Information Processing and Management 54 (2018) 273–290

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Discourse-aware rumour stance classification in social media using sequential classifiers

Arkaitz Zubiaga*[a], Elena Kochkina[a,b], Maria Liakata[a,b], Rob Procter[a,b], Michal Lukasik[c], Kalina Bontcheva[c], Trevor Cohn[d], Isabelle Augenstein[e]

[a] University of Warwick, Department of Computer Science, Gibbet Hill Road, Coventry CV4 7AL, United Kingdom
[b] Alan Turing Institute, 96 Euston Rd, Kings Cross, London NW1 2DB, United Kingdom
[c] University of Sheffield, Department of Computer Science, Regent Court, 211 Portobello, Sheffield S1 4DP, United Kingdom
[d] University of Melbourne, Computing and Information Systems, Melbourne VIC 3010, Australia
[e] University of Copenhagen, Department of Computer Science, Sigurdsgade 41, 2200 Copenhagen N, Denmark

ARTICLE INFO

Keywords:
Stance classification
Social media
Breaking news
Veracity classification

ABSTRACT

Rumour stance classification, defined as classifying the stance of specific social media posts into one of supporting, denying, querying or commenting on an earlier post, is becoming of increasing interest to researchers. While most previous work has focused on using individual tweets as classifier inputs, here we report on the performance of sequential classifiers that exploit the discourse features inherent in social media interactions or 'conversational threads'. Testing the effectiveness of four sequential classifiers – Hawkes Processes, Linear-Chain Conditional Random Fields (Linear CRF), Tree-Structured Conditional Random Fields (Tree CRF) and Long Short Term Memory networks (LSTM) – on eight datasets associated with breaking news stories, and looking at different types of local and contextual features, our work sheds new light on the development of accurate stance classifiers. We show that sequential classifiers that exploit the use of discourse properties in social media conversations while using only local features, outperform non-sequential classifiers. Furthermore, we show that LSTM using a reduced set of features can outperform the other sequential classifiers; this performance is consistent across datasets and across types of stances. To conclude, our work also analyses the different features under study, identifying those that best help characterise and distinguish between stances, such as supporting tweets being more likely to be accompanied by evidence than denying tweets. We also set forth a number of directions for future research.

1. Introduction

Social media platforms have established themselves as important sources for learning about the latest developments in breaking news. People increasingly use social media for news consumption (Hermida, Fletcher, Korell, & Logan, 2012; Mitchell, Gottfried, & Matsa, 2015; Zubiaga, Spina, Martinez, & Fresno, 2015), while media professionals, such as journalists, increasingly turn to social media for news gathering (Zubiaga, Ji, & Knight, 2013) and for gathering potentially exclusive updates from eyewitnesses (Diakopoulos, De Choudhury, & Naaman, 2012; Tolmie, Procter, Rouncefield, Liakata, & Zubiaga, 2017a). Social media platforms such as Twitter are a fertile and prolific source of breaking news, occasionally even outpacing traditional news media organisations (Kwak, Lee, Park, & Moon, 2010). This has led to the development of multiple data mining applications for mining and discovering

* Corresponding author.
E-mail addresses: a.zubiaga@warwick.ac.uk, arkaitz@zubiaga.org (A. Zubiaga).

https://doi.org/10.1016/j.ipm.2017.11.009
Received 31 May 2017; Received in revised form 26 November 2017; Accepted 28 November 2017
0306-4573/

events and news from social media (Dong, Mavroeidis, Calabrese, & Frossard, 2015; Stilo & Velardi, 2016). However, the use of social media also comes with the caveat that some of the reports are necessarily rumours at the time of posting, as they have yet to be corroborated and verified (Mendoza, Poblete, & Castillo, 2010; Procter, Crump, Karstedt, Voss, & Cantijoch, 2013; Procter, Vis, & Voss, 2013). The presence of rumours in social media has hence provoked a growing interest among researchers for devising ways to determine veracity in order to avoid the diffusion of misinformation (Derczynski et al., 2015).

Resolving the veracity of social rumours requires the development of a rumour classification system and we described in Zubiaga, Aker, Bontcheva, Liakata, and Procter (2017), a candidate architecture for such a system consisting of the following four components: (1) detection, where emerging rumours are identified, (2) tracking, where those rumours are monitored to collect new related tweets, (3) stance classification, where the views expressed by different tweet authors are classified, and (4) veracity classification, where knowledge garnered from the stance classifier is put together to determine the likely veracity of a rumour.

In this work we focus on the development of the third component, a stance classification system, which is crucial to subsequently determining the veracity of the underlying rumour. The stance classification task consists in determining how individual posts in social media observably orientate to the postings of others (Qazvinian, Rosengren, Radev, & Mei, 2011; Walker, Anand, Abbott, & Grant, 2012a). For instance, a post replying with "no, that's definitely false" is *denying* the preceding claim, whereas "yes, you're right" is *supporting* it. It has been argued that aggregation of the distinct stances evident in the multiple tweets discussing a rumour could help in determining its likely veracity, providing, for example, the means to flag highly disputed rumours as being potentially false (Mendoza et al., 2010). This approach has been justified by recent research that has suggested that the aggregation of the different stances expressed by users can be used for determining the veracity of a rumour (Derczynski et al., 2015; Liu, Nourbakhsh, Li, Fang, & Shah, 2015).

In this work, we examine in depth the use of so-called sequential approaches to the rumour stance classification task. Sequential classifiers are able to utilise the discursive nature of social media (Tolmie et al., 2017a), learning from how 'conversational threads' evolve for a more accurate classification of the stance of each tweet. The use of sequential classifiers to model the conversational properties inherent in social media threads is still in its infancy. For example, in preliminary work we showed that a sequential classifier modelling the temporal sequence of tweets outperforms standard classifiers (Lukasik et al., 2016b; Zubiaga, Kochkina, Liakata, Procter, & Lukasik, 2016). Here we extend this preliminary experimentation in four different directions that enable exploring further the stance classification task using sequential classifiers: (1) we perform a comparison of a range of sequential classifiers, including a Hawkes Process classifier, a Linear CRF, a Tree CRF and an LSTM; (2) departing from the use of only local features in our previous work, we also test the utility of contextual features to model the conversational structure of Twitter threads; (3) we perform a more exhaustive analysis of the results looking into the impact of different datasets and the depth of the replies in the conversations on the classifiers' performance, as well as performing an error analysis; and (4) we perform an analysis of features that gives insight into what characterises the different kinds of stances observed around rumours in social media. To the best of our knowledge, dialogical structures in Twitter have not been studied in detail before for classifying each of the underlying tweets and our work is the first to evaluate it exhaustively for stance classification. Twitter conversational threads are identifiable by the relational features that emerge as users respond to each others' postings, leading to tree-structured interactions. The motivation behind the use of these dialogical structures for determining stance is that users' opinions are expressed and evolve in a discursive manner, and that they are shaped by the interactions with other users.

The work presented here advances research in rumour stance classification by performing an exhaustive analysis of different approaches to this task. In particular, we make the following contributions:

- We perform an analysis of whether and the extent to which use of the sequential structure of conversational threads can improve stance classification in comparison to a classifier that determines a tweet's stance from the tweet in isolation. To do so, we evaluate the effectiveness of a range of sequential classifiers: (1) a state-of-the-art classifier that uses Hawkes Processes to model the temporal sequence of tweets (Lukasik et al., 2016b); (2) two different variants of Conditional Random Fields (CRF), i.e., a linear-chain CRF and a Tree CRF; and (3) a classifier based on Long Short Term Memory (LSTM) networks. We compare the performance of these sequential classifiers with non-sequential baselines, including the non-sequential equivalent of CRF, a Maximum Entropy classifier.
- We perform a detailed analysis of the results broken down by dataset and by depth of tweet in the thread, as well as an error analysis to further understand the performance of the different classifiers. We complete our analysis of results by delving into the features, and exploring whether and the extent to which they help characterise the different types of stances.

Our results show that sequential approaches do perform substantially better in terms of macro-averaged $F1$ score, proving that exploiting the dialogical structure improves classification performance. Specifically, the LSTM achieves the best performance in terms of macro-averaged $F1$ scores, with a performance that is largely consistent across different datasets and different types of stances. Our experiments show that LSTM performs especially well when only local features are used, as compared to the rest of the classifiers, which need to exploit contextual features to achieve comparable – yet still inferior – performance scores. Our findings reinforce the importance of leveraging conversational context in stance classification. Our research also sheds light on open research questions that we suggest should be addressed in future work. Our work here complements other components of a rumour classification system that we implemented in the PHEME project, including a rumour detection component (Zubiaga, Liakata, & Procter, 2016; Zubiaga, Liakata, & Procter, 2017), as well as a study into the diffusion of and reactions to rumour (Zubiaga, Liakata, Procter, Wong Sak Hoi, & Tolmie, 2016).