profiles and rumour thread structure can lead to satisfactory results for classifying the stance of a response to a

[1]The complete source code is available at: https://github.com/Bani57/rumourEval2019.

Model	Accuracy	Precision	Recall	F1	ROC AUC	Log loss
Task A - Comment	0.767632	0.773194	0.767632	0.766449	0.767632	/
Task A - Query	0.784186	0.790382	0.784186	0.783002	0.784190	/
Task A - Support/Deny	0.682915	0.687301	0.682915	0.681009	0.682702	/
Task B - Veracity	0.681707	0.693103	0.681707	0.677606	0.681707	0.600458

Table 1: Evaluation metrics on a randomly-created validation set

rumourous tweet, we have conducted two ablation studies, to find evidence of the performance gains that could be contributed to each category of features on both tasks.

The present study suggests that text analysis of rumourous tweets is the most important, yet not the sole constituent element when detecting the veracity of rumourous tweet and distinguishing the stance of the social response. While we are not in position to ascribe sound theoretical reason to all effects, we present the trends that appeared interesting and highlight the sensitivity of the performance results towards a particular category of features.

Table 2 displays the F1 results of ablation study when training the models for task A, removing one category of features at a time. The results affirm that capturing the metadata relating to the users profile and structural properties of the tree-like threads of tweet exchange complement the linguistic features and improve the predictive accuracy, especially for distinguishing between query, support and denial stance in rumour replies.

The ablation analysis strongly demonstrates that Twitter metadata has the most dramatic effect on distinguishing **comments** from all other rumour responses - removal of this category results in lowering the F1 values by 0.17. It appears that user's historical and behavioral metadata add to the prediction performance complementing the relevant n-grams in the content of the tweet and user profile. Some of the most relevant indicators in the rumourous comment include: end of sentence punctuation, hashtags, user mentioned in the tweet, pronouns and words such as: *reported*, *happening*, as well as n-grams extracted from the user profile, such as: *blog*, *I am*, *concerned citizen*, *culture*, *enthusiast*, *living*, etc.

Language model and word vectors representation of the content of tweets and user profiles were indicated as better predictors when identifying replies in the form of **queries**, resulting in 0.12 decrease in F1 values, if removed from the model. Some of the most relevant features were not surprisingly related to detecting question forms: *why*, *where*, *question mark at the end of a sentence*, *who was*, *what is*, *confirm*, *need*, and a number of word2vec clusters. In addition, language style, sentiment and Twitter interaction threads have also ranked in the top 10% of the most relevant features.

Sentiment and structural features have a more notable effect on discriminating between **supportive and denying responses**. We could hypothesize that the affective content of a tweet is a crucial indicator when distinguishing positive (confirmative) vs. negative (opposing) opinion toward the source tweet with rumourous claim. The top 10% of the best predictive indicators were the sentiment words, and n-grams, such as: *not*, *believe*, *know*, *oh*, *ugh*, *such*, *yeah*, *understand*, socially offensive words etc.

Table 3 shows the cross-entropy loss yielded by the models after each ablation step trained for predicting the **veracity** of a rumour (task B). The findings highlight the ability of language indicators to model the truthfulness or deception of a claim. The features with most predictive power were language model and word vector clusters, especially numbers and URLs in the tweet text, and words obtained from the user profile, such as: *delivering you*, *insightful analysis*, *breaking news*, *contact*, *facebook*, *latest*, *tweets*, *we*, *views*, EMAIL, *online news*, *around the world*, *channel*, *bbc*, *bbcsport*, *cnn*, etc.

The current findings demonstrate surprisingly low accuracies, (F1 = 0.21645 for Task A, F1 = 0.3326 for Task B) when evaluated on the testing dataset, although in line with the results of previous tasks on the same dataset Derczynski et al. (2017). Collecting larger quantities of Twitter data and optimization techniques could improve the consistency of the results obtained on the training and validation sets. Importance of close inspection of data, and comparative analysis with other research on the same task could better support the

Ablation	Comment F1	Query F1	Support/Deny F1
Baseline	0.795715	0.804499	0.707801
Without language style	0.795715	0.794216	0.707801
Without n-grams and embeddings	0.785013	0.681081	0.618934
Without sentiment	0.784082	0.791145	0.666029
Without network structure	0.793228	0.789274	0.693583
Without Twitter metadata	0.628665	0.804499	0.691252

Table 2: Evaluation results from the ablation experiment for Task A performed on a randomly-created validation set

Ablation	Veracity cross-entropy loss
Baseline	0.591780
Without language style	0.605128
Without n-grams and embeddings	0.667380
Without sentiment	0.600889
Without network structure	0.591780
Without Twitter metadata	0.596146

Table 3: Evaluation results from the ablation experiment for Task B performed on a randomly-created validation set

interpretation of the results. We defer such discussion and directions for future research, until a detailed analysis of misclassified cases is done and proper treatment and improvements of such scenarios could be speculated.

5 Conclusion

The present research explores a hybrid approach to the problem of analyzing the veracity of rumours and the support for rumours on social media platforms. Following the results of previous research in this field, different combinations of features were examined, while also leveraging a variety of tangible indicators not accounted for in related research. The recurrent challenges in fully elucidating the language ambiguities of complex phenomena such as rumour spreading led us in a direction of including distal contextual indicators. In particular, the models were augmented with language indicators extrapolated from the content of user profiles, Twitter metadata, and thread structural characteristics of rumourous tweets. Their relevance and predictive effects have been confirmed with the results, providing exciting directions for further research on the problem.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Prashant Bordia and Nicholas Difonzo. 2004. Problem solving in social interactions on the internet: Rumor as social cognition. *Social Psychology Quarterly*, 67(1):33–49. https://doi.org/10.1177/019027250406700105.

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, Vancouver, Canada. Association for Computational Linguistics. http://www.aclweb.org/anthology/S17-2006.

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. https://arxiv.org/pdf/1809.06683.pdf.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA.

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*.

Michal Lukasik, Trevor Cohn, and Kalina Bontcheva. 2015. Classifying tweet level judgements of rumours in social media. *Proceedings of the 2015 Conference on Empirical Methods in Natural Lan-*

guage Processing. http://dx.doi.org/10.18653/v1/D15-1311.

Tsvetomila Mihaylova, Preslav Nakov, Lluis Marquez, Alberto Barron-Cedeno, Mitra Mohtarami, Georgi Karadzhov, and James Glass. 2018. Fact checking in community forums. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. https://arxiv.org/pdf/1803.03178.pdf.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Warren A. Peterson and Noel P. Gist. 1951. Rumor and public opinion. *American Journal of Sociology*, 57(2):159–167. https://doi.org/10.1086/220916.

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*, EMNLP '11, pages 1589–1599, Stroudsburg, PA, USA. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2145432.2145602.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016a. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3):1–29. https://doi.org/10.1371/journal.pone.0150989.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016b. Pheme rumour scheme dataset: journalism use case. https://figshare.com/articles/PHEME_rumour_scheme_dataset_journalism_use_case/2068650/2.