Hawkes Process Classification through Discriminative Modeling of Text

Rohan Tondulkar
IIT Hyderabad
cs17mtech11028@iith.ac.in

Manisha Dubey
IIT Hyderabad
cs17resch11003@iith.ac.in

Srijith P. K.
IIT Hyderabad
srijith@iith.ac.in

Michal Lukasik*
Google Inc.
mlukasik@google.com

2022 International Joint Conference on Neural Networks (IJCNN) | 978-1-7281-8671-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/IJCNN55064.2022.9892868

***Abstract*—Social media such as Twitter has provided a platform for users to gather and share information and stay updated with the news. However, restriction on the length, informal grammar and vocabulary of the posts pose challenges to perform classification from textual content alone. We propose models based on the Hawkes process (HP) which can naturally incorporate additional cues such as the temporal features and past labels of the posts, along with the textual features for improving short text classification. In particular, we propose a discriminative approach to model text in HP, where the text features parameterize the base intensity and the triggering kernel of the intensity function. This allows textual content to determine influence from past posts and consequently determine the intensity function and class label. Another major contribution is to model the kernel as a neural network function of both time and text, permitting more complex influence functions for Hawkes process. This will maintain the interpretability of Hawkes process models along with the improved function learning capability of the neural networks. The proposed HP models can easily consider pre-trained word embeddings to represent text for classification. Experiments on the rumour stance classification problems in social media demonstrate the effectiveness of the proposed HP models.**

***Index Terms*—Hawkes Process, Discriminative Modeling, Text Classification**

I. Introduction

Social media provides a platform for common users to share information, generally in the form of short snippets of text, with a prominent example being Twitter. Mining and classification of social media posts can be useful in addressing various real world problems. For instance, it can help in rescue and relief operations during disasters, understanding the stance or opinion of people towards a product etc. However, modeling tweets is a difficult task since tweets involve frequent use of informal grammar as well as irregular vocabulary e.g. abbreviations, typographical errors and hashtags. Another associated problem is that the tweets (or micro-blogs) have a 280 character limit imposed by Twitter. Since these texts are short in nature, and has noisy word patterns, classification of social media posts is extremely challenging.

Text classification problems related to social media like stance classification have been receiving a lot of interest due to the necessity to debunk contentious claims circulated in social media. For instance, social media has become the starting point for many rumours and fake news, and rumour stance classification will help to determine the veracity of rumours. The rumour stance classification task intends to assist in this verification process by determining the type of support expressed in different tweets discussing the same rumour [39]. In this task, we classify the stance of the posts following a would-be rumour post as *supporting*, *denying*, *questioning* or *commenting* about the rumour. We can observe that the stance associated with a post depends on the labels associated with the past posts, e.g. in rumour stance classification if the past tweets have *questioning* label, then the current or future tweets will have *denying* labels. Approaches based on sequence labeling models such as LSTMs [23] consider additional cues in the form of past labels to improve rumour stance classification. However, it is also important to consider the temporal aspects associated with the posts. If the past post has happened long before (time difference of posts or inter-arrival times is large) then the influence of their labels on the current post will be less. Though LSTM models consider sequences in which posts arrive, they don't naturally consider the exact times associated with the posts. Though time can be considered as an additional input feature, it may not be effective when considered along with a high dimensional text data. Moreover, these deep learning models are black box models which suffer from interpretability issues. For instance, one cannot learn *label-label* influences or *post-post* influences, though attention models try to overcome this to some extent. Statistical models such as Hawkes process excels in these aspects and can overcome these limitations of deep learning models. Hawkes process have been proposed to perform rumour stance classification in [22], considering text, time, and past labels associated with social media posts.

In social networking platforms like Twitter, previous tweets can influence a response in the form of another tweet and consequently the label associated with it. Such characteristics like self or mutual excitation can be modeled easily using a Hawkes process (HP) [1] with an appropriate intensity function. They have been extensively used for solving various problems arising in social media [7], [31]. The Hawkes process approach to stance classification of social media posts [22] considers the intensity of a post as a function of past labels and time. The effect of text is modeled following a generative approach by considering an additional class conditional distribution over text along with standard HP likelihood. However, there are various disadvantages associated with this approach.

*Work done prior to joining Google

978-1-7281-8671-9/22/$31.00 ©2022 IEEE

Firstly, this generative model is restrictive as it does not consider text in determining the intensity function for a post but only label and time. Further, text can also play a major role in determining the influence from past labels. Posts with similar textual content tend to have higher influence in determining the stance of the current post. Moreover, stance classification or any text classification can also benefit a lot by considering an appropriate representation of text using word embeddings such as Glove and Word2vec [32], [33]. The HP based stance classification model in [22] fails to capture these aspects. We propose Hawkes process models based on discriminative modeling of text which addresses these concerns and can provide several other advantages in terms of modeling capability.

The proposed HP models for stance classification consider textual features as a part of the intensity function. It can capture the influence of past posts not only based on their time of occurrence but also based on textual contents through the use of kernels which are functions of both text and time. Moreover, we propose to use a neural kernel in Hawkes process which can learn the functional form of the influence from data rather than predefining it as exponential function as in the prior works. Incorporating neural kernels in HP models will provide dual benefit: the interpretability advantage of HP and universal approximation capability of neural networks. Also, the proposed HP model can easily consider word embeddings to represent text which will further help in improving stance classification of social media posts. We show the usefulness of the proposed HP models for the rumour stance classification problem on Twitter. The proposed models are generic and can be applied to any text classification problems involving a temporal dimension.

Contributions Our main contributions can be summarized as follows:

- We propose HP models which consider discriminative modeling of text for stance classification through the intensity function.
- We consider the effect of textual content in determining the influence of historical posts through the use of kernels which consider text and time.
- We propose using neural networks to model kernels in HP intensity function in order to learn complex non-linear influences.
- Use of word embeddings to represent text through proposed HP classification models.

II. Related Work

Effective classification of short text in social media requires considering additional cues such as past labels. Modeling stance classification problems using sequence labeling approaches such as LSTMs helps to capture the effect of past labels in determining current stance. Several LSTM based models [17], [23], [27] were proposed to perform stance classification of social media posts. However, black box deep learning models suffer from interpretability and do not consider exact posting times and influences which can be useful to perform the stance classification task.

Multivariate Hawkes processes are found to be very useful in modeling problems in social media [7]. However, there exist very few works in literature where Hawkes process is used for language modeling problems. [34] focusses on the problem of inferring the diffusion of information together with the topics characterizing the information using Hawkes process and topic modeling. Another work related to topic modeling is [35] where authors have proposed Hidden Markov Hawkes Process that incorporates topical Markov Chains within Hawkes processes to jointly model topical interactions along with user-user and user-topic patterns. [36] has used the combination of Dirichlet process and Hawkes process for clustering document streams. They have been used for various applications like detecting fake retweets [30] and modeling of COVID-19 Twitter narratives [37].

A closely related work is [22] where hawkes process is developed to perform stance classification of social media posts. However, previous approaches model temporal part and language modeling part as two separate likelihoods and consider their product as the joint likelihood. For instance, [22] used a multivariate Hawkes Process (MHP) to model the influence of stance labels and time, and used a separate class conditional density (multinomial distribution) to model short texts. This generative model is restrictive and does not allow text to be represented using word embeddings and to be used for determining influence from past posts. We propose HP models which can overcome these limitations by considering text in a discriminative manner through the intensity function. Moreover, we propose a hybrid model where the influence function is modeled as a neural network, and can enjoy interpretability of HP and function learning capability of neural networks. This is different from the previous works [9]–[11] used for diffusion modeling but not language modeling, where the full intensity function is modeled using a neural network losing the interpretability advantage of HP models.

III. Problem Statement

We consider tweets associated with D topics (or statements or claims) of interest for stance classification. Each tweet is represented as a tuple $d_j = (t_j, X_j, m_j, y_j)$, which includes the following information: t_j is the posting time of the tweet, X_j is the text message, m_j is the topic (or rumour) category and y_j is the stance of the tweet towards a topic (or rumour). In particular, we consider rumour stance classification where $y_j \in Y$= *{supporting, denying, questioning, commenting}*. The stance classification task is to classify the tweet d_j to a stance class $y_j \in Y$.

IV. Background

A. Point Process

A point process is a random process which models the occurrence of a set of points in some space. A point process

is characterized by its conditional intensity function defined as -

$$\lambda(t|\mathcal{H}_t) = \lim_{h \to 0} \frac{P(N_{t+h} - N_t) = 1|\mathcal{H}_t)}{h} \tag{1}$$

where $\mathcal{H}_t$ is the history of the events up to time t, occurring at times $\{t_1, t_2, ...t_n\}$ and N_t is the count of events until time t. The intensity function models the instantaneous occurrence of an event at time t. There exist different types of point processes such as Poisson process and Hawkes process depending on the way the intensity function is defined.

B. Hawkes Process

A Hawkes process [5] is a point process with self-triggering property, i.e occurrence of the previous events trigger occurrences of future events. Conditional intensity function for univariate Hawkes process is defined as $\lambda(t) = \mu + \sum_{t_k<t} k(t - t_k)$, where μ is the base intensity function and $k(\cdot)$ is the triggering kernel function capturing the influence from previous events. The summation over $t_k < t$ represents the effect of all the events prior to time t and will contribute in computing the intensity at time t. Typically, $k(\cdot)$ is considered as an exponentially decaying function of time capturing that the influence of past events decreases exponentially over time.

C. Hawkes Process for Stance Classification

[22] proposed a Hawkes process based approach for stance classification of posts in social media. The approach used a multivariate Hawkes process (MHP) to capture the influence of past labels and their time of occurrences. The intensity function at time t for a post belonging to stance y and topic (rumour) m is given by

$$\lambda_{y,m}(t) = \mu_y + \sum_{t_\ell<t} \mathbf{I}_{m_\ell=m} \alpha_{y_\ell,y} k(t - t_\ell) \tag{2}$$

where $\mathbf{I}_{m_\ell=m}$ is the indicator function taking value 1 when m_ℓ is m, otherwise 0. The base intensity μ_y is a constant per stance label and the triggering kernel $k(t - t_\ell) = \omega e^{-\omega(t-t_\ell)}$ captures the influence from the past events. The matrix α of size $|Y| \times |Y|$ captures the influence between various stance labels. For instance, *Support* label may have less influence on the *Deny* label but has higher influence on the *Comment* label. However, this influence can be low if the future tweets are happening far ahead in time. This is captured by multiplying the influence matrix with the exponentially decaying kernel.

In order to capture the effect of textual contents of a post on the stance label, Multinomial distribution is used to model the class conditional generation of text $p(X_n|y_n)$. The final likelihood is obtained by multiplying the intensity function likelihood with the class conditional probability $p(X_n|y_n)$. We can observe that the intensity function does not consider the text data and consequently does not consider text in determining the influence from the past events. This generative model further restricts using word embeddings to represent text or require considerable changes in the existing model.

V. Discriminative Modeling of Text

We aim to overcome drawbacks of the existing approaches through discriminative modeling of text along with time in the intensity function of the Hawkes process. We discuss different ways to model text in the intensity function (2), through the base intensity as well as through the triggering kernels. Along with time based kernels, text based kernels can model the impact of historical events better. We also introduce a methodology where we use a neural network, which is a universal function approximator, as a kernel to model text and time. We discuss the proposed models (Figure 1) in detail in the following sections.

A. Base Textual HP: Modeling base intensity using textual features

The base intensity influences the arrival of events due to exogenous factors. In a standard Hawkes process model, base intensity is constant and learnt from the data. However, we propose a model (*Base Textual HP*) where base intensity considers the textual features. Along with this, we capture the influence from previous tweets using the kernel over time.

1) Intensity Function: The proposed *Base Textual HP* model consider any representation of text in the base intensity and the base intensity is defined as follows :

$$\mu_{y,t} = \frac{\exp(W_y \times X_t)}{\sum_{i=1}^{|Y|} \exp(W_i \times X_t)} \tag{3}$$

Please note that the base intensity is no longer a constant and depends on the textual content of the post at time t. The base intensity is normalized across all labels to avoid it from having a dominating influence on the intensity function.

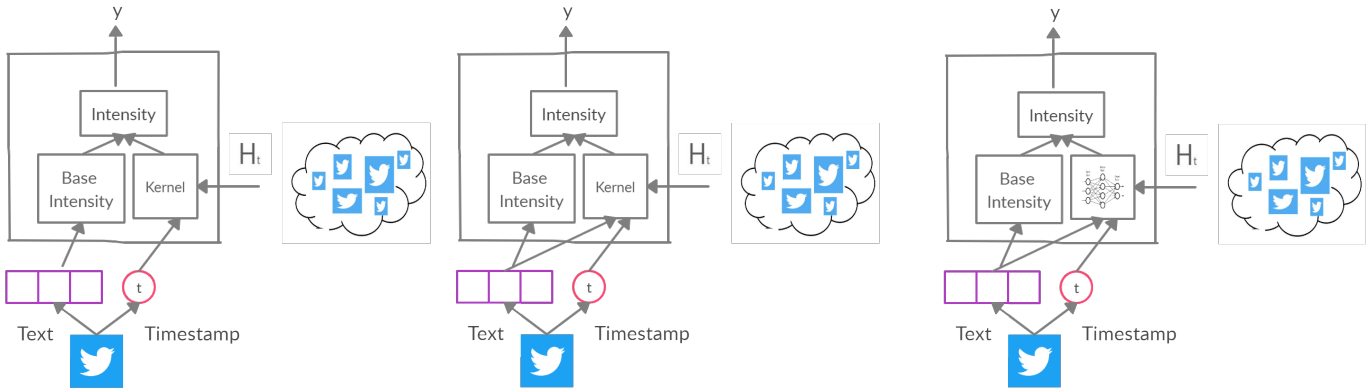
- X_t is a V-dimensional text representation of tweet at time t.
- W of size $|Y| \times |V|$ are the weights associated with the classes.

Using the proposed base intensity, we can write *Base Textual HP* intensity function for a stance label y as the sum of the base intensity and kernel function values over previous tweet times:

$$\lambda_{y,m}(t|H_{t^-}) = \mu_{y,t} + \sum_{t_\ell<t} \mathbf{I}_{m_\ell=m} \alpha_{y_\ell,y} \kappa(t - t_\ell) \tag{4}$$

Here, we consider the kernel $\kappa(t - t_\ell) = \omega \exp(-\omega(t - t_\ell))$, an exponentially decaying kernel capturing the influence of previous tweets. We can observe that base intensity will be higher for posts whose textual content is closely related to the stance label. Consequently it favours posts with this particular stance if the influence of past labels weighted by time are also favourable. In this way, we augment Hawkes process intensity with both time and text based information.

2) Likelihood function: The parameters of the proposed model (weight vector W and influence matrix α) can be learnt



(a) Textual HP Model (b) Fully Textual HP Model (c) Neural Kernel HP Model

Fig. 1: Framework for the proposed models. 1a) displays text being used as a part of base intensity 1b) displays text being used for base intensity as well as kernel 1c) displays the kernel is modeled as a neural network and text contributes to both base intensity and neural kernel function.

by maximizing the likelihood. The HP likelihood function is defined as

$$L(t, y, X, m) = \Big(\prod_{n=1}^{N} \lambda_{y_n, m_n}(t_n)\Big) \times \exp\Big(-\sum_{y=1}^{|Y|}\sum_{m=1}^{M}\int_0^T \lambda_{y,m}(s)\Big) \tag{5}$$

where the intensity function is defined as in (4). After expanding individual components of the above equation, we get log likelihood as

$$LL(t, y, X, m) = \sum_{n=1}^{N} \log \lambda_{y_n, m_n}(t_n) - C||W||^2 - |R|\sum_{n=1}^{N+1}(t_n - t_{n-1}) - \sum_{y=1}^{|Y|}\sum_{l=1}^{N}\alpha_{y_\ell, y}K(T - t_l) \tag{6}$$

where $K(T - t_l) = 1 - exp(-\omega(T - t_\ell))$ and $|R|$ represents number of topic (or rumour) categories. We add a regularization term over the weights of text with C as the regularization constant for better generalization of model. The parameters α and W are obtained by maximizing the log-likelihood function in (6). We find parameters using joint gradient based optimization over α and W, using partial derivatives of log-likelihood.

B. Fully Textual HP: Using text-based kernel

We propose a model (*Fully Textual HP*) where we use a text based kernel in combination with the temporal kernel in addition to text based base intensity. The text based kernel can help in representing the influence of past events/tweets based on their textual similarity.

1) Intensity Function: In this case, our model will consist of text-based base intensity, an exponentially decaying kernel to model time of tweets and a text-based kernel to consider text content of tweet as well. We can use different types of text kernels like the Gaussian kernel, or a polynomial kernel. Similar to Equation (4), we write the intensity function for the proposed model as:

$$\lambda_{y,m}(t) = \mu_{y,t} + \sum_{t_\ell < t} \mathbf{I}_{m_\ell = m}\alpha_{y_\ell, y}\kappa(t - t_\ell)\kappa(X_t, X_\ell)$$

The base intensity $\mu_{y,t}$ is the same as the one used in equation (3) while we use a Gaussian kernel over text to capture its influence

$$\kappa(X_t, X_\ell) = \exp\Big(-\frac{||X_t - X_\ell||^2}{2\sigma^2}\Big)$$

where σ is the hyper-parameter. When the textual contents of the post at time t (X_t) is similar to a past post text (X_ℓ) then $\kappa(X_t, X_\ell)$ will be higher and consequently the influence of the corresponding label will be higher. We define the likelihood for this model similar to Section V-A2.

VI. Neural Kernel Hawkes Process

A restriction with the previous approaches is that the past influence is specified through a predefined exponentially decaying kernel. Often these influences can take a form other than exponential decay. We intend to capture the functional form of the influence (kernel) through the proposed neural kernel Hawkes process. Here, we model kernels using a neural network which is theoretically capable of modeling any function (universal approximator). This allows us to learn the complex non-linear relationships between historical events and the current event. Proposed approach is different from previous works on HP for information diffusion [9], [10], where a recurrent neural network is used to model the full intensity function. Since we are only modeling kernels using neural networks, we continue to maintain the advantage of interpretability of Hawkes process, for e.g. label-label influences through the influence matrix. This model enables us to learn a more generalized version of Hawkes process keeping its causality property intact.

A. Intensity Function:

The intensity function for the neural Kernel Hawkes process is defined as :

$$\lambda_{y,m}(t) = \mu_{y,t} + \sum_{t_\ell < t} \mathbf{I}_{m_\ell = m} \alpha_{y_\ell, y} F([t_\ell, X_\ell], [t, X_t]; W_{nn})$$

where W_{nn} are the weights in the NN kernel, the text and time are input together and the base intensity $\mu_{y,t}$ is defined in (3).

All the parameters including the neural kernel parameters are learnt by maximizing the likelihood defined in (5). However, likelihood computation is challenging here due to intractable integral arising from the neural kernel. We approximate the intractable integral using the Monte Carlo approximation. It computes average intensity over uniformly sampled time and multiplies with time period to get the integral value. Backpropagation is applied on (5) after Monte Carlo approximation to learn parameters of the neural kernel. Prediction is done by evaluating the intensity function across all the classes at the time of the post and choosing the class with the highest intensity function.

VII. Experiments

A. Dataset

We use the PHEME dataset [21] for rumour stance classification. It considers tweets belonging to nine noteworthy events that occurred around the world. Along with tweets, it also considers retweets, and replies to form a tweet thread. Each thread contains a source tweet as well as replies to that tweet. Every tweet is assigned a stance from - *Supporting*, *Denying*, *Questioning*, *Commenting* classes w.r.t. the source tweet. The detailed statistics of the dataset is mentioned in Table (I). A characteristic of the dataset is that the distribution of categories is skewed which makes the task challenging. We use standard pre-processing of data and use 100-dimensional word2vec (Google News) representation of words and averaging over them to get tweet representation.

B. Baselines

We have considered following baselines:

- **Hawkes Process [22]** : The authors considered two approaches for learning parameters, gradient based (HP Grad) and closed form approximation (HP Approx). They have been shown to perform better than several machine learning models including conditional random fields.
- **LSTM [23]** : The authors have used the sequential structure of conversational threads using LSTM.

TABLE I: Statistics of the PHEME dataset, where S represents support class, D represents Deny class, Q represents Question class, and C represents Comment class. T is the total number of Tweets.

Dataset	S	D	Q	C	T
Ottawa Shoot	161	76	64	481	782
Ferguson Riots	161	82	94	680	1017
Prince in Toronto	19	7	11	59	96
Charlie Hebdo	236	56	51	710	1053
Ebola Essien	6	6	1	21	34
Germanwings crash	177	12	28	169	386
Putin missing	17	7	5	33	62
Sydney Siege	89	223	99	713	1124

C. Experimental Setup

The experiments are built in a way where we depict real world scenarios as closely as possible. In the real world, new rumours arise on a regular basis. We try to perform something similar in our experiments where we train our model on old rumours and then use them for stance classification on new rumours. The experimental setup can be categorized into following two types.

1) ***Leave one out - Thread****:* Following prior work [38], we consider 4 events - Ottawa, Ferguson, Charlie Hebdo and Sydney Siege, the largest events from PHEME (each with approximately 1000 tweets per event). Every event in the data set has multiple tweet threads $(50 - 70)$, where each thread is a new rumour generated when the event occurred. We train on $n-1$ rumour threads and test on the n^{th} rumour. We perform this n times, testing on a different rumour each time. This helps in getting the overall performance across all rumours.

2) ***Leave one out - Event****:* Here, a dataset of top 8 events is considered and then combined to form a bigger data set, with 4554 tweets in total. We consider training on 7 events at a time and testing on the 8^{th} one. This is repeated 8 times, and an average score is reported.

D. Evaluation Metrics

We use the popular metrics for multi-class classification i.e. accuracy and F1-score. We consider micro-averaged accuracy and macro averaged F1-score as reported in the previous work.

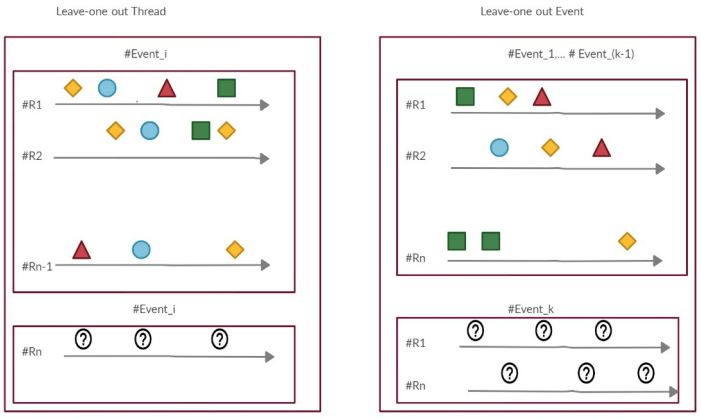


Fig. 2: Experimental set-up (left) shows Leave one out - Thread and (right) shows Leave one out - Event

TABLE II: Results from the baselines (bottom rows) and our proposed approaches (top rows). First four results columns are from individual four rumor datasets.

	Ottawa		Ferguson		Charlie Hebdo		Sydney Siege	
	Acc.(%)	F1	Acc.(%)	F1	Acc.(%)	F1	Acc.(%)	F1
Base Textual HP	**70.2**	0.312	**71.12**	0.329	69.5	0.324	**71.63**	0.324
Fully textual HP	62.4	0.329	62.34	0.259	69.04	0.304	64.32	0.318
Neural Kernel HP	56.01	0.153	66.7	0.193	59.18	0.169	56.32	0.168
HP Grad [22]	63.43	0.424	63.23	0.331	71.79	0.419	62.99	0.395
HP Approx. [22]	67.77	0.32	68.44	0.26	**72.93**	0.325	68.59	0.349
LSTM on HPfeatures [23]	66.67	**0.487**	69.73	**0.409**	70.99	**0.513**	69.51	**0.496**

Considering number of stances to be K, the formulae for macro F1-score can be written as follows -

$$Macro\text{-}Precision = \frac{\sum_{i=1}^{K} Precision}{K}$$

$$Macro\text{-}Recall = \frac{\sum_{i=1}^{K} Recall}{K}$$

$$Macro\text{-}F1Score = \frac{2 * Macro\text{-}Precision * Macro\text{-}Recall}{Macro\text{-}Precision + Macro\text{-}Recall}$$

VIII. Results and Analysis

A. Results

The proposed approach is compared against the Hawkes process [22] and LSTM [23] based approaches for rumour stance classification. The *Base Textual HP* which used discriminative modeling of text with normalized base intensity outperforms the benchmarks for all events except for Charlie Hebdo in terms of micro-accuracy, demonstrating that incorporation of textual features as part of intensity function helps to improve results. In comparison with the LSTM approach [23] for this setup, we can see that our *Base Textual HP* model gives better accuracy in all datasets except Charlie Hebdo. This shows usefulness of HP based models over modern neural networks especially when dataset size is small. The *Fully textual HP* which uses text in base intensity as well as kernel, gives comparable results to benchmark models, but doesn't outperform *Base Textual HP*. This means that influence arising through text similarity is not very useful for predictions at thread level, with typical thread size being 10. Here, dissimilar text belonging to different classes (e.g. deny and support tweets) can have higher influence, which is restricted through the text kernel. Although, this shows another successful way of augmenting Hawkes process with text. We also observe that *Neural Kernel HP* did not give good performance. In Figure (5) we show an example function learned by the neural kernel against text similarity and in general, and we find a decrease w.r.t cosine similarity. This supports the observations from fully textual HP. However, Neural kernel HP did not perform well overall, presumably due to the small sized rumour stance data (1000 tweets per event).

We can see the results for Leave one out - event approach explained in Section VII-C2 in Table III. The *Fully Textual HP* gives the best results beating the benchmarks, showing the importance of considering text similarities between posts, as opposed to only similarities between categories from [22]. The *Neural Kernel HP* model performs better in this setup, however is still limited by the small data set size. On the other hand, using the inductive bias of Hawkes process assumption helps perform better under this data scarce scenario.

TABLE III: Result comparison in Leave one out - Event setup

	Accuracy (%)	Macro F1
Base Textual HP	64.70	0.269
Fully textual HP	**69.10**	**0.329**
Neural kernel HP	59.44	0.233
HP Grad.	-	0.309
HP Approx.	-	0.307
LSTM on HPFeatures	-	0.318

B. Analysis

1) Analysis of Influence Matrix α: We analyze the values learnt by the influence matrix α. It is a 4×4 dimensional matrix which learns the influence of different classes of tweets on others. For example, it learns the impact of a previous tweet being of class *Support* on the next tweet being of class *Deny*. In Figure 3, we see sample values of the influence matrix belonging to the Fully Textual HP model. The bold face values show the best result in a row while italics show the second best. An interesting observation is that the *Deny* class has the highest influence on the *Question* and *Deny* classes. This means that a *Deny* tweet is often followed by a *Question* tweet or a *Deny* tweet, which makes sense in rumour stance classification. The diagonal values are relatively high which means that each class influences the next tweet to belong to that class, i.e. *Support* or *Question* is likely to attract more *Support* or *Question* tweets respectively than tweets belonging to other classes. The values in the last column are also usually quite high in the row. The last column belongs to *Comment* class. This tells that it is very likely for a *Comment* tweet to follow tweets belonging to other classes. This also is quite expected as the data set has a class imbalance with more than 60% tweets belonging to *Comment* class.

2) Intensity plots: Tweets are associated with posting times, and in Figure(4) we plot the intensity value of tweets at their posting times for support and deny classes. The intensity function values are obtained by considering temporal and textual information as discussed in Equation (2). In Figure (4), we find that intensity value is higher for the tweets of respective

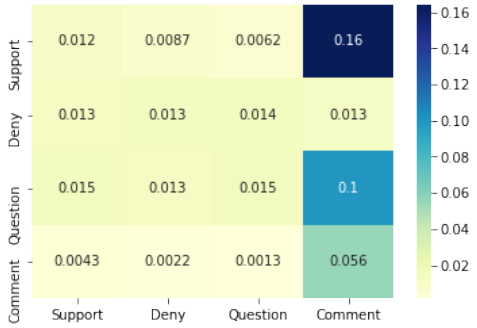


Fig. 3: Influence matrix of Fully Textual HP model

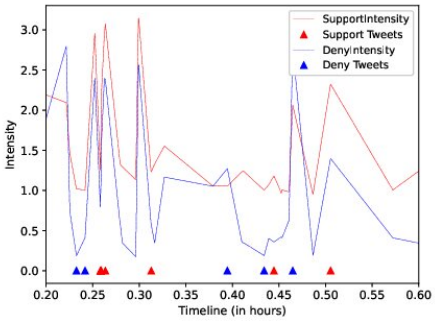


Fig. 4: A snapshot of posts and intensities for 2 classes, Support and Deny, at post times using *Base Textual HP* on Sydney Siege data. The intensity for a class becomes higher as some tweet occurs from that class through textual features and temporal influences

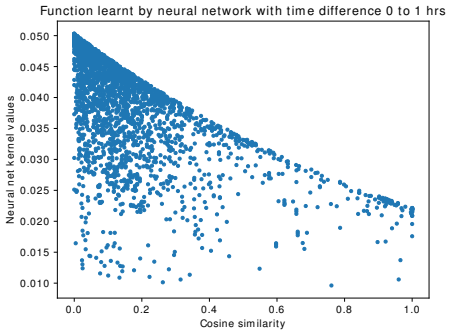


Fig. 5: Neural kernel values for text features of an input pair of tweets vs. cosine similarity between these text features. Dots accumulated over the time difference range 0-1 hours.

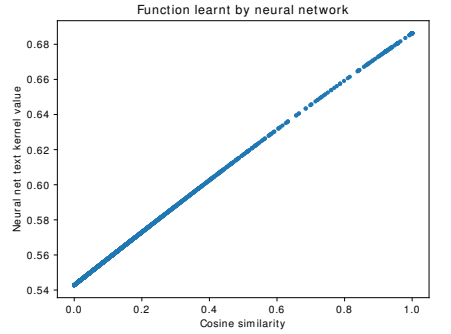


Fig. 6: Function learnt by independent text neural network when we consider neural network with two kernels

classes. Figure(5) shows the kernel function learnt against cosine similarity of text for Neural Kernel HP Model. The pairs of tweets are selected such that they belong to the same thread. We compute the cosine similarity (using text) and the neural kernel values (using text and time) between them. The difference in their time of occurrence ranges between 0-1hrs. Hence there can be multiple pairs with the same cosine similarity. We can observe here that in general neural kernel value is decreasing with increasing cosine similarity.

IX. Ablation Study

We carry out an ablation study on the Neural Hawkes Process model by using text and time based kernels in different ways. Thereby, we learn the nature of the function learnt by neural networks. Figure 7 shows the time kernel learnt when we use input just the difference of time for the neural network. We can observe that it learns a function similar to exponentially decaying kernel, which explains the relevance of exponential kernel for such settings. In another variation, we use two separate kernels for time and text. Figure 5 displays the function learnt for text under such settings. The value of function learnt increases with increasing similarity of text. Figure 6 displays the function learnt by independent text neural network while we consider separae kernels for text and time. However, results of the variants were not as good as the proposed model.

X. Conclusion

We proposed a novel method for text classification based on Hawkes processes, where we consider textual features in the intensity function and overcome limitations of the existing HP models. We propose using kernels (exponential and neural network) over text and time, providing more flexible approaches to model influences between posts. This enables us

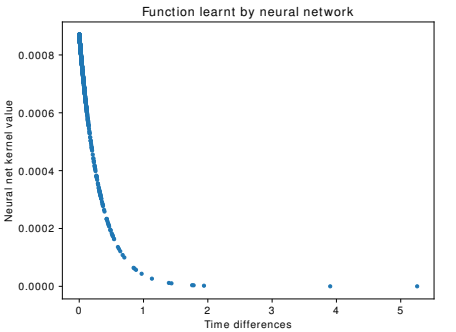


Fig. 7: Function learnt for time by neural network kernel

to capture the influence among tweets not only using time but also using textual content of tweets. Neural kernel can learn the functional form of the influence rather than predefining it as an exponential function. The proposed models also allow one to easily consider pre-trained word embeddings. The experiments on rumour stance classification showed the effectiveness of the proposed approaches, with discriminative textual HP models performing better than generative textual HP models. We also show that traditional HP based approaches can perform better than neural network based HP on smaller datasets due to its inductive bias. The proposed approach for time and text sensitive sequence classification is generic and can be used for other tweet classification tasks and in domains where text and time co-exists.

REFERENCES

[1] Hawkes, Alan G., and David Oakes. "A cluster process representation of a self-exciting process." Journal of Applied Probability 11.3 (1974): 493-503.

[2] Ozaki, Tohru. "Maximum likelihood estimation of Hawkes' self-exciting point processes." Annals of the Institute of Statistical Mathematics 31.1 (1979): 145-155.

[3] Liniger, Thomas Josef. "Multivariate hawkes processes." PhD diss., ETH Zurich, 2009.

[4] Dassios, Angelos, and Hongbiao Zhao. "Exact simulation of Hawkes process with exponentially decaying intensity." Electronic Communications in Probability 18 (2013).

[5] Hawkes, Alan G. "Spectra of some self-exciting and mutually exciting point processes." Biometrika 58.1 (1971): 83-90.

[6] Embrechts, Paul, Thomas Liniger, and Lu Lin. "Multivariate Hawkes processes: an application to financial data." Journal of Applied Probability 48.A (2011): 367-378.

[7] Rizoiu, Marian-Andrei, et al. "A tutorial on hawkes processes for events in social media." arXiv preprint arXiv:1708.06401 (2017).

[8] Laub, Patrick J., Thomas Taimre, and Philip K. Pollett. "Hawkes processes." arXiv preprint arXiv:1507.02822 (2015).

[9] Du, Nan, et al. "Recurrent marked temporal point processes: Embedding event history to vector." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

[10] Xiao, Shuai, et al. "Modeling the intensity function of point process via recurrent neural networks." Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[11] Mei, Hongyuan, and Jason M. Eisner. "The neural hawkes process: A neurally self-modulating multivariate point process." Advances in Neural Information Processing Systems. 2017.

[12] Upadhyay, Utkarsh, Abir De, and Manuel Gomez Rodriguez. "Deep reinforcement learning of marked temporal point processes." Advances in Neural Information Processing Systems. 2018.

[13] Xiao, Shuai, et al. "Wasserstein learning of deep generative point process models." Advances in Neural Information Processing Systems. 2017.

[14] Lukasik, Michal, et al. "Using Gaussian processes for rumour stance classification in social media." arXiv preprint arXiv:1609.01962 (2016).

[15] Santosh, T. Y. S. S., Srijan Bansal, and Avirup Saha. "Can Siamese Networks help in stance detection?." Proceedings of the ACM India Joint International Conference on Data Science and Management of Data. 2019.

[16] Zubiaga, Arkaitz, et al. "Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations." arXiv preprint arXiv:1609.09028 (2016).

[17] Kochkina, Elena, Maria Liakata, and Isabelle Augenstein. "Turing at semeval-2017 task 8: Sequential approach to rumour stance classification with branch-lstm." arXiv preprint arXiv:1704.07221 (2017).

[18] Aker, Ahmet, et al. "Stance classification in out-of-domain rumours: A case study around mental health disorders." International Conference on Social Informatics. Springer, Cham, 2017.

[19] Aker, Ahmet, Leon Derczynski, and Kalina Bontcheva. "Simple open stance classification for rumour analysis." arXiv preprint arXiv:1708.05286 (2017).

[20] Ferreira, William, and Andreas Vlachos. "Emergent: a novel data-set for stance classification." Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies. 2016.

[21] Zubiaga, Arkaitz, et al. "Analysing how people orient to and spread rumours in social media by looking at conversational threads." PloS one 11.3 (2016).

[22] Lukasik, Michal, et al. "Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2016.

[23] Zubiaga, Arkaitz, et al. "Discourse-aware rumour stance classification in social media using sequential classifiers." Information Processing & Management 54.2 (2018): 273-290.

[24] Lukasik, Michal, Trevor Cohn, and Kalina Bontcheva. "Classifying tweet level judgements of rumours in social media." arXiv preprint arXiv:1506.00468 (2015).

[25] Lukasik, Michal, Trevor Cohn, and Kalina Bontcheva. "Point process modelling of rumour dynamics in social media." Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2015.

[26] Kochkina, Elena, Maria Liakata, and Arkaitz Zubiaga. "All-in-one: Multi-task learning for rumour verification." arXiv preprint arXiv:1806.03713 (2018).

[27] Veyseh, Amir Pouran Ben, et al. "A temporal attentional model for rumor stance classification." Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017.

[28] Lukasik, Michal. "Probabilistic Modeling of Rumour Stance and Popularity in Social Media." PhD diss., University of Sheffield, 2017.

[29] Jebara, Tony, and Alex Pentland. "Maximum conditional likelihood via bound maximization and the CEM algorithm." Advances in neural information processing systems. 1999.

[30] Dutta, H.S., Dutta, V.R., Adhikary, A. and Chakraborty, T., 2020. HawkesEye: Detecting fake retweeters using Hawkes process and topic modeling. IEEE Transactions on Information Forensics and Security, 15, pp.2667-2678.

[31] Srijith, P.K., Lukasik, M., Bontcheva, K. and Cohn, T., 2017, July. Longitudinal modeling of social media with Hawkes process based on users and networks. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (pp. 195-202).

[32] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.

[33] Pennington, J., Socher, R. and Manning, C.D., 2014, October. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[34] He, X., Rekatsinas, T., Foulds, J., Getoor, L. and Liu, Y., 2015, June. Hawkestopic: A joint model for network inference and topic modeling from text-based cascades. In International conference on machine learning (pp. 871-880). PMLR.

[35] Choudhari, J., Dasgupta, A., Bhattacharya, I. and Bedathur, S., 2018, November. Discovering topical interactions in text-based cascades using hidden markov hawkes processes. In 2018 IEEE International Conference on Data Mining (ICDM) (pp. 923-928). IEEE.

[36] Du, N., Farajtabar, M., Ahmed, A., Smola, A.J. and Song, L., 2015, August. Dirichlet-hawkes processes with applications to clustering continuous-time document streams. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 219-228).

[37] Sha, H., Al Hasan, M., Mohler, G. and Brantingham, P.J., 2020. Dynamic topic modeling of the COVID-19 Twitter narrative among US governors and cabinet executives.

[38] Lukasik, M., Bontcheva, K., Cohn, T., Zubiaga, A., Liakata, M. and Procter, R., 2016. Using Gaussian Processes for Rumour Stance Classification in Social Media.

[39] Qazvinian, V., Rosengren, E., Radev, D. and Mei, Q., 2011, July. Rumor has it: Identifying misinformation in microblogs. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (pp. 1589-1599).