Hawkes Processes for Continuous Time Sequence Classification: an Application to Rumour Stance Classification in Twitter

Michal Lukasik, P.K. Srijith, Duy Vu, Kalina Bontcheva, Arkaitz Zubiaga, and Trevor Cohn

1Department of Computer Science, The University of Sheffield
2Department of Computing and Information Systems, The University of Melbourne
3Department of Computer Science, The University of Warwick

{m.lukasik, pk.srijith, k.bontcheva}@shef.ac.uk
{duy.vu, t.cohn}@unimelb.edu.au
a.zubiaga@warwick.ac.uk

Abstract

Classification of temporal textual data sequences is a common task in various domains such as social media and the Web. In this paper we propose to use Hawkes Processes for classifying sequences of temporal textual data, which exploit both temporal and textual information. Our experiments on rumour stance classification on four Twitter datasets show the importance of using the temporal information of tweets along with the textual content.

1 Introduction

Sequence classification tasks are often associated with temporal information, where the timestamp is available for each of the data instances. For instance, in sentiment classification of reviews in forums, opinions of users are associated with a timestamp, indicating the time at which they were posted. Similarly, in an event detection task in Twitter, tweets being posted on a continuous basis need to be analysed and classified in order to detect the occurrence of some event. Nevertheless, traditional sequence classification approaches (Song et al., 2014; Gorrell and Bontcheva, 2016) ignore the time information in these textual data sequences. In this paper, we aim to consider the continuous time information along with the textual information for classifying sequences of temporal textual data. In particular, we consider the problem of rumour stance classification in Twitter, where tweets provide temporal information associated with the textual tweet content.

Rumours spread rapidly through social media, creating widespread chaos, increasing anxiety in society and in some cases even leading to riots. For instance, during an earthquake in Chile in 2010, rumours circulating on Twitter stated that a volcano had become active and there was a tsunami warning, which were later proven false. Denials and corrections of these viral pieces of information might often come late and without the sufficient effect to prevent the harm that the rumours can produce (Lewandowsky et al., 2012). This posits the importance of carefully analysing tweets associated with rumours and the stance expressed in them to prevent the spread of malicious rumours. Determining the stance of rumour tweets can in turn be effectively used for early detection of the spread of rumours, as well as for flagging rumours as being potentially false when a large number of people are found to be countering them. The rumour stance classification task has been previously defined as that in which a classifier needs to determine whether each of the tweets is supporting, denying or questioning a rumour (Qazvinian et al., 2011). Here we add a fourth label, commenting, which is assigned to tweets that do not add anything to the veracity of a rumour.

In this paper, we propose to use Hawkes Processes (Hawkes, 1971), commonly used for modelling information diffusion in social media (Yang and Zha, 2013; De et al., 2015), for the task of rumour stance classification. Hawkes Processes (HP) are a self-exciting temporal point process ideal for modelling the occurrence of tweets in Twitter (Zhao et al., 2015). The model assumes that the occurrence of a tweet will influence the rate at which future tweets will arrive. Figure 1 shows the behaviour of the intensity functions associated with a multivariate Hawkes Process. Note the intensity spikes at the points of tweet occurrences. In applications such as stance classification, different labels can influence one another. This can be modelled effectively using the mutually exciting behaviour of Hawkes Processes. In the end, we demonstrate how the information gar-
We define the stance classification task as that in which each tweet \( d_j \) needs to be classified into one of the four categories, \( y_j \in Y \), which represents the stance of the tweet \( d_j \) with respect to the rumour \( R_i \) it belongs to.

We consider the Leave One Out (LOO) setting, introduced by Lukasik et al. (2015a), where for each rumour \( R_i \in D \) we construct the test set equal to \( R_i \) and the training set equal to \( D \setminus R_i \). The final performance scores we report in the paper are averaged across all rumours. This represents a realistic scenario where a classifier has to deal with a new, unseen rumour.

3 Data

We consider four Twitter rumour datasets with tweets annotated for stance (Zubiaga et al., 2016). The authors relied on a slightly different scheme for the annotation, given that they annotated tree-structured conversation threads where a source tweet initiates a rumour and a number of replies follow responding to it. Given this structure, the source tweet of a Twitter conversation is annotated as supporting, denying or underspecified, and each subsequent tweet is annotated as agreed, disagreed, appeal for more information (questioning) or commenting with respect to the source tweet. We convert these labels into our set of four including supporting, denying, questioning and commenting, which extends the set of three labels used before in the literature (Qazvinian et al., 2011; Lukasik et al., 2015a) adding the new label commenting. To perform this conversion, we first remove rumours where the source tweet is annotated as underspecified, keeping the rest of source tweets as supporting or denying. For the subsequent tweets, we keep their label as is for the tweets that are questioning or commenting. To convert those tweets that agree or disagree into supporting or denying, we apply the following set of rules: (1) if a tweet agrees to a supporting source tweet, we label it supporting, (2) if a tweet agrees to a denying source tweet, we label it denying, (3) if a tweet disagrees to a supporting source tweet, we label it denying and (4) if a tweet disagrees to a denying tweet, we label it supporting. The latter enables to infer stance with respect to the rumour from the original annotations that instead refer to agreement with respect to the source.

While the authors annotated and released 9 datasets, here we make use of 4 sufficiently large datasets.
Figure 2 shows examples of tweets taken from the dataset along with our inferred annotations.

We summarise the statistics of the resulting dataset in Table 1. Note that the commenting label accounts for the majority of the tweets.

| Dataset            | Rumours | Tweets | Supporting | Denying | Questioning | Commenting |
|--------------------|---------|--------|------------|---------|-------------|------------|
| Ottawa shooting    | 58      | 782    | 161        | 76      | 64          | 481        |
| Ferguson riots      | 46      | 1017   | 161        | 82      | 94          | 680        |
| Charlie Hebdo       | 74      | 1053   | 236        | 56      | 51          | 710        |
| Sydney siege        | 71      | 1124   | 89         | 223     | 99          | 713        |

Table 1: Statistics and distribution of labels for the four datasets used in our experiments. Each dataset consists of multiple rumours, and the rest of the columns offer the aggregated counts for all rumours within that dataset.

4 Model

Hawkes Processes are a probabilistic framework for modelling self-exciting phenomena, which has been used for modelling memes and their spread across social networks (Yang and Zha, 2013). They have been used to model the generation of tweets over a continuous time domain (Zhao et al., 2015). The frequency of tweets generated by them is determined by an underlying intensity function which considers the influence from past tweets. The intensity function models the self-exciting nature by adding up the influence from past tweets. We use a multi-variate Hawkes process for modelling the mutually exciting phenomena between the tweet labels. In this section we describe how we apply the Hawkes Process framework for rumour stance classification.

Intensity Function In the intensity function formulation, we assume that all previous tweets associated with a rumour influence the occurrence of a new tweet. This allows to use information on all the other tweets that have been posted about a rumour. We consider the intensity function to be summation of base intensity and the intensities associated with all the previous tweets,

\[
\lambda_{y,m}(t) = \mu_y + \sum_{t_\ell < t} \Pi_{(m_\ell = m)} \kappa(t - t_\ell),
\]

where the first term represents the constant base intensity of generating label \(y\). The second term represents the influence from the tweets that happen prior to time of interest. The influence from each tweet decays over time and is modelled using an exponential decay term \(\kappa(t - t_\ell) = \omega \exp(-\omega(t - t_\ell))\). The matrix \(\alpha\) of size \(|Y| \times |Y|\) encodes the degrees of influence between pairs of labels assigned to the tweets, e.g. a questioning label may influence the occurrence of a rejecting label in future tweets differently from how it would influence a commenting label.

Likelihood function The parameters governing the intensity function are learnt by maximizing the likelihood of generating the tweets. The complete likelihood function is given by

\[
L(t, y, m, W) = \prod_{n=1}^N p(W_n|y_n) \prod_{n=1}^N \lambda_{y_n, m_n}(t_n) \times p(E_T),
\]

where the first term provides the likelihood of generating text given the label and is modelled as a multinomial distribution conditioned on the label,

\[
p(W_n|y_n) = \prod_{v=1}^V \beta_{y_n, v}^{W_n|v},
\]

where \(V\) is the vocabulary size and \(\beta\) is the matrix of size \(|Y| \times V\) specifying the language model for each label. The second term provides the likelihood of occurrence of tweets at times \(t_1, \ldots, t_n\) and the third term provides the likelihood that no tweets happen in the interval \([0, T]\) except at times \(t_1, \ldots, t_n\). We estimate the parameters of the model by maximizing the log-likelihood,

\[
l(t, y, m, W) = -\sum_{y=1}^{|Y|} \sum_{m=1}^{|D|} \int_0^T \lambda_{y,m}(s) ds + \sum_{n=1}^N \log \lambda_{y_n, m_n}(t_n) + \sum_{n=1}^N \sum_{v=1}^V W_{nv} \log \beta_{y_n, v}.
\]

The integral term in Equation (4) is easily computed for the intensity function since the exponential decay function and the constant function are easily integrable.
Rumour 1 - u1: We understand there are two gunmen and up to a dozen hostages inside the cafe under siege at Sydney. ISIS flags remain on display #7News [supporting]
Rumour 1 - u2: @u1 sorry - how do you know it’s an ISIS flag? Can you actually confirm that? [questioning]
Rumour 2 - u1: These are not timid colours; soldiers back guarding Tomb of Unknown Soldier after today’s shooting #StandforCanada –PICTURE– [supporting]
Rumour 2 - u2: @u1 This photo was taken this morning, before the shooting. [denying]
Rumour 2 - u3: @u1 More on situation at Martin Place in Sydney, AU –LINK– [commenting]

Figure 2: Examples of rumour tweets associated with two different rumours.

Note that \( \beta \) is independent from the dynamics part, and a closed form solution after applying Laplacian smoothing takes form

\[
\beta_{yv} = \frac{\sum_{n=1}^{N} \mathbb{I}(y_n = y)W_{nv} + 1}{\sum_{n=1}^{N} \sum_{v=1}^{V} \mathbb{I}(y_n = y)W_{nv} + V}.
\]

In one approach to \( \mu \) and \( \alpha \) optimization (HP Approx.) we approximate the log term in Equation (4) by taking the log inside the summation terms in Equation (1). This approximation leads to closed form updates for \( \mu \) and \( \alpha \),

\[
\mu_y = \frac{\sum_{n=1}^{N} \mathbb{I}(y_n = y)}{T|D},
\]

\[
\alpha_{ij} = \frac{\sum_{n=1}^{N} \sum_{l=1}^{n} \mathbb{I}(m_l = m_n)\mathbb{I}(y_l = i)\mathbb{I}(y_n = j)}{\sum_{k=1}^{N} \mathbb{I}(y_k = i)K(T - t_k)},
\]

where \( K(T - t_k) = 1 - \exp(-\omega(T - t_k)) \) arises from the integration of \( \kappa(t - t_k) \).

In a different approach (HP Grad.) we find parameters using joint gradient based optimization over \( \mu \) and \( \alpha \), using derivatives of log-likelihood \( \frac{d}{d\mu} \) and \( \frac{d}{d\alpha} \). In optimization, we operate in the log-space of the parameters in order to ensure positivity, and employ L-BFGS approach to gradient search. Moreover, we initialize parameters with those found by the HP Approx. method.

Similar to Yang and Zha (2013), we fix the decay parameter \( \omega \) in our case to 0.1.

**Prediction** We predict the most likely label for each test tweet as the label which maximises the likelihood of occurrence of the tweet from Equation (2), or the approximated likelihood in case of HP Approx. The likelihood considers both the textual information and the temporal dynamics in predicting the label for the tweet. The predicted labels are then considered while predicting the labels for next tweets in the test data. Thus, we follow a greedy sequence classification approach.

## 5 Experiments

We conduct experiments using the rumour datasets described in Table 1. We consider our Hawkes Process model described in Section 4 as well as a set of baseline and benchmark approaches.

### 5.1 Baselines

We compare our model against baselines:

**Language Model** considers only the textual information through multinomial distribution defined in Equation (3).

**Majority vote** classifier based on the training label distribution.

**Naive Bayes** models the text using a multinomial likelihood and a prior over label frequencies (Manning et al., 2008).

Note that Multinomial, Majority vote and Naive Bayes approaches are special cases of our Hawkes Process model for classification, where a particular subset of parameters is fixed to 0.

### 5.2 Benchmark models

We compare our model against the following competitive benchmark models:

**SVM** Support Vector Machines with the cost coefficient selected via nested cross-validation.

**GP** Gaussian Processes have been shown by Lukasik et al. (2015a) to work well, particularly in supervised settings where a multi-task learning kernel has been used to learn correlations across different rumours. Here we use a single task kernel (linear) as we consider the fully unsupervised setting.

**CRF** Conditional Random Field (Lafferty et al., 2001) over temporally ordered sequences using both text and neighbouring label features. The model is trained using \( \ell_2 \) penalized log-likelihood where the regularisation parameter...
commenting causes a spike in intensity for questioning lifespan a ample around 1 hour and 50 minutes into the ru-

tweet from a different label is posted, for ex-
sity functions for different labels at times when
self-exciting property, with spikes in the inten-
tensity function of the HP Grad.

helpful than the simple frequency based prior from
ral dynamics based prior provided by HP is more
time-dependent prior. This shows, that the tempo-
tive according to F
GP baseline, which turns out to be very competi-
t metrics on all datasets. This is the case also for
HP Grad.

notice that different rankings of methods. Nevertheless, we can
ment for rumour #1 from the Ferguson dataset. Notice the
self-exciting property, with spikes in the intensity
functions for different labels at times when tweets occur. Moreover, spikes occur even when
a tweet from a different label is posted, for ex-
ample around 1 hour and 50 minutes into the ru-
mour lifespan a questioning tweet is posted which
causes a spike in intensity for commenting tweets.

Another issue is the approximation used in HP
Approx. which might lead to violation of the
Hawkes Process mutual-excitation property. In
particular, we noticed that in some scenarios oc-
currences of tweets cause decrease in the intensity
value rather than spikes. However, the accuracy
metric which has been used in previous work for
this task (Lukasik et al., 2015a) yielded by this
method turns out to be the best, although when
measuring F1 the relative ordering changes with
the GP performing best (Lukasik et al., 2015a)
closely followed by other techniques including HP
Grad. which is competitive on all datasets.

### 6 Conclusions

We proposed a novel model based on Hawkes Processes for sequence classification of stances in Twitter which takes into account temporal information in addition to text. Using four Twitter datasets and experimenting on rumour stance classification of tweets, we have shown that HP is a competitive approach, which outperforms a range of strong benchmark methods by providing the multinomial language model with an informative prior based on temporal dynamics. Our experiments posit the importance of making use of temporal information available in tweets, which along with the textual content provide valuable information for the model to perform well on the task.

### Acknowledgments

The work was supported by the European Union under grant agreement No. 611233 PHEME. Cohn was supported by an ARC Future Fellowship scheme (project number FT130101105).
References

Abir De, Isabel Valera, Niloy Ganguly, Sourangshu Bhattacharya, and Manuel Gomez-Rodriguez. 2015. Modeling opinion dynamics in diffusion networks. *CoRR*, abs/1506.05474.

Genevieve Gorrell and Kalina Bontcheva. 2016. Classifying twitter favorites: Like, bookmark, or thanks? *Journal of the Association for Information Science and Technology*, 67(1):17–25.

Alan G. Hawkes. 1971. Spectra of Some Self-Exciting and Mutually Exciting Point Processes. *Biometrika*, 58(1):83–90.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. of International Conference on Machine Learning (ICML)*, pages 282–289.

Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131.

Michal Lukasik, Trevor Cohn, and Kalina Bontcheva. 2015a. Classifying Tweet Level Judgements of Rumours in Social Media. In *Proc. of Empirical Methods in Natural Language Processing (EMNLP)*, pages 2590–2595.

Michal Lukasik, Trevor Cohn, and Kalina Bontcheva. 2015b. Point process modelling of rumour dynamics in social media. In *Proc. of the 53rd ACL*, vol. 2, pages 518–523.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proc. of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1589–1599.

Ge Song, Yunming Ye, Xiaolin Du, Xiaohui Huang, and Shifu Bie. 2014. Short text classification: A survey. *Journal of Multimedia*, 9(5).

Shuang-Hong Yang and Hongyuan Zha. 2013. Mixture of mutually exciting processes for viral diffusion. In *Proc. of International Conference on Machine Learning (ICML)*, volume 28, pages 1–9.

Qingyuan Zhao, Murat A. Erdogdu, Hera Y. He, Anand Rajaraman, and Jure Leskovec. 2015. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proc. of International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1513–1522.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Gerdine 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.