Sequential Approach to Rumour Stance Classification

Elena Kochkina[12], **Maria Liakata**[12], **Isabelle Augenstein**[3]
[1] University of Warwick, Coventry, United Kingdom
[2] Alan Turing Institute, London, United Kingdom
[3] University College London, London, United Kingdom
{E.Kochkina, M.Liakata}@warwick.ac.uk, I.Augenstein@ucl.ac.uk

Abstract

Rumour stance classification is a task that involves identifying the attitude of Twitter users towards the truthfulness of the rumour they are discussing. Stance classification is considered to be an important step towards rumour verification, therefore performing well in this task is expected to be useful in debunking false rumours. In this work we classify a set of Twitter posts discussing rumours into either supporting, denying, questioning or commenting on the underlying rumours. We propose an LSTM-based sequential model that, through modelling the conversational structure of tweets, obtains state-of-the-art accuracy on the SemEval-2017 RumourEval dataset.

1 Introduction

In stance classification one is concerned with determining the attitude of the author of a text towards a target (Mohammad et al., 2016). Targets can range from abstract ideas, to concrete entities and events. Stance classification is an active research area that has been studied in different domains (Ranade et al., 2013; Chuang and Hsieh, 2015). Here we focus on stance classification of tweets towards the truthfulness of rumours circulating in Twitter conversations in the context of breaking news. Each conversation is defined by a tweet that initiates the conversation and a set of nested replies to it that form a conversation thread. The goal is to classify each of the tweets in the conversation thread as either *supporting*, *denying*, *querying* or *commenting* (*SDQC*) on the rumour initiated by the source tweet. Being able to detect stance automatically is very useful in the context of events provoking public resonance and associated rumours, as a first step towards verification of early reports (Zhao et al., 2015). For instance, it has been shown that rumours that are later proven to be false tend to spark significantly larger numbers of denying tweets than rumours that are later confirmed to be true (Mendoza et al., 2010; Procter et al., 2013; Derczynski et al., 2014; Zubiaga et al., 2016).

Here we focus on exploiting the conversational structure of social media threads for stance classification and introduce a novel LSTM-based approach to harness conversations.

2 Dataset

We use the dataset of Twitter conversation threads associated with rumours around ten different events in breaking news, including the Paris shootings in Charlie Hebdo, the Ferguson unrest, the crash of a Germanwings plane[1]. These events include 325 conversation threads consisting of 5568 underlying tweets annotated for stance at the tweet level as either *supporting*,*denying*, *querying* or *commenting* on a rumour.

3 Method

3.1 Features

We use the following features:

- **Word vectors:** we use a word2vec (Mikolov et al., 2013) model pre-trained on the Google News dataset (300d) using the gensim package (Řehůřek and Sojka, 2010).
- **Tweet lexicon:** (1) count of negation words[2] and (2) count of swear words.[3]
- **Punctuation:** (1) presence of a period, (2) presence of an exclamation mark, (3) pres-

[1] http://alt.qcri.org/semeval2017/task8/index.php?id=data-and-tools

[2] A presence of any of the following words would be considered as a presence of negation: not, no, nobody, nothing, none, never, neither, nor, nowhere, hardly, scarcely, barely, don't, isn't, wasn't, shouldn't, wouldn't, couldn't, doesn't

[3] A list of 458 bad words was taken from http://urbanoalvarez.es/blog/2008/04/04/bad-words-list/

	Accuracy	Macro F	S	D	Q	C
Development	0.782	0.561	0.621	0.000	0.762	0.860
Testing	**0.784**	0.434	0.403	0.000	0.462	0.873

Table 1: Results on the development and testing sets. Accuracy and F1 scores: macro-averaged and per class (S: *supporting*, D: *denying*, Q: *querying*, C: *commenting*).

ence of a question mark, (4) ratio of capital letters.

- **Attachments:** (1) presence of a URL and (2) presence of images.
- **Relation to other tweets** (1) Word2Vec cosine similarity wrt source tweet, (2) Word2Vec cosine similarity wrt preceding tweet, and (3) Word2Vec cosine similarity wrt thread
- **Content length:** (1) word count and (2) character count.
- **Tweet role:** whether the tweet is a source tweet of a conversation.

Tweet representations are obtained by averaging word vectors in a tweet and then concatenating with the additional features into a single vector, at the preprocessing step. We found this set of features to be the best compared to using word2vec features on their own or any of the combinations of subsets of these features.

3.2 Branch - LSTM Model

To tackle the task of rumour stance classification, we propose *branch-LSTM*, a neural network architecture that uses layers of LSTM units (Hochreiter and Schmidhuber, 1997) to process the whole branch of tweets, thus incorporating structural information about the conversation (see the illustration of the *branch-LSTM* on Figure 1). The input at each time step i of the LSTM layer is the representation of the tweet as a vector. We record the output of each time step so as to attach a label to each tweet in a branch[4]. This output is fed through several dense ReLU layers, a 50% dropout layer, and then through a softmax layer to obtain class probabilities.

The model uses tweet representation as the mean average of word vectors concatenated with extra features described above. Due to the short length of tweets, using more complex models for learning tweet representations, such as an LSTM that takes each word as input at each time step

[4]For implementation of all models we used Python libraries Theano (Bastien et al., 2012) and Lasagne (Dieleman et al., 2015).

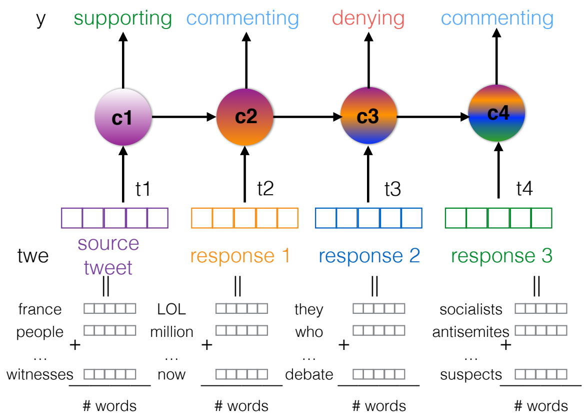

Figure 1: Illustration of the input/output structure of the branch-nestedLSTM model.

and returns the representation at the final time step, does not lead to a noticeable difference in the performance based on cross-validation experiments on the training and development sets, while taking significantly longer to train.

4 Results

The performance of our model on the testing and development set is shown in Table 1. Together with the accuracy we show macro-averaged F-score and per-class macro-averaged F-scores since these metrics account for class imbalance. The difference in accuracy between the test and development sets is minimal, however we see significant difference in Macro-F score due to different class balance in these two sets. Macro-F score could be improved if we used it as a metric for optimising hyper-parameters. The *branch-LSTM* model predicts *commenting*, the majority class well, however it is unable to pick out any *denying*, the most-challenging under-represented class.

5 Conclusions

Our method decomposes the tree structure of conversations into linear sequences, achieves an accuracy of 78.4% on the test set and constitutes the state-of-the-art for rumour stance classification. In future work we plan to explore different methods for modelling tree-structured conversations.

References

Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian Goodfellow, Arnaud Bergeron, Nicolas Bouchard, David Warde-Farley, and Yoshua Bengio. 2012. Theano: new features and speed improvements. *arXiv preprint arXiv:1211.5590* .

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

Leon Derczynski, Kalina Bontcheva, Michal Lukasik, Thierry Declerck, Arno Scharl, Georgi Georgiev, Petya Osenova, Toms Pariente Lobo, Anna Kolliakou, Robert Stewart, et al. 2014. Pheme: computing veracity: the fourth challenge of big social data. In *Proceedings of ESWC EU Project Networking*.

Sander Dieleman, Jan Schlüter, Colin Raffel, Eben Olson, Sren Kaae Sønderby, Daniel Nouri, Daniel Maturana, Martin Thoma, Eric Battenberg, Jack Kelly, Jeffrey De Fauw, Michael Heilman, Diogo Moitinho de Almeida, Brian McFee, Hendrik Weideman, Gbor Takács, Peter de Rivaz, Jon Crall, Gregory Sanders, Kashif Rasul, Cong Liu, Geoffrey French, and Jonas Degrave. 2015. Lasagne: First release. https://doi.org/10.5281/zenodo.27878.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. Twitter under crisis: Can we trust what we RT? In *1st Workshop on Social Media Analytics*. SOMA'10, pages 71–79.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Saif M Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the International Workshop on Semantic Evaluation, SemEval*. volume 16.

Rob Procter, Farida Vis, and Alex Voss. 2013. Reading the riots on twitter: methodological innovation for the analysis of big data. *International journal of social research methodology* 16(3):197–214.

Sarvesh Ranade, Rajeev Sangal, and Radhika Mamidi. 2013. Stance classification in online debates by recognizing users' intentions. In *Proceedings of the SIGDIAL 2013 Conference*. pages 61–69.

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*. ELRA, Valletta, Malta, pages 45–50. `http://is.muni.cz/publication/884893/en`.

Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, pages 1395–1405.

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