Can Rumour Stance Alone Predict Veracity?

Sebastian Dungs[♪], **Ahmet Aker**[♪♩], **Norbert Fuhr**[♪] **and Kalina Bontcheva**[♩]
[♪] University of Duisburg-Essen, Lotharstraße 65, 47057 Duisburg, Germany
[♩] University of Sheffield, 211 Portobello, Sheffield S1 4DP, UK
firstname.lastname@uni-due.de
k.bontcheva@dcs.shef.ac.uk

Abstract

Prior manual studies of rumours suggested that crowd stance can give insights into the actual rumour veracity. Even though numerous studies of automatic veracity classification of social media rumours have been carried out, none explored the effectiveness of leveraging crowd stance to determine veracity. We use stance as an additional feature to those commonly used in earlier studies. We also model the veracity of a rumour using variants of Hidden Markov Models (HMM) and the collective stance information. This paper demonstrates that HMMs that use stance and tweets' times as the only features for modelling true and false rumours achieve F_1 scores in the range of 80%, outperforming those approaches where stance is used jointly with content and user based features.

1 Introduction

Social media are rife with rumours, which are fast-spreading, unverified pieces of information (Zubiaga et al., 2018). With fake news and misinformation now widely recognised as a major problem for journalists, media, online platforms, and citizens, automatic rumour detection and analysis has become a hot research topic too. Rumour analysis research has focused on Twitter in particular, as it has established itself as the go-to social platform for real-time news (Hu et al., 2012). Twitter's unmoderated nature is also the perfect ground for spreading rumours (Qazvinian et al., 2011). A key focus of prior work on rumour analysis has been *rumour stance classification*, where the stance of each tweet on a given rumour is classified as supporting, denying, questioning or commenting on the rumour (Procter et al., 2013).

Our work builds on the hypothesis that, as rumours evolve over time, so does the stance expressed by the public towards those rumours. In the early stages of a rumour, its actual veracity tends to be unknown. However, as new evidence emerges over time, Twitter users take more pronounced and continuously evolving stance towards the information asserted in the rumour. For instance, in the early stages of a rumour supporting tweets might prevail, simply due to the lack of information to the contrary. However, when authoritative sources or reliable evidence emerge either for or against the rumour, a similar trend tends to be observed in the collective rumour stances. This was first noted in a manual analysis by Mendoza et al. (2010), who found that true rumours tended to be affirmed more than 90% of the time, whereas false rumours were challenged (questioned or denied) 50% of the time. This encourages the use of crowd (or collective) stance as a feature in an automatic rumour veracity classifier.

A rumour consists of a source tweet and several other tweets that responded to the source one—the source tweet contains the rumour. Each of those responding tweets is associated with a particular stance (supporting, denying, questioning or commenting). In this work we make use of the stances of those responding tweets to judge the veracity of the rumour. We refer to the stances of those tweets as crowd or collective stance.

Hidden Markov Models (HMM) are well known for their application in temporal pattern recognition. By regarding the individual stances over a rumour's lifetime as an ordered sequence of observations, we

Proceedings of the 27th International Conference on Computational Linguistics, pages 3360–3370
Santa Fe, New Mexico, USA, August 20-26, 2018.

can then model the actual veracity of a rumour by the hidden part of the HMM. The general assumption is that true and false rumours would have different patterns in the stance distribution over time. Therefore, after training models for *true* and *false* rumours, we can build a binary veracity classifier by comparing sequence occurrence probabilities for the two cases.

This paper investigates whether and to what extent rumour veracity classification can be predicted on the basis of crowd stance. While there is a body of work on automatic veracity classification such as (Castillo et al., 2011; Kwon et al., 2013; Vosoughi, 2015; Wu et al., 2015; Ma et al., 2015; Lukasik et al., 2016), only few have applied stance information as a feature (Liu et al., 2015; Enayet and El-Beltagy, 2017), and none of these studies investigated the collective power of the crowd as a source of stance information, which can then be used as a feature in rumour veracity classification. Here we argue that the content of the posts is not necessarily helpful towards determining the veracity of a story, as the content can be either misleading or inaccurate. We believe it is the aggregation of stances which can provide useful information to determine the veracity. Despite the fact that some of the users will be inevitably mistaken, and sharing the wrong stance, we argue that the aggregation of stances will correct itself towards being a useful feature for veracity classification. The novel contribution of this paper is in demonstrating that collective stance together with the tweets' time as features can indeed boost performance results. We demonstrate that HMM using only stance and time features achieve an F_1 score of around 80%, rendering significantly better results than traditional rich feature engineered approaches that entails stance as additional feature. We also show that our HMM are indeed applicable in case of early rumour detection.

2 Related Work

The veracity classification task aims to determine whether a given rumour can be confirmed as true, debunked as false, or in some cases its truth value is yet to be resolved (i.e. unknown). Related work has tackled the problem in a supervised fashion by applying state-of-the-art machine learning algorithms on features extracted from rumour datasets. One of the key differences between the approaches is the set of features proposed to tackle the problem. The pioneering paper of Castillo et al. (2011) proposed message, user, topic and propagation-based features. In most subsequent studies these features have been used as baselines. Following these feature sets, Kwon et al. (2013) and Kwon et al. (2017) proposed a new set of feature categories: temporal, structural and linguistic and showed their importance in fighting with rumours. Other than Twitter, the Chinese microblogging platform Sina Weibo has been also analysed for rumours. For this Yang et al. (2012) proposed client and location-based features and showed that these help to increase prediction accuracy.

Liu et al. (2015) used the approaches reported by Yang et al. (2012) and Castillo et al. (2011) as baseline systems and compared them against a new approach based on verification features, which include stance of individual tweets, rather than collective stance. Ma et al. (2015) adapted features from earlier studies and proposed to model them over time. Wu et al. (2015) extracted features from message propagation trees. Three categories of features were considered: message, user and report-based. The idea of message propagation was also investigated by Wang and Terano (2015). Vosoughi (2015) tackled veracity classification using three categories of features (linguistic, user oriented and temporal propagation related) and speech recognition inspired machine learning approaches, such as Dynamic Time Wrapping and Hidden Markov Models.

Chen et al. (2016) treated rumour veracity classification as an anomaly detection problem where false rumours are regarded as anomalies. Several features related to the content, crowd opinion and post propagation were used. Chang et al. (2016) put the emphasis on the characteristics of users who post the rumours to determine the veracity. Unlike previous studies, Tong et al. (2017) aimed at blocking rumours rather than detecting them or marking tweets as true or false. Motivated by the fact that later corrections are not as effective, the authors argued that the first post seen by a user is influential for their future opinion and thus it is important to show users rumours only once they are confirmed to be true. Based on this they proposed a reverse-tuple based randomised algorithm to block rumours. The algorithm aimed at producing positive seeds to be shown to users first.

Rumour veracity classification has also been studied in the RumourEval shared task at SemEval 2017 (Derczynski et al., 2017a). Subtask B consisted in determining if each of the rumours in the dataset were true, false or remained unverified. It considered two different settings, one *closed* where participants could not make use of external knowledge bases and another *open* where use of external resources was allowed. Participants viewed the task either as a three-way (Enayet and El-Beltagy, 2017; Wang et al., 2017; Singh et al., 2017) or two-way (Chen et al., 2017; Srivastava et al., 2017), single tweet classification task. The winning system (Enayet and El-Beltagy, 2017) added features more specific to the distribution of stance labels in the tweets replying to the source tweet (percentage of reply tweets classified as either support, deny or query). The survey paper of Zubiaga et al. (2018) provides an extensive summary of current work on rumour verification but also related task such as detection of rumours as well as stance classification of messages involved in rumours.

In this work we propose to use stance only to tackle the rumour verification task. As highlighted earlier, we aim to capture collective stance information for veracity classification. First we investigate the impact of using sequences of individual stances only. In our second approach, we integrate stance and time information into a unified model. Both approaches are based on modelling the state-changes of stances using HMM. Thus the most related studies to ours are Ma et al. (2015) because they model features over time, Vosoughi (2015) due to the use of HMM as well as Chang et al. (2016), Liu et al. (2015) and Enayet and El-Beltagy (2017) because they add the reactions (stances) of the users as additional feature. However, we differ from those because we rely only on the power of collective stance, model this using HMM and apply this model to predict the veracity of rumours.

3 Dataset

The rumour dataset used in this paper (Zubiaga et al., 2016) is the only rumour stance and veracity classified tweet dataset which is publicly available. The authors identified rumours associated with major news events as well as tweets associated with those rumours and then annotated each of the tweets for stance and the rumours for their overall veracity. The dataset consists of rumours that emerged during eight different events. However, three of these events evoked less than five rumours consisting of five or more tweets. Therefore, we limit our experiments to the events detailed in Table 1.

Event	Rumours	True / False
Charlie Hebdo	46	24 / 22
Ferguson Riots	34	2 / 32
Germanwings Crash	12	2 / 10
Ottawa Shooting	31	20 / 11
Sydney Siege	50	33 / 17
Total	173	81 / 92

Table 1: Rumours in five events. Each rumour has at least 10 tweets.

Each of the remaining events consists of several rumours and each rumour of several tweet threads. In the data set each tweet is identified by its unique ID, is associated with exactly one event and a unique rumour ID, has a time stamp and is assigned a stance label. Possible stances σ are *supporting, denying, questioning* and *commenting*. Rumours are marked with their veracity values (*true* or *false*).

4 Method: HMM for Veracity Classification

A Hidden Markov Model is a stochastic model in which the system is presumed to be a Markov process including unobservable hidden states. The hidden system undergoes discrete state changes which trigger the emission of observable signals. Based on these signals the hidden states' properties can be learnt.

4.1 Model Generation

More formally, a HMM describes two random processes

$$\{X_t\}_{t\in\mathbb{N}} \text{ and } \{Y_t\}_{t\in\mathbb{N}} \tag{1}$$

of which only the latter is directly observable. A HMM is defined as a 5-tupel

$$\lambda = \{S, E, A, B, \pi\} \tag{2}$$

where $S = \{s_1; \ldots; s_n\}$ is the set of N possible hidden states, $E = \{e_1; \ldots; e_m\}$ is the alphabet of possible observations—i.e. the system's emissions, $A \in \mathbb{R}^{n\times n}$ is the hidden state transition matrix where a_{ij} denotes the probability of the system changing from state i to j, $B \in \mathbb{R}^{n\times m}$ is the emission probability matrix where $b_i(e_j)$ denotes the probability of observing emission $e_j \in \{Y\}$ when the system is in state s_i and finally the starting state probability vector $\pi \in \mathbb{R}^n$, where $p_i = P(X_1 = s_i)$.

For rumour classification we consider tweets related to five events (see Section 3). Consequently, for classification the system λ is defined as follows:

State Space S

Since there is no a-priori solution to determine the optimal size of S (Rabiner, 1990) we consider various hidden state counts $N = \{n \in \mathbb{N} \mid 1 \leq n \leq 15\}$ and repeat calculations accordingly.

Observation Alphabet E

Unlike prior veracity classification approaches, we do not consider the tweets' textual content. Instead, classification is based on stance information alone. Therefore, we set the observation alphabet to

$$E = \{support, deny, question, comment\} \tag{3}$$

Transition Probabilities

Transition matrix A, emission matrix B and start state probability vector π are chosen at random. Baum-Welch parameter optimization algorithm is used with 10 iterators to learn models' final properties. To avoid particularly suboptimal start value configurations and local optima 100 configurations are tested for every $n \in N$. The best performing model is kept while the others are discarded.

4.2 Class Assignment ε

Training data (see Section 5 for details) is used to learn model λ_{false} for false and model λ_{true} for true rumours. For all remaining rumours ε_i (i.e. the testing data) their respective class $C(\varepsilon_i)$ is assigned as follows:

$$C(\varepsilon_i) = \underset{c\in\{false,true\}}{\operatorname{argmax}} \; P(\varepsilon_i|\lambda_c) \tag{4}$$

In (4) the expression $P(\varepsilon_i|\lambda_c)$ is calculated by using the Forward-Algorithm for Hidden Markov Models.

4.3 Multi Spaced HMM Including Tweet Time

As described above, the data set contains tweets' time stamps which are used to build ordered sequences of tweets beginning with the first reply to a rumourous tweet and ending with the last, while the actual distance in time between tweets is disregarded.

Therefore, a straightforward extension of the classification framework is the inclusion of actual posting times. Since basic HMM implementations do not support combining discrete (stance) and continuous (time) emissions into a unified model, we applied Multi Spaced Hidden Markov Models (MSHMM) for this purpose.

MSHMM were introduced by Tokuda et al. (2002) and originally used in speech synthesis tasks. While HMM use probability matrices in the discrete case or one dimensional distribution functions in the continuous case, MSHMM use multi spaced observation probability distributions where the sample space Ω contains G spaces:

$$\Omega = \bigcup_{g=1}^{G} \Omega_g \tag{5}$$

Each Ω_g is a n-dimensional real space R^{n_g} with a space index g. Furthermore, each space has a probability w_g, where

$$\sum_{g=1}^{G} w_g = 1 \tag{6}$$

and an observation probability density function

$$N_g(x), x \in \mathbb{R}^1, \text{ where } \int N_g(x)dx = 1 \tag{7}$$

Established algorithms for HMM can be adapted to multi spaced observation probability functions including Baum-Welch-, Viterbi- and Forward-Backward-Algorithm.

For classification of tweets including their time stamps a MSHMM λ' is defined. Here each stance σ is assigned its own 1-dimensional real space $\Omega_\sigma = R^{1_\sigma}$, while space weights w_σ are determined by stances' occurrence counts. Space probability density functions are learned based on the timestamps which have been pre-processed to represent only the time elapsed since the start of the rumour until a replying tweet occurred.

Accordingly, observation alphabet of MSHMM λ' is defined as a random vector $o = (X, x)$, where X is a space index (i.e. a stance) and $x \in \mathbb{R}^1$ the processed timestamp.

Multi spaced observation probability of o is defined as

$$b(o) = \sum_{g \in X} w_g N_g(x) \tag{8}$$

Additionally, state set S, state transition matrix A and starting state probability vector π of system λ' are defined analogue to system λ.

5 Evaluation Settings

To evaluate the performance of the classification framework leave-one-event-out cross validation was performed, following the methodology of Lukasik et al. (2015). This means that we train on $n - 1$ events (on all rumours within these events) and test it on an unseen n^{th} event (on all rumours within this event). We use F_1 score to measure classifier performance and compare our results against two baselines. The first one does not make use of stance, whereas the second one integrates stance as one of a set of features. These two baselines have been selected to simulate the impact of collective stance i.e., that collective stance have positive impact on rich traditional feature engineered approaches. However, our results (see Section 6) show that it is important how this collective stance information is used to judge the veracity of rumours. We describe these baselines in more details next.

5.1 Stance Unaware Baseline: B1

As described in Section 2, prior work on tweet veracity classification investigated a wide range of features and classification methods. Most influential is the feature set proposed by Castillo et al. (2011), which we have adopted for training a classification model. We experimented with various classifiers such as simple decision trees, k-nn, etc. but achieved best performance with Random Forests, more precisely, we use a range of 33 different features varying from syntactical, semantic, indicator, user-specific and message-specific categories. Details of our features can be obtained from (Aker et al., 2017b).

5.2 Stance Aware Baseline: B2

Following the approach of Liu et al. (2015), we integrate stance as several features, additional to those used in the first baseline. Since there are four different stance classes, the following additional feature(s) are used:

Relative-Stance-Score

Percentage of supporting, denying, questioning, and commenting stances extracted from tweets within a rumour. To obtain this feature we count e.g. how many individual tweets express support for the rumour and divide it by the total number of tweets. We have in total four features, one of each class. Similar to above, we use Random Forest as the machine learning algorithm.

6 Results

Table 2 details a performance comparison for both our systems, as well as B1 and B2. For completeness we also include a simplistic majority-vote baseline on event level. Using system λ we achieved an F_1 score of 0.756 across the 5 events including 173 rumours. The multi spaced system λ' shows superior performance with an F_1 score of 0.804. Both our systems significantly outperform both baselines' F_1 scores of 0.553 and 0.557 respectively.

System	Precision	Recall	F_1
B1	0.650	0.481	0.553
B2	0.661	0.481	0.557
λ	**0.747**	0.765	0.756*
λ'	0.690	**0.963**	**0.804***
majority-vote	0.059	0.025	0.035

Table 2: Overall classification scores.

* indicates significant difference to B1 and B2 (Tukey's HSD $p < 0.05$)

Breaking down the results into precision and recall, we can observe that system λ achieves the highest precision score of 0.747, outperforming both baselines (0.650 and 0.661) as well as system λ'. The latter also gives results slightly more precise than the baselines (0.690). Regarding recall system λ' outperforms all other systems by a great margin achieving a score of 0.963. System λ has also much higher recall of 0.765 comparing to the baseline score of 0.481 for B1 and B2.

Additionally, we look at the individual results for each event (Table 3). It can be seen that the performance varies substantially across events and classifiers. This becomes most obvious in case of the Ferguson event, where B1 and B2 fail to deliver a single true positive result—hence the F_1 score of 0. In other events such as Ottawa Shooting and Sydney Siege all systems have acceptable precision in their results. However, only our proposed (MS-)HMM also have a high to very high recall leading to a vastly superior F_1 score in these events. Another interesting observation can be made regarding the Germanwings Crash event where system λ yields perfect classification while B2 achieves exactly the same results as the overall best performing system λ'. However, it also has to be noted that this particular event contains only 12 rumours and these results therefore have limited expressiveness.

Finally, we also examined system λ''s noticeable low F_1 score of 0.4 in the Ferguson event. Looking at the classification outcome on rumour level we observe that this score is largely caused by the unbalanced nature of the event with only 2 of the 34 rumours being actually true. Out of these true rumours the system managed to capture one (λ: 2; B1: 0; B2: 0) while overall misclassifying only three rumours (λ: 1; B1: 4; B2: 5). In fact, the performance of all classifiers concerning this event is much more similar than the particular F_1 scores suggest.

6.1 Early detection of rumours

For an eventual practical application of our rumour classification system it is also important to consider its performance as a function of time, i.e. how many tweets are necessary to achieve a reasonable classification performance. Therefore, we also explore early detection of rumours by confining our systems to the first ten (first five) tweets only during classification. Classification results using shortened

Charlie Hebdo			
System	Precision	Recall	F_1
B1	0.634	0.792	0.704
B2	**0.667**	0.667	0.667
λ	0.643	0.750	0.692
λ'	0.605	**0.958**	**0.742**
Ferguson			
System	Precision	Recall	F_1
B1	0	0	0
B2	0	0	0
λ	**0.667**	**1**	**0.800**
λ'	0.333	0.500	0.400
Germanwings			
System	Precision	Recall	F_1
B1	0.333	0.500	0.400
B2	0.500	**1**	0.667
λ	**1**	**1**	**1**
λ'	0.500	**1**	0.667
Ottawa			
System	Precision	Recall	F_1
B1	0.818	0.450	0.581
B2	**0.909**	0.500	0.645
λ	0.882	0.750	0.811
λ'	0.792	**0.950**	**0.864**
Sydney Siege			
System	Precision	Recall	F_1
B1	0.714	0.303	0.426
B2	0.647	0.333	0.440
λ	**0.758**	0.758	0.758
λ'	0.750	**1**	**0.857**

Table 3: Performance across events

sequences are summarized in Table 4. As expected the best scores can be achieved by using complete sequences, which feature a median tweet count of 18. For sequences shortened to the first 10 tweets performance drops down to an F_1 score of 0.658 for system λ and 0.642 for system λ' respectively. Further reducing sequences' length to five tweets leads to worse classification performance. However, performance decrease is considerably lower for system λ' with an F_1 score of 0.618 (λ: 0.524).

Additionally, we also retrained our best performing model λ' on shortened sequences using first 10 and first 5 tweets only. Comparing F_1 scores between training conditions we only find marginal decrease of roughly 1% when using the first 10 tweets. However, narrowing down to the first 5 tweets during model training gives unsatisfactory classification results with an F_1 score 0.529.

Overall it can be seen that given only the first 10 tweets at classification time our models still perform better than the baselines. Using even shorter sequences is only reasonable when utilizing the full potential of MSHMM which were still able to beat the baselines while using only the first 5 tweets.

System	All tweets	First 10	First 5
λ	0.756	**0.658**	0.524
λ'	**0.804**	0.642	**0.618**

Table 4: Early detection F_1 scores

6.2 Using Automatic Stance Labels

Results in Table 2 are obtained using gold stance labels. However, this restricts the idea of stance-based rumour verification to only data where human stance labels are available. To overcome this limitation, we have adopted the state-of-the-art stance classifier of Aker et al. (2017a). It has been evaluated on the RumourEval'2017 shared task A on rumour stance classification (Derczynski et al., 2017b) and achieved the best results, namely 79.02% accuracy. The classifier's performance measure was obtained by the RumourEval organizers. Note that for evaluating our models a subset of the data from shared task A has been used. The system performs standard feature engineering such as the extraction of bag-of-words, sentiments, indicator features, etc, but also adds features that are specific to modelling the stance classification problem such as whether the tweet entails some surprise, doubt, certainty and support terms. Apart from feature extraction the system does not require any further parameters to set nor any domain knowledge.

We label automatically all tweets in our dataset with tweet-level stance information using this classifier. Once these stance labels are obtained, we repeat the evaluation of (MS-)HMMs λ and λ'. Results for systems λ_a and λ'_a with automatic stance labels are shown in Table 5.

Both variants show only slight changes in F_1 score compared to their gold data counterparts. In terms of precision and recall λ_a shows a shift towards recall compared to λ—combined with a corresponding loss in precision. However, system λ'_a remains stable in precision and recall while using automatically generated labels. This demonstrates that our veracity classification approach based on the features stance and time has viable practical applications.

System	Precision	Recall	F_1
λ_a	0.632	0.888	0.738
λ'_a	**0.669**	**0.975**	**0.794**

Table 5: Overall scores using automatic labels

7 Discussion

In our results in Table 2 we showed that overall collective stance indeed is an important feature to consider for the purpose of veracity prediction. However, this depends on how this collective feature is used. When stance is added as additional feature to those features reported by related work (setting B2) we could only gain a neglectable improvement. On the other side, when collective stance was used within HMM we observe superior results indicating that using HMM is a better way for capturing the crowd stance wisdom and applying this effectively on the veracity classification task. In case of baseline B2 we used rather a crude way of capturing the stance wisdom by counting different stance types. However, collective stance might obey some specific patterns of development, as indicated by Mendoza et al. (2010), and capturing these patterns is an important factor. This is where our systems λ and λ' shine and the power of capturing this is reflected in the ample classification performance increase. However, this increase is not distributed equally across all events. In the following paragraphs we are going to raise a few points that likely have contributed to these outcomes:

Our MSHMM λ' overall produces results which are especially distinct from the other models' results by being able to correctly classify 12 rumours where all other models fail. Note that this occurs in only two and one cases for baselines B1 and B2 respectively. Further investigation of these 12 correctly classified rumours shows that they tend to be shorter than average with a median length of 11 tweets. Interestingly, λ' correctly classifies true rumours in 9 out of the 12 cases, although they contain only 1 to 3 *supporting* stances each.

On the other hand, there are also 15 rumours that are solely misclassified by our HMM λ'. Again, this is less often the case for the baselines with 8 occurrences each for model B1 and B2. As it is also indicated by the comparably lower precision scores, model λ's produces more false positive classifications than the other models, especially concerning the events *Sydney siege* and *Ferguson* (B1: 15; B2: 14; λ: 18; λ': 26). Consequently, all 15 misclassified rumours are false positives. Further investigation of the individual rumours' properties remained inconclusive.

Overall the baselines have a negative bias, i.e. they tend to classify rumours as false. While the actual true/false ratio of all rumours is 46.8% true, model B1 classifies 35% of the rumours as true (B2 34%). On the contrary, λ shows no bias with a ratio of 48%, while λ' has a positive bias with a ratio of 64.3%. This observation is in line with the baselines F_1 score of 0 for rumours related to the *Ferguson riots* event where only 2 out of all 34 rumours are true. Out of these two true rumours one was found by both our models while the other was correctly classified solely by model λ.

We investigated our models' performance on short sequences of tweets to accommodate for the fact that a timely detection of rumours would be of great practical benefit. Naturally, we experience a performance drop when reducing sequences' length. We believe that this is due to the collective stance being expected to stabilize only over time (see Section 1). However, we indeed could show that especially our MSHMM is capable of capturing a rumours' veracity very fast with a high level of recall and sufficient precision, still outperforming the baselines even when using only the first five tweets of a rumour.

Especially the high recall affirms our intend to adapt our framework for the task of rumour detection, which is a necessary step before rumour veracity classification can be performed. It will be interesting to observe in future work whether the model can perform equally well at this task.

As described in Section 6 our model λ' including tweets' times achieved best performance values on this particular data set. This result—as well as the strong result of model λ—shows our HMM-based method's general applicability to the problem at hand. Furthermore, we have demonstrated the general suitability and classification stability of our automated stance annotation framework. Since we seem to have overcome the need for manual annotation when using MSHMM in future work the data sets can be extended to more recent events featuring potentially large amounts of tweets.

However, it is also reasonable to assume that events are heterogeneous in their stance distribution patterns, which might have an impact on classification performance and generalizability across events. Therefore, in subsequent analysis different event types should be considered when training MSHMM. A basic event distinction might be *sharp*, sudden events that also feature a definite ending vs. *soft* events where multiple sub-events occur that trigger new developments in the corresponding discussion threads.

Finally, note that the discussed HMM approach achieves results in the 80% margin in terms of F_1 when it uses gold standard stance. This performance drops insignificantly when the stance information is obtained by automatic methods. With that the stance information is the only dependent variable for the HMM and any performance improvement on that site will, as the results on gold standard stance show, also increase the performance of the HMM-based classifier. This is also valid for new rumours where it can be expected that the current automatic stance detection approach does not perform as well as in the SemEval2017 data and thus have negative impact on the verification results. This is a known problem that there is a performance drop when the system moves to new unseen data. However, the performance gap can be closed with extending training data (obtained either manually or using some distance learning) and focusing more on domain independent features.

8 Conclusions

This paper investigated whether rumour stance alone and combined with tweets' times can be used to predict rumour veracity automatically. We followed two different strategies in applying stance for the verification task. In the first strategy we added stance as an additional feature to those commonly used in earlier studies. We demonstrated that including collective stances observed in sequences of tweets leads to only neglectable performance increase when using approaches adapted from the literature. The second strategy was to use stance as the only feature to model true and false rumours with discrete HMM. Finally, we employed multi-spaced HMM to jointly model the temporal changes in stance information. We demonstrated that already using stance-based HMM without time information leads to substantially better classification results over the first strategy. Though, the extension to multi-spaced HMM with time incorporated has led to even superior results with an F_1 score of 80.4%.

Next, we will investigate further the integration of temporal information for the rumour detection task. Furthermore, we will explore if veracity classification can be improved further by combining our HMM-based classifier with other state-of-the-art approaches. In addition, we plan to adapt the crowd stance along with HMM for the fake news detection task.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media", the European Union funded COMRADES project (grant agreement No. 687847), and an EPSRC career acceleration fellowship (EP/I004327/1). Last but not least we thank all anonymous reviewers for their valuable feedback.

References

Ahmet Aker, Leon Derczynski, and Kalina Bontcheva. 2017a. Simple open stance classification for rumour analysis. *CoRR*, abs/1708.05286.

Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, Anna Kolliakou, Rob Procter, and Maria Liakata. 2017b. Stance classification in out-of-domain rumours: A case study around mental health disorders. In *International Conference on Social Informatics*, pages 53–64. Springer.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM.

Cheng Chang, Yihong Zhang, Claudia Szabo, and Quan Z Sheng. 2016. Extreme user and political rumor detection on twitter. In *Advanced Data Mining and Applications: 12th International Conference, ADMA 2016, Gold Coast, QLD, Australia, December 12-15, 2016, Proceedings*, pages 751–763. Springer.

Weiling Chen, Chai Kiat Yeo, Chiew Tong Lau, and Bu Sung Lee. 2016. Behavior deviation: An anomaly detection view of rumor preemption. In *Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2016 IEEE 7th Annual*, pages 1–7. IEEE.

Yi-Chin Chen, Zhao-Yand Liu, and Hung-Yu Kao. 2017. IKM at SemEval-2017 Task 8: Convolutional Neural Networks for Stance Detection and Rumor Verification. In *Proceedings of SemEval*. ACL.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017a. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of SemEval*. ACL.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017b. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of SemEval*. ACL.

Omar Enayet and Samhaa R. El-Beltagy. 2017. NileTMRG at SemEval-2017 Task 8: Determining Rumour and Veracity Support for Rumours on Twitter. In *Proceedings of SemEval*. ACL.

Mengdie Hu, Shixia Liu, Furu Wei, Yingcai Wu, John Stasko, and Kwan-Liu Ma. 2012. Breaking news on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2751–2754. ACM.

Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th International Conference on Data Mining*, pages 1103–1108. IEEE.

Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. 2017. Rumor detection over varying time windows. *PLOS ONE*, 12(1):e0168344.

Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. 2015. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1867–1870. ACM.

Michal Lukasik, Trevor Cohn, and Kalina Bontcheva. 2015. Classifying tweet level judgements of rumours in social media. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 2590–2595.

Michal Lukasik, PK 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 54th Annual Meeting of the Association for Computational Linguistics*, pages 393–398. Association for Computational Linguistics.

Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1751–1754. ACM.

Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. Twitter under crisis: can we trust what we rt? In *Proceedings of the first workshop on social media analytics*, pages 71–79. ACM.

Rob Procter, Jeremy Crump, Susanne Karstedt, Alex Voss, and Marta Cantijoch. 2013. Reading the riots: What were the police doing on twitter? *Policing and society*, 23(4):413–436.

Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of EMNLP*, pages 1589–1599.

Lawrence R. Rabiner. 1990. Readings in speech recognition. chapter A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, pages 267–296. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Vikram Singh, Sunny Narayan, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharya. 2017. IITP at SemEval-2017 Task 8: A Supervised Approach for Rumour Evaluation. In *Proceedings of SemEval*.

Ankit Srivastava, Rehm Rehm, and Julian Moreno Schneider. 2017. DFKI-DKT at SemEval-2017 Task 8: Rumour Detection and Classification using Cascading Heuristics. In *Proceedings of SemEval*, pages 486–490. ACL.

Keiichi Tokuda, Takashi Masuko, Noboru Miyazaki, and Takao Kobayashi. 2002. Multi-space probability distribution HMM. *IEICE TRANSACTIONS on Information and Systems*, 85(3):455–464.

Guangmo Tong, Weili Wu, Ling Guo, Deying Li, Cong Liu, Bin Liu, and Ding-Zhu Du. 2017. An efficient randomized algorithm for rumor blocking in online social networks. *arXiv:1701.02368*.

Soroush Vosoughi. 2015. *Automatic detection and verification of rumors on Twitter*. Ph.D. thesis.

Shihan Wang and Takao Terano. 2015. Detecting rumor patterns in streaming social media. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 2709–2715. IEEE.

Feixiang Wang, Man Lan, and Yuanbin Wu. 2017. ECNU at SemEval-2017 Task 8: Rumour Evaluation Using Effective Features and Supervised Ensemble Models. In *Proceedings of SemEval*, pages 491–496. ACL.

Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st International Conference on Data Engineering*, pages 651–662. IEEE.

Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. 2012. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, page 13. ACM.

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):1–29, 03.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):32.