

Turing at SemEval-2017 Task 8: Sequential Approach to Rumour Stance Classification with Branch-LSTM

Elena Kochkina^{1,2}, Maria Liakata^{1,2}, Isabelle Augenstein³

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

This paper describes team Turing’s submission to SemEval 2017 RumourEval: Determining rumour veracity and support for rumours (SemEval 2017 Task 8, Subtask A). Subtask A addresses the challenge of rumour stance classification, which 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 a LSTM-based sequential model that, through modelling the conversational structure of tweets, which achieves an accuracy of 0.784 on the RumourEval test set outperforming all other systems in Subtask A.

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., 2016b).

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 Related Work

Single Tweet Stance Classification Stance classification for rumours was pioneered by Qazvinian et al. (2011) as a binary classification task (support/denial). Zeng et al. (2016) perform stance classification for rumours emerging during crises. Both works use tweets related to the same rumour during training and testing.

A model based on bidirectional LSTM encoding of tweets conditioned on targets has been shown to achieve state-of-the-art on the SemEval-2016 task 6 dataset (Augenstein et al., 2016). However the RumourEval task is different as it addresses conversation threads.

Sequential Stance Classification Lukasik et al. (2016) and Zubiaga et al. (2016a) consider the sequential nature of tweet threads in their works. Lukasik et al. (2016) employ Hawkes processes to classify temporal sequences of tweets. They show the importance of using both the textual content and temporal information about the tweets, disregarding the discourse structure. Zubiaga et

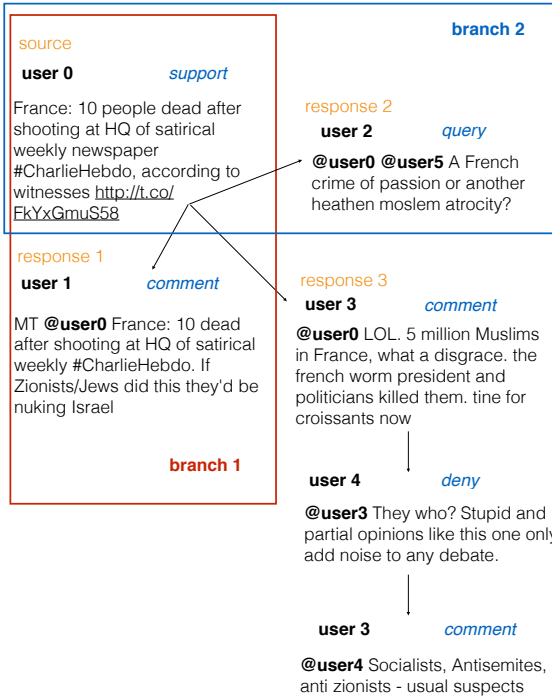


Figure 1: Example of a conversation thread from the dataset with three branches, two of which are highlighted. The conversation has a tree structure, which can be split into individual branches by taking each leaf node with all its direct parents.

al. (2016a) model the conversational structure of source tweets and subsequent replies: as a linear chain and as a tree. They use linear- and tree- versions of a CRF classifier, outperforming the approach by Lukasik et al. (2016).

3 Dataset

The dataset provided for this task contains 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. These events include 325 conversation threads consisting of 5568 underlying tweets annotated for stance at the tweet level (breakdown between training, testing and development sets is shown in Table 1) (Derczynski et al., 2017).

	# threads	# branches	# tweets
Development	25	215	281
Testing	28	772	1049
Training	272	3030	4238
Total	325	4017	5568

Table 1: Number of threads, branches and tweets in the training, development and testing sets.

	S	D	Q	C
Development	69	11	28	173
Testing	94	71	106	778
Training	841	333	330	2734
Total	1004	415	464	3685

Table 2: Per-class distribution of tweets in the training, development and testing sets.

Each thread includes a source tweet that initiates a conversation and nested tweets responding to either the source tweet or earlier replies. The thread can be split into linear branches of tweets, where a branch is defined as a chain of tweets that starts with a leaf tweet including its direct parent tweets, all the way up to the source tweet. Figure 1 shows an example of a conversation along with its annotations represented as a tree structure with highlighted branches. The depth of a tweet is the number of its parents starting from the root node. Branches 1 and 2 in Figure 1 have depth one whereas branch 3 has depth three. There is a clear class imbalance in favour of *commenting* tweets (66%) and *supporting* tweets (18%), whereas the *denying* (8%) and *querying* classes (8%) are under-represented (see Table 2). While this imbalance poses a challenge, it is also indicative of the realistic scenario where only a few users question the veracity of a statement.

4 System Description

4.1 Features

Prior to generating features for the tweets, we perform a pre-processing step where we remove non-alphabetic characters, convert all words to lower case and tokenise texts.¹ Once tweet texts are pre-processed, we extract the following features:

- **Word vectors:** we use a word2vec (Mikolov et al., 2013) model pre-trained on the Google News dataset (300d)² using the gensim package (Řehůřek and Sojka, 2010).
- **Tweet lexicon:** (1) count of negation words³ and (2) count of swear words.⁴

¹For implementation of all pre-processing routines we use Python 2.7 with the NLTK package.

²We have also tried using Glove word embeddings trained on Twitter dataset, but it lead to a decrease in performance on both development and testing sets comparing to the Google News word vectors

³A 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

⁴A list of 458 bad words was taken from <http://urbanoalvarez.es/blog/2008/04/04/bad-words-list/>

[As querying] @username Weren't you the one who abused her?

[As supporting] "Go online & put down 'Hillary Clinton illness,'" Rudy says. Yes – but look up the truth – not health smears https://t.co/EprqiZhAxM

[As supporting] @username I demand you retract the lie that people in #Ferguson were shouting "kill the police", local reporting has refuted your ugly racism

[As commenting] @FoxNews six years ago... real good evidence. Not!

Figure 2: Examples of misclassified denying tweets.

- **Punctuation:** (1) presence of a period, (2) presence of an exclamation mark, (3) presence 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. This set of features have shown to be the best comparing to using word2vec features on their own or any of the reduced combinations of these features.

4.2 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 of the conversation (see the illustration of the *branch-LSTM* on the Figure 3). The input at each time step i of the LSTM layer is the representation of the tweet as a vector. We record the

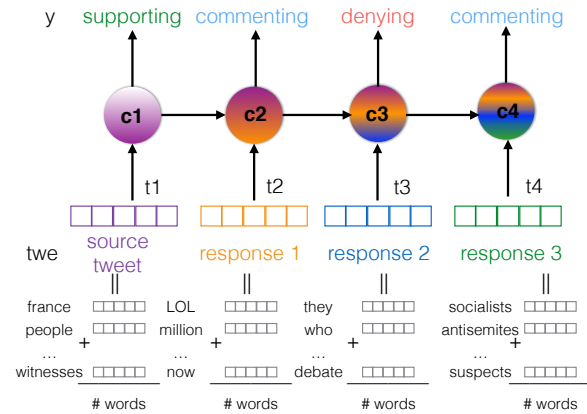


Figure 3: Illustration of the input/output structure of the branch-nested LSTM model.

output of each time step so as to attach a label to each tweet in a branch⁵. This output is fed through several dense ReLU layers, a 50% dropout layer, and then through a softmax layer to obtain class probabilities. We use zero-padding and masks to account for the varying lengths of tweet branches. The model is trained using the categorical cross entropy loss function. Since there is overlap between branches originating from the same source tweet, we exclude the repeating tweets from the loss function using a mask at the training stage. 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 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. We experimented with replacing the unidirectional LSTMs with bidirectional LSTMs but could observe no improvements in accuracy (using cross-validation results on the training and development set).

5 Experimental Setup

The dataset is split into training, development and test sets by the task organisers. We determined the optimal set of hyperparameters via testing the performance of our model on the development set for different parameter combinations. We used the

⁵For implementation of all models we used Python libraries Theano (Bastien et al., 2012) and Lasagne (Dieleman et al., 2015).

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

Depth	# tweets	# S	# D	# Q	# C	Accuracy	MacroF	S	D	Q	C
0	28	26	2	0	0	0.929	0.481	0.963	0.000	0.000	0.000
1	704	61	60	81	502	0.739	0.348	0.000	0.000	0.550	0.842
2	128	3	6	7	112	0.875	0.233	0.000	0.000	0.000	0.933
3	60	2	1	5	52	0.867	0.232	0.000	0.000	0.000	0.929
4	41	0	0	3	38	0.927	0.481	0.000	0.000	0.000	0.962
5	27	1	0	1	25	0.926	0.321	0.000	0.000	0.000	0.961
6+	61	1	2	9	49	0.803	0.223	0.000	0.000	0.000	0.891

Table 4: Number of tweets per depth and performance at each of the depths.

Tree of Parzen Estimators (TPE) algorithm ⁶ to search the parameter space, which is defined as follows: the number of dense ReLU layers varies from one to four; the number of LSTM layers is one or two; the mini-batch size is either 32 or 64; the number of units in the ReLU layer is one of {100, 200, 300, 400, 500}, and in the LSTM layer one of {100, 200, 300}; the strength of the L2 regularisation is one of {0.0, 1e-4, 3e-4, 1e-3} and the number of epochs is selected from {30, 50, 70, 100}. We performed 100 trials of different parameter combinations optimising for accuracy on the development set in order to choose the best combination. We fixed hyperparameters to train the model on combined training and development sets and evaluated on the held out test set.

6 Results

The performance of our model on the testing and development set is shown in Table 3. Together with the accuracy we show macro-averaged F-score and per-class macro-averaged F-scores as these metrics account for the class imbalance. The difference in accuracy between testing and development set is minimal, however we see significant difference in Macro-F score due to different class balance in these 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. Most *deny-*

⁶We used the implementation of the TPE algorithm in the hyperopt package (Bergstra et al., 2013)

Label \ Prediction	Prediction			
	C	D	Q	S
Commenting	760	0	12	6
Denying	68	0	1	2
Querying	69	0	36	1
Supporting	67	0	1	26

Table 5: Confusion matrix for testing set predictions

ing instances get misclassified as *commenting* (see Table 5), with only one tweet misclassified as *querying* and two as *supporting* (Figure 2). An increased amount of labelled data would be helpful to improve performance of this model. As we were considering conversation branches, it is interesting to analyse the performance distribution across different tweet depths (see Table 4). Maximum depth/branch length in the testing set is 13 with most tweets concentrated at depths from 0 to 3. Source tweets (depth zero) are usually *supporting* and the model predicts these very well, but performance of *supporting* tweets at other depths decreases. The model does not show a noticeable difference in performance on tweets of varying lengths.

7 Conclusions

This paper describes the Turing system entered in the SemEval-2017 Task 8 Subtask A. Our method decomposes the tree structure of conversations into linear sequences and achieves accuracy 0.784 on the testing set and sets the state-of-the-art for rumour stance classification. In future work we plan to explore different methods for modelling tree-structured conversations.

Acknowledgments

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. Cloud computing resource were kindly provided through a Microsoft Azure for Research Award. Work by Elena Kochkina was partially supported by the Leverhulme Trust through the Bridges Programme.

References

- Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance Detection with Bidirectional Conditional Encoding. In *Proceedings of EMNLP*.
- 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*.
- James Bergstra, Daniel Yamins, and David D Cox. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. *ICML (1)* 28:115–123.
- 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, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. [Semeval-2017 task 8: Rumoureal: Determining rumour veracity and support for rumours](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 69–76. <http://www.aclweb.org/anthology/S17-2006>.
- Leon Derczynski, Kalina Bontcheva, Michal Lukasik, Thierry Declerck, Arno Scharl, Georgi Georgiev, Petya Osenova, Toms Pariente Lobo, Anna Kollidakou, 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 Maturation, 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 Degraeve. 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.
- Michal Lukasik, P. K. Srijith, Duy Vu, Kalina Bontcheva, Arkaitz Zubiaga, and Trevor Cohn. 2016. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. In *Proceedings of the 54th Meeting of the Association for Computational Linguistics*. Association for Computer Linguistics, pages 393–398.
- 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.
- 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.
- 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>.
- Li Zeng, Kate Starbird, and Emma S Spiro. 2016. #unconfirmed: Classifying rumor stance in crisis-related social media messages. In *Tenth International AAAI Conference on Web and Social Media*.
- 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, 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. In *Proceedings of COLING, the International Conference on Computational Linguistics*.

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