# Hot Topic-Aware Retweet Prediction with Masked Self-attentive Model

Renfeng Ma*, Xiangkun Hu*, Qi Zhang†, Xuanjing Huang, Yu-Gang Jiang
School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing
Fudan University, Shanghai, P.R.China 201203
{rfma17,xkhu17,qz,xjhuang,ygj}@fudan.edu.cn

## ABSTRACT

Social media users create millions of microblog entries on various topics each day. Retweet behaviour play a crucial role in spreading topics on social media. Retweet prediction task has received considerable attention in recent years. The majority of existing retweet prediction methods are focus on modeling user preference by utilizing various information, such as user profiles, user post history, user following relationships, etc. Yet, the users exposures towards real-time posting from their followees contribute significantly to making retweet predictions, considering that the users may participate into the hot topics discussed by their followees rather than be limited to their previous interests. To make efficient use of hot topics, we propose a novel masked self-attentive model to perform the retweet prediction task by perceiving the hot topics discussed by the users' followees. We incorporate the posting histories of users with external memory and utilize a hierarchical attention mechanism to construct the users' interests. Hence, our model can be jointly hot-topic aware and user interests aware to make a final prediction. Experimental results on a dataset collected from Twitter demonstrated that the proposed method can achieve better performance than state-of-the-art methods.

## CCS CONCEPTS

• **Computing methodologies** → *Natural language processing*; • **Information systems** → Social recommendation; Semantic web description languages;

## KEYWORDS

Retweet prediction; Hot topic; Social Medias

**ACM Reference Format:**
Renfeng Ma*, Xiangkun Hu*, Qi Zhang†, Xuanjing Huang, Yu-Gang Jiang. 2019. Hot Topic-Aware Retweet Prediction with Masked Self-attentive Model. In *42nd Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3331184.3331236

*Both authors contributed equally to this research.
†Corresponding author.


*SIGIR '19, July 21–25, 2019, Paris, France*

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331236

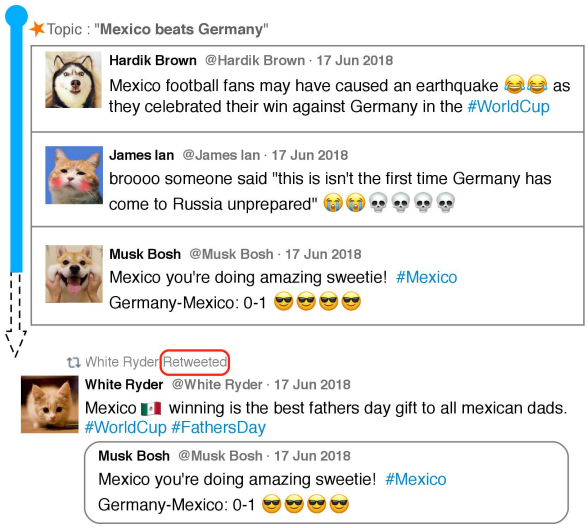


**Figure 1: Example of the influence of a user's social exposures covering hot topics. When Mexico stunned defending champion Germany for a brilliant win at the 2018 FIFA World Cup, Mexicans who knew nothing about football would also joined in the celebration. Similarly, users would join in the topics discussed in the users' social exposures from their followees, which are irrelevant to users' interests.**

## 1 INTRODUCTION

With the boom of social media platforms (e.g., Instagram, Facebook and Twitter), it has become incredibley convenient for users to focus on worldwide hot topics and share their personal insights and opinions. Anyone who wants to express personal views on some events or promote a product can post a tweet on a Twitter-like social media site. Nevertheless, retweeting is the most straightforward and crucial way to spread information or participate in topic discussions. According to a recommendation from a multimedia journalism professor, retweeting is a great way to express your thoughts and good retweeting really does involve adding a few comments of your own to the original post[1]. Hence, retweet prediction plays a important role in various application scenarios, such as stock prediction [4, 52], opinion mining [3, 26], public health analysis [31, 43], real-time event detection [34], etc.

Currently, research on retweet prediction in social media takes primarily two kinds of information. In the first part, some methods

[1]https://www.lifewire.com/retweet-with-a-comment-on-twitter-2655355

construct the prediction models through the social network maps, which may contain the probability of information dissemination. [30] built an information propagation tree, while other researchers have considered various social relationships [27, 32, 47]. Second, some contextual features are incorporated to perform this task [10, 37], such as content features, hashtags, and users' profiles. Recently, some other methods have attempted to construct users' interests from their posting histories [2, 11, 49, 50]. With the volume of user-generated image tweets growing tremendously, the multimodal content information is the latest to be introduced to deal with retweet prediction. [53] proposed a multimodal model to utilize image tweets to deal with this problem. In addition to feature engineering-based machine learning models, [50] proposed an attention-based deep neural network to calculate the similarity between the content of the tweet and the user's interests. The attention-based convolutional neural network has achieved better performances than other kinds of methods.

Although a variety of studies have been conducted on the task of automatically predicting retweet behaviour in social networks, the majority of these methods focus on modelling user preference by utilizing various information, such as user profiles, user post history, user's following relationships, etc. Considering the convenience of sharing real-time information on social media sites, it is obvious the hot topics discussed in a user's social exposures would affect the user's retweet behaviour, interpreted as participating into those topics. Figure 1 shows an example of the influence of an unexpected win at the 2018 FIFA World Cup. After Mexico beat the defending champion Germany and became the dark horse, Mexicans who knew nothing about football would also joined in the celebration.

To overcome the above difficulties in retweet prediction tasks, we propose a novel masked self-attentive model to incorporate the hot topics discussed by the users' followees. Some previous works have been successfully to utilized revised self-attentive mechanisms to capture the key information among the contextual information. For example, [54] utilized a masked transformer to help a video caption model to focus on key events. [24] proposed a gated self-attentive mechanism to utilize intent detection to improve slot filling task. In this paper, a novel masked self-attentive model is applied to the user social exporsure. After treating the recent posts of a user's followees as an interrelated module, we firstly utilized self-attention to construct the context-aware representation. Further, we utilized the context-aware representation to generate a mask, and each channel of the mask represents the relevance of a common topic. More specifically, a hierarchical attention mechanism is utilized to model the users' interests. In a word, our model take the content of a tweet, history of its author, history interest of the candidate retweet user, and the social exposures of the candidate retweet user into consideration, simultaneously. Hence, our model would make a better retweet prediction as considering both users' interests and the hot topic discussed among the users' followees.

To demonstrate the effectiveness of our model, we performed experiments on a large data set collected from Twitter. Experimental results showed that the proposed method could achieve better performance than state-of-the-art methods ignoring potential hot topics in the user's social exposures. The main contributions of our work can be summarized as follows.

- We introduced an integrated framework to not only consider users' interest similarity, but also the hot topics discussed in users' social exposures to perform retweet prediction task.
- We proposed a novel masked self-attentive network that can perceive the hot topics discussed among the users exposures towards real-time posting from their followees.
- Experimental results using a dataset constructed by us from Twitter demonstrated that our model could achieve significantly better performance than current state-of-the-art methods.

## 2 RELATED WORK

### 2.1 Retweet Prediction and Social Media Recommendation

With the continuous development of social media, there are dramaticly increasing requirements coming out. Vary tasks have been proposed for different issues on social media, such as content recommendation [6, 21], community recommendation [51], tag recommendation [9, 12, 22], music recommendation [35], mention recommendation [13, 16, 40], stock prediction [4, 52], opinion mining [3, 26], public health analysis [31, 43], real-time event detection [34] and so on.

There is a broad spectrum of retweet-related research includes prediction of retweets, retweeters, retweet counts, information spread flow of tweets as well as tweet recommendation. Some works try to explore, analyze, and predict user's retweet behaviour, some are focus on finding out potential retweeters. These methods can be categorized into three versions as follows: 1. Which tweet will be retweeted by the user? 2. Who will retweet the target tweet? 3.Why do some tweets get more retweets?

For the first kind of research area, [7]incorporated some important features such as user's prior interaction with author, author's tweeting rate, content of tweet on user's retweeting behavior. In the second sub-category, the goal is to not only analyze which feature may influence user's retweeting behavior but also make retweet prediction models based on their investigated information. Early works such as [5, 37] studied a wide range of features that could affect the retweetability of a tweet. [2, 10] used probabilistic models to predict the retweeting probability based on the user retweet behavior and user interest. [32] explored content influence, network influence, and temporal decay factor on users' retweeting decision and proposed Conditional Random Field (CRF) based retweet prediction model using features that define tweet's content influence, user's network influence, and temporal influence on user's retweet decision. [45] analyzed different features to develop retweet prediction model from the perspective of individual users. [46] incorporated the social influence and breaking news on user's retweeting behavior and incorporated these influences in their proposed mixture latent topic retweet prediction model. [47, 48] studied the influence of users' social information on their retweeting behavior. Recent works [41, 50] applied deep neural network (DNN) architectures to this task to perform feature extraction, and realized significant performance improvements. In the last sub-category, the goal is to judge the influence of information flow by retweet behaviour. [18] and [42] proposed matrix factorization retweet prediction models.

**Table 1: Statistics of our dataset.**

| | |
|---|---|
| # Tweets | 4,903,398 |
| # Users | 80,575 |
| # Positive Data | 18,937 |
| # Negative Data | 94,495 |

In this work, we mainly study the second sub-category as making a prediction about whether the user will retweet the query tweet. [50] was the state-of-the-art approach performing retweet prediction task as the same way with us. We not only incorporate the posting history of the users and the authors, but also take into consideration of the hot topics discussed in user's social exposures.

## 2.2 Attention mechanism and Memory Network

Attention mechanisms allow models to focus on necessary parts of inputs at each step of a task. And the idea is come from visual attention mechanism found in humans. Human visual attention suggests that our brain usually focuses on selective parts of the whole perception regions according to demand. Futhermore, attention mechanism has been proved to be significantly effective in both visual related tasks and natural language processing tasks, such as machine translation [1], question answering [36, 44], image classification [29], etc. Particularly, the self-attention [39] based model BERT [8] has achieved great success in many NLP tasks. Its effectiveness results from the assumption that human recognition does not tend to process whole texts or images in their entirety. In other words, humans attempt to focus on the important parts of the whole perception regions according to demand. Hence, the attention mechanism is proposed to extract important information from the inputs space, which can provide the model to focus more on processing the important information and achieve a better performance on tasks. And the self-attention mechanism is proposed to address long-range dependencies challenge, and construct a better context-aware representation. Therefore, we make a improved version based on self-attention mechanism to construct the hot topic aware information among the social exposure for users.

Recently, variants of Memory Networks have been proposed to deal with various NLP tasks. Especially for question answering task, [38] proposed a end-to-end memory network with a recurrent attention model over a possibly large external memory, [23] proposed a dynamic memory network to treat various NLP tasks as question answering problem. [25] incorporate the reinforce learning to construct dynamic memory and utilize the dynamic memory to deal with question answering task. In this work, we combine the hierarchical attention with end-to-end memory network to model the interest similarity between user and author.

# 3 PRELIMINARY

## 3.1 Dataset Construction

We constructed a large dataset to evaluate our proposed model. We crawled a large amount of Twitter data published before June 5, 2017 from Twitter API [2]. First, we randomly selected 1,500 users, who have published more than 1,000 tweets and have more than 20 followees, as the starting users. Then we crawled all the tweets posted by them and their followees. In this step, we crawled totally 411,054 users and 36,807,681 tweets. Second, we removed non-English Tweets, cleaned the remaining tweets by removing URLs and special characters; then we removed the tweets less than 5 words. After these removal, we checked the users iteratively, making sure that each user had at least 50 tweets, more than 20 followees, and retweet data.

We sorted the tweets of every user by time. Using the follow-fans information, we then reverted the home timeline of each user. We used the retweets by the user in the timeline as his/her positive data; used 30 tweets immediately before the retweet as timeline data; and randomly chose 30 tweets posted before the retweet by the user and the author separately, as their posting history. For each retweet, we randomly chose 5 tweets in the timeline that not reweeted by the user as his/her negative data, and constructed the timeline and posting histories as before. In total, we got 18,937 positive data and 94,495 negative data. The detailed statistics of the dataset are listed in Table 1.

## 3.2 Data Analysis

Generally speaking, social media users usually participate in the discussion of tweets related to their interests. Particularly, [17] proved that social media user' interest is concentrated in a limited range and tends to remain stability over a period of time. Furthermore, the topics that a user wants to join will be similar to the tweets he or she has posted. However, human beings are essentially tend to join in social groups and full of exploring spirit. There is no wonder that both Twitter and Facebook provide trending lists (e.g., "in case you missed it" in Twitter and "top stories" in Facebook) to help users quickly find popular topics and improve the diversity of recommended content.

**Hypothesis :** *User will also take retweet behaviour and participate in the hot topics discussed by his or her followees, which is unrelated to user's history interest.*

To verify the hypothesis, we analyze the difference among user posting history, user retweet history and real-time posting of corresponding user's followees. In order to figure out interests among these tweet sets, we utilize Latent Dirichlet Allocation with collapsed Gibbs sampling [14]. With the help of LDA, each document (here, tweet collection) may be viewed as a mixture of various words that attempt to co-occur in similar documents. Each set of co-occur words is named as "topics" in LDA. In this experiment, we utilize these topics to represent user's posting interest, user's retweet interest and social interest of user's social exposures.

Specifically, for each user, we randomly selected 100 posted tweets as the user's posting documents, 100 tweets from retweeted tweets as the the user's retweet documents and sampled 100 tweets from the the real-time post from user's followees as the user's social exposure documents (both retweet or not retweet may contain in the user's social exposure documents). We set the number of LDA topics to 50 and utilized Kullback-Leibler (KL) divergence to measure the differences between these three distributions. Based on statistics, we found that tweet the user will retweet have similar interests with the user's posts. Because the KL-divergence between

[2] https://developer.twitter.com

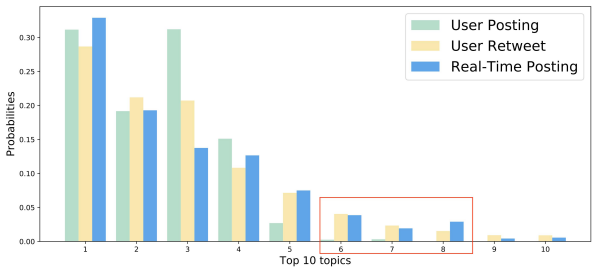


**Figure 2: An example of LDA topic distributions of among user postings, user retweets and user's social exposures. The number of LDA topics is set to 50, and the probabilities of top 10 topics are shown in the figure. Topics that a user wants to join have a distribution similar to the tweet sets he or she posted, whereas there still remain some retweets are unrelated to the user's posting history. Accordingly, taking the social exposures of users into consideration is a great choice to fix such issue.**

these two distributions is below than threshold for more than 89% of users. However, there are also some topics that are irrelevant to user's post history. Just like the example shown in Figure 2, the probabilities of top 10 topics are shown in the figure. Topics that a user wants to join have a similar distribution to the tweet sets he or she posted, whereas there still remain some top topics are out of the rule. In the other words, taking the social exposures of users into consideration is a great choice to meet the challenge. Certainly, according to our statistics, we discovered that more than 43% of users meet the following conditions on at least one top topic,

1). $0.8 \leq P_{retweet}(topic = i)/P_{social}(topic = i) \leq 1.25$,

2). $P_{retweet}(topic = i)/P_{post}(topic = i) \geq 50$,

where $P_{retweet}(topic = i)$ denotes the retweet probability at topic $i$, and the same to $P_{social}(topic = i), P_{post}(topic = i)$.

From the above analysis, we can conclude that user retweet behaviours are sometimes influenced by the topics discussed among his or her followees. Thus, compared with only construct interest similarity between query tweet and users' posting history, it is feasible to incorporate users' social exposures as a crucial complementary part to perform retweet prediction.

## 4 APPROACH

**Problem Definition and Notation**

First of all, we describe some preliminary of our approach, such as the problem definition and corresponding notation. In this work, given a query tweet $t_q$ and a user $u$, our task is to predict whether the user $u$ would retweet the tweet $t_q$ based on three parts of information. And the details of these external information are described as follows: 1. the posting history of the author of the query tweet $H_a = \{t_{a_1}, t_{a_2}, \cdots, t_{a_N}\}$, which can represent author's interests, 2. the posting history of the user $H_u = \{t_{u_1}, t_{u_2}, \cdots, t_{u_N}\}$, which can represent user's interests, 3. the real-time tweets posted by followees of the user $H_l = \{t_{l_1}, t_{l_2}, \cdots, t_{l_N}\}$, which may contain hot topics unfamiliar with user' interests and author's interests. (And $N$ denotes the amount of tweets for each collection. )

The social exposures of each user can be viewed as an information flow within a small community, and may contain common topics due to potential common interests among the small community. Furthermore, one user's social exposures are composed of the real-time posting from different followees, and the information flow does not have time sequence consistency. Hence, the self-attention mechanism with a maximum length of 1 is a better choice to construct the context aware representation than Recurrent Neural Network. (The shorter these paths between any combination of positions in the input and output sequences, the easier it is to learn long-range dependencies [15, 39].) Particularly, in order to filter out hot topic among the user's social exposures, we incorporate two parallel Transformer encoder and the final layer of first Transformer encoder is utilized to generate a mask vector, where the value in each dimension of the mask vector represents the similarity for corresponding tweet. By taking element-wise multiplication of mask vector and each stack layer of the second Transformer encoder, unrelated information will be filtered after multiplying a small value many times. To model the history interest similarity between user and author, the query tweet $t_q$ is incorporated to make a hierarchical mechanism over the author posting history $H_a$ and user posting history $H_u$. Based on the query tweet $t_q$, author posting history $H_a$, user posting history $H_u$, and user's social exposures $H_l$, the output layer can represent the probability of retweeting behaviour. The overall architecture of the model is illustrated in Figure 3, and the masked self-attentive mechanism is the right part in Figure 3.

In this section, we will introduce the basic framework of our approach in detail. Firstly, we utilize bi-directional long short-term memory networks (Bi-LSTM) to encode each tweet. To meet the need of different levels of semantic representation, we maintain all word-level hidden vectors for each tweet in author posting history $H_a$ and user posting history $H_u$, but utilize the last hidden vector for query tweet $t_q$ and each tweet in user's social exposures $H_l$. After performing multi-hop hierarchical attention over author posting history $H_a$ and user posting history $H_u$, we can formulate the history interest similarity vector $O^H$. Meanwhile, the hot-topic aware social exposure vector $O^l$ is constructed by our proposed masked self-attentive mechanism. Then a concatenation layer is applied to combine the history interest similarity $O^H$ and hot-topic aware social exposures $O^l$. Finally, the retweet behaviour is predicted by a fully connected softmax layer. We describe our models in four parts. The tweet feature representation is described in Section 4.1. The masked self-attentive mechanism and hierarchical attention mechanism are described in Section 4.2 and Section 4.3, respectively. The last Section 4.4 is the description of final prediction.

### 4.1 Tweet Feature Representation

At the beginning, we convert each word $word_i$ in a given tweet $t$ to a one-hot vector in the size of the vocabulary. Then, we utilize a simple embedding layer to encode each one-hot vector to a word vector $w_i$ distributed in a continuous space: $w_i = Mword_i$ (here, we utilize pre-trained Twitter glove embedding [33]). The size of the embedding layer is $d \times |V|$, where $d$ is the embedding dimension

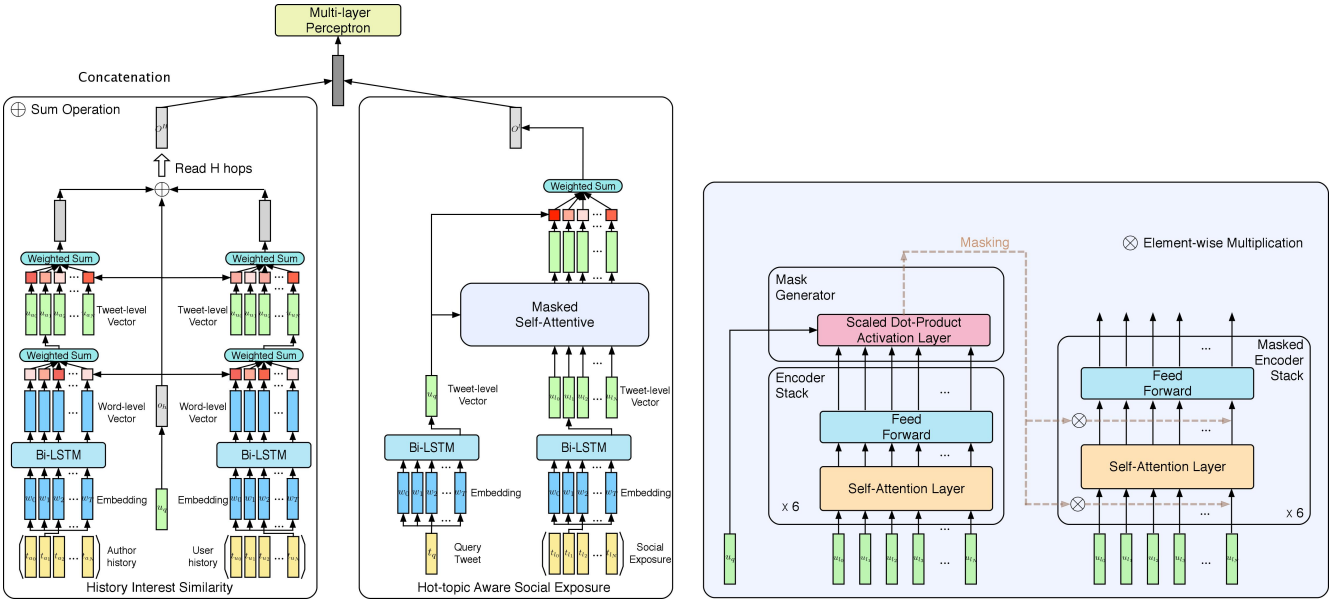


**Figure 3: Overall architecture of our proposed AUT-MSAM: (1) Modeling of History Interests Similarity between Author and User, (2) Modeling of Hot-topic Aware User's Social Exposures. Here, we denote $u_q$ as the representation of query tweet. For the first componet, we utilize $u_q$ as the initial representation for $o^h$ to query author posting history $H_a$ and user posting history $H_u$. And the hierarchical attention constructing procesure will stack $H$ times to formulate the final represenation $o^H$, which represents history interests similarity between user and author. For the second component, $u_q$ is incorporated into the masked self-attentive mechanism, and final representation $o^l$ of social exposure can aware whether the query tweet meets the hot-topic in social exposure. These two components are joint to make a final prediction.**

and $|V|$ is the size of the vocabulary. Hence, we get a word-level tweet feature representation: $t = \{w_1, w_2, \cdots, w_T\}$, where $T$ is the maximum tweet length. More specifically, each sentence with length less than $T$ is padded with zero vectors.

In view of that tweets are limited to 140 characters in length and tend to be short in content. Hence, we utilize the bidirectional LSTM to construct a sentence-level tweet features representation. At each time step, the bidirectional LSTM unit takes the word embedding vector $w_t$ as an input vector and outputs a hidden state $h_t$. The details are illustrated as follows:

$$h_t{}^{(f)} = LSTM^{(f)}(w_t, h_{t-1}{}^{(f)}), \tag{1}$$

$$h_t{}^{(b)} = LSTM^{(b)}(w_t, h_{t+1}{}^{(b)}), \tag{2}$$

where $h_t{}^{(f)}$ and $h_t{}^{(b)}$ represent the hidden states at time step $t$ from the forward and backward LSTMs, respectively. Finally, we construct a set of text feature vectors $u_T = \{u_1, u_2, \cdots, u_T\}$ by concatenating the two hidden state vectors at each time step:

$$u_t = [h_t{}^{(f)} : h_t{}^{(b)}], \tag{3}$$

where $u_t$ is the representation vector of the $t$-th word in the context of the entire sentence. Specifically, the word embedding matrix and the bidirectional LSTMs are trained end-to-end over the whole model.

Considering the following two attention mechanism focus on different semantic level representation, we will maintain all hidden vectors and the last hidden vector from bi-lstm respectively. The last hidden vector is utilized to construct the sentence-level representation for the query tweet $t_q$ and all tweets in the user's social exposures $H_l$. And all hidden vectors will be stored in the corresponding user history memory. More details will be described in the next two sections.

## 4.2 Hierarchical Attention Memory Network

Intuitively, a user's history interest will be reflected in his post history. Furthermore, the similarity between user's history interest and author's interest can influence the retweet behaviour when given a query tweet. Hence, we propose a hierarchical attention memory network to model the interest similarity between the author and the user. As shown in the left part of Figure 3, both user posting history set $H_u$ and author posting history set stored in the memory have a hierarchical structure. Take $H_u$ as an example, there are many tweets in user posting history set: $H_u = \{t_{u_1}, t_{u_2}, \cdots, t_{u_N}\}$ and these are tweet-level information stored in the user history memory. Next, there are many words in each tweet: $t_{u_i} = w_1, w_2, \cdots, w_T$ and these are word-level information stored in the user history memory. There is no wonder that not all tweets in the history memory contribute equally to modelling user's interest, nor do all words in each tweet. Hence, we propose a hierarchical architecture to model the interest similarity between the user and the author.

Specifically, in order to construct a high-level representation of the similarity between author's interest and user's interest, we stack $H$ layer of hierarchical attention on user's history memory and author's history memory. We denote the above query tweet

feature representation $u_q$ as the initial query vector $O^0$ for $H$ layer stacked hierarchical attention. Then, we utilize $O^h$ to formulate textual attention probabilities over the author's tweet histories and the user's tweet histories, and construct the next query vector $O^{h+1}$. And the final interest similarity vector $O^H$ is constructed after $H$ layer hierarchical attention.

**Word-level encoder**

Give an tweet history set $t_1, t_2, ..., t_N$, first, each word $word_{i,j}$ of $t_i$ is embedded into a vector $w_{i,j}$ (dimension of $w_{i,j}$ is $d$) using an embedding matrix A (size of A is $d \times |V|$). Then, after a bi-lstm layer, each word hidden vector $u_{i,j}$ will be stored into the corresponding history memory. Leveraging the advantage of filtering irrelevant words, at the $h$-th hierarchical-attention layer, we use the last step of the query vector $O^{h-1}$ to generate attention probabilities over the word-level part of user's history memory. The detail of match between input memory vector $u_{i,j}$ and $O^{h-1}$ is illustrated as follows:

$$z^h_{i,T} = (W^h_O O^{h-1})^{tr} W^h_T u_{i,T}, \tag{4}$$

$$a^h_{i,T} = softmax(z^h_{i,T}/\sqrt{d_k}), \tag{5}$$

$$\widetilde{u}^h_i = \sum_{j=0}^{T} a^h_{i,j} u_{i,j}, \tag{6}$$

where $T$ is the max length of each tweet, $u_{i,T} \in \mathbb{R}^{d \times T}$ is the representation matrix of tweet $i$, "$tr$" denotes matrix transpose operation, and $\sqrt{d_k}$ denotes the scaling factor as $W^h_O \in \mathbb{R}^{d_k \times d}$, $d_k < d$.

**Tweet-level encoder**

Following the above procedure, we formulate a new representation $\widetilde{u}^h_i$ for each history tweet $t_i$ based on a word-level attention mechanism. Similarly, not all tweets are equally relevant to constructing a user's interests. Hence, in order to model the whole interest of a user, we also utilize the last step of the query vector $O^{h-1}$ to query the new representations of each history tweet $\widetilde{u}^h_i$. Modelling the representation of a user's tweet histories based on the tweet-level attention probability distributions:

$$z^h_N = (W^h_{ON} O^{h-1})^{tr} W^h_N \widetilde{u}_N, \tag{7}$$

$$a^h_N = softmax(z^h_N/\sqrt{d_k}), \tag{8}$$

$$\widetilde{u}^h_* = \sum_{i=0}^{N} a^h_i \widetilde{u}_i, \tag{9}$$

where $N$ is the amount of tweets stored in user history memory, $\widetilde{u}_N \in \mathbb{R}^{d \times N}$ is the representation matrix formulated by the new representation vector of each tweet $i$, the label "$*$" can represent the author or the user in the tweet history interest representation.

Further, the $h$-th global representation vector is used to denote the interest similarity between the author and the user according to a query tweet. In each hop $h$, the global representation is updated by combining the high-level representation constructed from both user's posting history and author's posting history: $O^h = \widetilde{u}^h_u + O^{h-1} + \widetilde{u}^h_a$. The hops operator allows the model to recurrently accumulate information from the supporting memory, ultimately producing a final joint representation for modelling the interest similarity between user and author.

## 4.3 Masked Self-Attentive Mechanism

Obviously, the real-time posting from a user's followees can cover various topics. Moreover, the followees of a user can be viewed as local community, and hot topics discussed in the community may attract user to participate in these topics by retweeting or replying. Therefore, the challenge is how to model the hot topics contained in the recent posting from the user's followees. Considering the non-sequential continuity of the recent posting from the user's followees, it is not appropriate to construct the context information with Recurrent Neural Network-like sequential models. Hence, the self-attention mechanism with a maximum length of 1 is a better choice to construct the context aware representation among the recent posting from the user's followees $H_l$.

**Preliminary**

Firstly, we introduce some background on Transformer [39], which contains the building block for our masked self-attentive mechanism. And we introduce the *scaled dot-product attention*, which is the foundation of Transformer. Given a query $q_i \in \mathbb{R}^d$ from all queries, a set of keys $k_t \in \mathbb{R}^d$ and values $v_t \in \mathbb{R}^d$ where $t = 1, 2, \cdots, T$, the scaled dot-product attention outputs a weighted sum of values $v_t$, where the attention probabilities are determined by the dot-products of query $q$ and keys $k_t$. And the formulation of attention output on query q is as follows:

$$Att(Q, K, V) = softmax(\frac{K^T Q}{\sqrt{d}})V, \tag{10}$$

The *multi-head attention* consists of $H$ paralleled scaled dot-product attention layers called "head", where each "head" is an independent *dot-product attention*. And the *multi-head attention* allows the model to jointly attend to information from different representation subspaces at different positions. The attention output from multi-head attention is as below:

$$MultiHead(Q, K, V) = Concat(head_1, head_2, \cdots, head_H)W^O, \tag{11}$$

$$head_i = Att(QW^Q_i, KW^K_i, VW^V_i) \tag{12}$$

where $W^Q_i, W^K_i, W^V_i \in \mathbb{R}^{d \times d_k}$ are the independent head projection matrices, $i = 1, 2, \cdots, H$ and $W^O \in \mathbb{R}^{hd_k \times d}$.

As shown in the right part of Figure 3, we utilize two parallel stacked encoders to construct the topic-aware representation among the social exposures of a user (also named the recent post from a user's followees). And the masked self-attention module consists of three components: 1) General Stacked Encoder, 2) Mask Generator, and 3) Masked Encoder Stack. As mentioned above, self-attention mechanism with a maximum length of 1 is a better choice to construct the context aware representation among the recent posting from the user's followees $H_l$. Hence, the primary task of both two parallel stacked encoder is focus on formulating the context-aware representation. However, the difference is that the masked encoder multiplies the input of each encoder layer by a mask matrix to perceive the hot topics. Moreover, the mask matrix is formulated by the outputs of the general stacked encoder and query tweet.

**General encoder**

Specifically, our general encoder is composed of a stack of $N = 6$ identical layers. And each layer is composed of the above multi-head self-attention layer and fully connected feed-forward network. Like [39], we also incorporate residual connection and layer normalization around each of two sub-layers. The fully connected feed-forward network consists of two linear transformations with a ReLU activation in between.

$$FFN(x) = ReLU(xW_1 + b_1)W_2 + b_2 \tag{13}$$

Where $W_1 \in \mathbb{R}^{d \times d_F}$, $W_2 \in \mathbb{R}^{d_F \times d}$, $d < d_F$, and the fully connected feed-forward network can be viewed as two convolutions with kernel size 1. In other words, the $FFN$ can increase the nonlinear characteristics while keeping the feature map scale unchanged.

And $i$-th layer of general encoder is formulate as follows:

$$\widetilde{u}^i_{l_N} = MultiHead(u^{i-1}_{l_N}, u^{i-1}_{l_N}, u^{i-1}_{l_N}), \tag{14}$$

$$u^i_{l_N} = FFN(\widetilde{u}^i_{l_N}), \tag{15}$$

where $u^i_{l_N}$ represents the $i$-th layer output of the identical encoder, and $u^0_{l_N}$ is initiated by the social exposure representation matrix $u_{l_N}$. Moreover, the difference is additional mask matrix which helps to perceive the hot topics among the recent posts from a user's followees:

$$\begin{cases} u^{i-1}_{l_N} = u^{i-1}_{l_N}, & if Mask = None \\ u^{i-1}_{l_N} = Mask u^{i-1}_{l_N}, & Else \end{cases} \tag{16}$$

Based on the context-aware representation of social exposure tweets, the mask generator module incorporates the query tweet $u_q$ to formulate the relevant probability for each social exposure tweet $u_{l_i}$:

$$Mask = \sigma(\frac{(u_q W_1)^T u_{l_N}}{\sqrt{d}}) \tag{17}$$

where $\sigma$ denotes the sigmoid activation function, which limits the value of each channel from zero to one. And $u_{l_N}$ is the last layer output of the general encoder stack.

After above procedure, we utilize the query tweet $u_q$ to query the output of masked encoder, generating the hot-topic aware representation $O^l$ based on attention probability distributions:

$$z_N = (W_N u_q)^{tr} W_N u_{l_N}, \tag{18}$$

$$a_N = softmax(z_N / \sqrt{d_k}), \tag{19}$$

$$O^l = \sum_{i=0}^{N} a_i u_{l_N}, \tag{20}$$

## 4.4 Prediction

Finally, we concatenate the representation of history interest similarity $O^H$ and hot-topic aware social exposures $O^l$ obtained from the above process, and we denote the final representation as $O^f$. Then, we utilize multi-layer perceptron followed by a single-layer softmax classifier to determine whether or not the user would retweet the query tweet:

$$f = \sigma(MLP(O^f)), \tag{21}$$

where $MLP$ are the multi-layer perceptron, and $\sigma$ is the non-linear activation function sigmoid.

The final prediction is made by the following equations:

$$p(y = i|f; \theta_s) = \frac{exp(\theta^i_s f)}{\sum_j exp(\theta^j_s f)}, \tag{22}$$

where $\theta^i_s$ is a weight vector of the $i$-th class and $j \in \{0, 1\}$.

In our work, the training objective function is formulated as follows:

$$J = \sum_{(t_q, a, u, l, i) \in D} -logp(i|t_q, a, u, l; \theta), \tag{23}$$

where $D$ is the training set. $i \in \{0, 1\}$ is the label of the quadruples $(t_q, a, u, l)$, and when $i = 1$, the user $u$ would retweet the query tweet $t_q$, and $i = 0$ represents the user $u$ would not take retweet behaviour. $\theta$ is the whole parameter set of our model.

To minimize the objective function, we use a stochastic gradient descent (SGD) with the Adam [20] update rule. And the detail of hyper-parameter will be illustrated in following section.

# 5 EXPERIMENT

## 5.1 Baseline and Setup

To analyze the effectiveness of our model, we evaluate some traditional and state-of-the-art methods as baselines as follows on the constructed corpus:

- **NB**: We applied NB to model the posterior probability of each query tweet given the posting history of the users and the authors and the recent tweets posted by the followees of the users. Additionally, we utilize GloVe's Twitter vectors [33] to represent each word and average all the words vectors as the feature of a tweet.
- **SVM**: We implemented the method proposed in [28], which used an SVM to solve retweet prediction problem. Similar to the Naive Bayes, we consider the same information and the same embedding matrix to make a prediction.
- **AUT-CNN**: As we defined the retweet prediction task as a binary classification problem, we used the public code of the method proposed in [19]. In this model, we utilize multiple window sizes (3,4,5) to model the quadruple (query tweet $t_q$, author posting history $H_a$, user posting history $H_u$ and user's social exposures $H_l$).
- **LSTM-ATT**: First, we utilize the LSTM to generate the representation of each tweet. Then, we incorporate attention mechanism to process the three parts of information to make a prediction.
- **DMN**: Dynamic Memory Networks (DMN) is proposed in [23] for natural language processing. We utilize the dynamic memory mechanism from DMN to model the interests of the authors and the users, and address the retweet prediction problem.
- **SUA-ACNN**: SUA-ACNN is proposed in [50]. It incorporates only the posting history of the users and the authors. This was the state-of-the-art approach used for the retweet prediction task.

In this work, the model was implemented in the Pytorch framework, At the training stage, we utilized Bi-GRU to construct the tweet representation. The cell dimension $d$ was set to be 100, and the stack layer of the GRU was 3. The embedding dimension in the

**Table 2: Comparison results on three versions of testing dataset. We divided the testing dataset into three categories based on the capacity of each collection (the posting history of the author $H_a = \{t_{a_1}, t_{a_2}, \cdots, t_{a_N}\}$, the posting history of the user $H_u = \{t_{u_1}, t_{u_2}, \cdots, t_{u_N}\}$ and the social exposure tweets posted $H_l = \{t_{l_1}, t_{l_2}, \cdots, t_{l_N}\}$). Category "$N = 10$" denotes that we randomly sample 10 tweets from users' history to form $H_a$ and $H_u$, respectively. And we also select 10 tweets immediately before the retweet as social exposure collection $H_l$. Both Category "$N = 20$" and Category "$N = 30$" are in the same way.**

| | $Dataset_{N=10}$ | | | $Dataset_{N=20}$ | | | $Dataset_{N=30}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| Target Rate | Precision | Recall | F-Score | Precision | Recall | F-Score | Precision | Recall | F-Score |
| NB | 0.128 | 0.462 | 0.200 | 0.155 | 0.496 | 0.236 | 0.179 | 0.531 | 0.268 |
| SVM [28] | 0.207 | 0.526 | 0.297 | 0.253 | 0.563 | 0.349 | 0.288 | 0.608 | 0.391 |
| AUT-CNN [19] | 0.658 | 0.587 | 0.620 | 0.704 | 0.638 | 0.669 | 0.757 | 0.632 | 0.689 |
| LSTM-ATT | 0.769 | 0.647 | 0.703 | 0.788 | 0.686 | 0.733 | 0.825 | 0.672 | 0.741 |
| DMN [23] | 0.785 | 0.684 | 0.731 | 0.773 | 0.717 | 0.744 | 0.812 | 0.701 | 0.752 |
| SUA-ACNN [50] | **0.787** | 0.691 | **0.736** | 0.772 | 0.712 | 0.741 | 0.799 | 0.705 | 0.749 |
| AUT-MSAM | 0.772 | **0.693** | 0.731 | **0.789** | **0.759** | **0.773** | **0.842** | **0.732** | **0.783** |

experiment was 100 and the embedding layer was initialized by GloVe's Twitter vectors [33]. The number of hops for the hierarchical memory layer was set to 3. For the masked self-attentive module (MSAM), both two encoders were composed of a stack of $L = 6$ identical layers. Specifically, we set the dimension of multi-head attention to 64 and utilized 8 head to compose the multi-heads. To increase the nonlinear characteristics while keeping the feature map scale unchanged, we set the dimension of inner-layer to 2048. And the learning rate was 0.001, mini-batches was 150, and the dropout rate was set to 0.2. Further, we used precision (P), recall (R), and F1-score (F1) to evaluate performance.

## 5.2 Results and Discussion

The performance of different methods on our datasets is listed in Table 2. There are three blocks, and the difference between blocks rely on the capacity of each collection (the posting history of the author $H_a = \{t_{a_1}, t_{a_2}, \cdots, t_{a_N}\}$, the posting history of the user $H_u = \{t_{u_1}, t_{u_2}, \cdots, t_{u_N}\}$ and the social exposures tweets posted $H_l = \{t_{l_1}, t_{l_2}, \cdots, t_{l_N}\}$). Category "$N = 10$" denotes that we randomly sample 10 tweets from users' history to form $H_a$ and $H_u$, respectively. And we also select 10 tweets immediately before the retweet as social exposure collection $H_l$. Both Category "$N = 20$" and Category "$N = 30$" are in the same way. When we have enough information, we can observe that our proposed model AUT-MSAM achieves a better performance than other methods of all metrics.

By compared with SUA-ACNN, which was the state-of-the-art method for the retweet prediction task, the proposed model (AUT-MSAM) achieves a relative improvement of 5.4% in precision, along with a 3.8% increase in recall and 4.5% increase in F-score when the "$N = 30$". Moreover, our model (AUT-MSAM) also achieves a relative improvement of 2.2% in precision, along with a 6.6% increase in recall and 4.1% increase in F-score when the "$N = 20$". Particularly, when the capacity of each tweet collection is set to 10, we can find that our proposed model only achieve a little better performance in recall. As SUA-ACNN was proposed to construct user interest by utilizing only 5 history posting tweets, hence, it can achieve a little better performance with limited information. But, the results can also prove our can address long-range dependencies challenge, and construct a better hot topic aware representation.

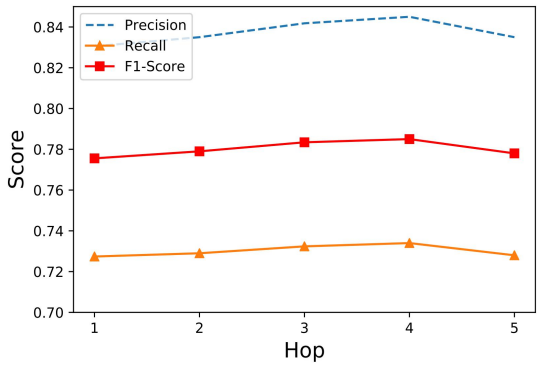


**Figure 4: Performance on different hop of hierarchical attention memory network**

In order to prove the effectiveness of incorporating users' tweet posting histories and the recent posting from users' followees, we also apply Dynamic Memory Network (DMN) [23] on our dataset. Moreover, this method is proposed for question answering task and can learn to dynamically construct external memory. From the results table, we can observe that our proposed model ( AUT-MSAM) achieves a better performance than DMN in all evaluation results. Compared with the DMN, our model achieves more than 3.6% relative improvements in precision, 4.4% relative improvements in recall and 4.1% relative improvements in F-score. Hence, by incorporating users' tweet posting histories and the recent posting from users' followees, our proposed model performs well on the retweet prediction task.

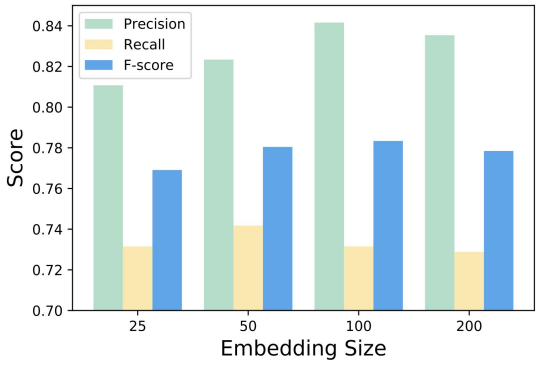


**Figure 5: Influence of Embedding size**

## 5.3 Parameter Influence

The proposed model contains several critical hyper-parameters. We analyzed the influence of critical parameters from the following perspectives: 1) the hops of hierarchical attention memory network, and 2) the embedding dimension. We vary one parameter and fix the others in turn to evaluate their influences. Based on the experimental results, we can observe that the proposed model could achieve stable performance, in the condition of various parameter settings.

The first parameter we evaluated is the hops of hierarchical attention memory network, which we varied from 1 to 5 in this experiment. In Figure 4, we draw the Precision, Recall and F-score curves to show the the influence of the hops. Along with the increase hops of the hierarchical attention memory layer, the results are better. We also obtained the best performance with the 4-layer hierarchical attention, which indicates the robustness of our model along with the depth deeper. Since increasing the number of layers of the network will make the model more complex with more parameters, and the model attempt to be overfitting. Hence, 4 hops of hierarchical attention memory layer is a better choice.

To evaluate how embedding dimension influence the performance, we fix the hops of the hierarchical attention memory layer to 4 and tried different embedding dimensions. Since the pre-trained Glove Twitter Embedding [33] only provides four kind of dimensions, we have tried 25, 50, 100, 200 respectively. The comparison results shown in Figure 5 demonstrate that the models with a medium embedding dimension performed better. The results improved when the dimension was increased from 25 to 100, and the result of the 200 embedding dimension was worse than the 100 dimension. This shows that the word representation in 100 dimension is enough to represent the semantic space in the dataset.

## 5.4 Ablation Study

We further analyze the major components that contribute a lot to the performance. The results are illustrated in Table 3. As mentioned before in Problem Definition, with the consideration of users' interest and potential hot topic information, AUT-MSAM is verified to make a considerate prediction of query tweet. As shown in Table 3, the performance of AUT-MSAM significantly decreased when removing any of them. Definitely, the user history is critical for making great prediction. However, the mask mechanism is shown to contributing more than 50% performance improvements, which is achieved after incorporating social exposures. Specifically, we simply utilize general encoder stack in AUT-MSAM (w/o Mask).

**Table 3: Ablation of our proposed model AUT-MSAM.**

| Method | Precision | Recall | F1 |
|---|---|---|---|
| AUT-MSAM (w/o Mask) | 0.828 | 0.725 | 0.773 |
| AUT-MSAM (w/o Social Exposures) | 0.811 | 0.722 | 0.764 |
| AUT-MSAM (w/o Author History) | 0.808 | 0.719 | 0.761 |
| AUT-MSAM (w/o User History) | 0.801 | 0.676 | 0.733 |
| **AUT-MSAM(ours)** | **0.842** | **0.732** | **0.783** |

## 6 CONCLUSION

With the rapid development of social media and the great richness of topic diversity, in this paper we introduce a integrated framework to incorporate the hot topics discussed in users' social exposures and users' posting history to perform retweet prediction task. Considering that the real-time posts from users' followees contain various topics, and these tweets come from different users, it is out of sequential continuity. Hence, we propose a novel masked self-attention to construct the hot topic aware representation among the recent tweets posted by a user's followees. Since user's posting history tweets are not equally important in modelling user's interest, we utilize a hierarchical attention memory network, which generates word-level attention and tweet-level attention sequentially. Therefore, our model can not only sense whether the query text matches the user's interest, but also the social influence of hot topics, which is a very important supplementary information for retweet prediction. We also constructed a large data collection retrieved from Twitter to evaluate the effectiveness of our model. Experimental results showed that the proposed method achieves better performance than state-of-the-art methods ignoring hot topics.

## ACKNOWLEDGMENTS

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by China National Key R&D Program (No. 2018YFC0831105,2017YFB1002104), National Natural Science Foundation of China (No. 61751201, 61532011), Shanghai Municipal Science and Technology Major Project (No.2018SHZDZX01), STCSM (No.16JC1420401,17JC1420200), ZJLab.

## REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *Computer Science* (2014).
[2] Bin Bi and Junghoo Cho. 2016. Modeling a retweet network via an adaptive bayesian approach. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 459–469.
[3] Johan Bollen, Huina Mao, and Alberto Pepe. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *Icwsm* 11 (2011), 450–453.

[4] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science* 2, 1 (2011), 1–8.

[5] Danah Boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*. IEEE, 1–10.

[6] Kailong Chen, Tianqi Chen, Guoqing Zheng, Ou Jin, Enpeng Yao, and Yong Yu. 2012. Collaborative personalized tweet recommendation. (2012), 661–670.

[7] Giovanni Comarela, Mark Crovella, Virgilio Almeida, and Fabricio Benevenuto. 2012. Understanding factors that affect response rates in twitter. In *Proceedings of the 23rd ACM conference on Hypertext and social media*. ACM, 123–132.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[9] Zhuoye Ding, Xipeng Qiu, Qi Zhang, and Xuanjing Huang. 2013. Learning Topical Translation Model for Microblog Hashtag Suggestion. In *International Joint Conference on Artificial Intelligence*. 2078–2084.

[10] Wei Feng and Jianyong Wang. 2013. Retweet or not?: personalized tweet re-ranking. In *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 577–586.

[11] Syeda Nadia Firdaus, Chen Ding, and Alireza Sadeghian. 2016. Retweet prediction considering user's difference as an author and retweeter. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE Press, 852–859.

[12] Yuyun Gong and Qi Zhang. 2016. Hashtag recommendation using attention-based convolutional neural network. In *International Joint Conference on Artificial Intelligence*. 2782–2788.

[13] Yeyun Gong, Qi Zhang, Xuyang Sun, and Xuanjing Huang. 2015. Who Will You "@"?. In *ACM International on Conference on Information and Knowledge Management*. 533–542.

[14] Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences* 101, suppl 1 (2004), 5228–5235.

[15] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al. 2001. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. (2001).

[16] Haoran Huang, Qi Zhang, Xuanjing Huang, Haoran Huang, Qi Zhang, and Xuanjing Huang. 2017. Mention Recommendation for Twitter with End-to-end Memory Network. In *Twenty-Sixth International Joint Conference on Artificial Intelligence*. 1872–1878.

[17] Haoran Huang, Qi Zhang, Jindou Wu, and Xuanjing Huang. 2017. Predicting Which Topics You Will Join in the Future on Social Media. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 733–742.

[18] Bo Jiang, Jiguang Liang, Ying Sha, and Lihong Wang. 2015. Message clustering based matrix factorization model for retweeting behavior prediction. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1843–1846.

[19] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[20] Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *Computer Science* (2014).

[21] Ioannis Konstas, Vassilios Stathopoulos, and Joemon M. Jose. 2009. On social networks and collaborative recommendation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 195–202.

[22] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. 2009. Latent dirichlet allocation for tag recommendation. In *ACM Conference on Recommender Systems, Recsys 2009, New York, Ny, Usa, October*. 61–68.

[23] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *International Conference on Machine Learning*. 1378–1387.

[24] Changliang Li, Liang Li, and Ji Qi. 2018. A Self-Attentive Model with Gate Mechanism for Spoken Language Understanding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3824–3833.

[25] Daniel Li and Asim Kadav. 2018. Adaptive Memory Networks. *arXiv preprint arXiv:1802.00510* (2018).

[26] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5, 1 (2012), 1–167.

[27] Yanbing Liu, Jinzhe Zhao, and Yunpeng Xiao. 2018. C-RBFNN: A user retweet behavior prediction method for hotspot topics based on improved RBF neural network. *Neurocomputing* 275 (2018), 733–746.

[28] Zhunchen Luo, Miles Osborne, Jintao Tang, and Ting Wang. 2013. Who will retweet me?: finding retweeters in twitter. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 869–872.

[29] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent models of visual attention. 3 (2014), 2204–2212.

[30] Yusuke Ota, Kazutaka Maruyama, and Minoru Terada. 2012. Discovery of interesting users in twitter by overlapping propagation paths of retweets. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on*, Vol. 3. IEEE, 274–279.

[31] Michael J Paul and Mark Dredze. 2011. You are what you Tweet: Analyzing Twitter for public health. *Icwsm* 20 (2011), 265–272.

[32] Huan-Kai Peng, Jiang Zhu, Dongzhen Piao, Rong Yan, and Ying Zhang. 2011. Retweet modeling using conditional random fields. In *2011 11th IEEE International Conference on Data Mining Workshops*. IEEE, 336–343.

[33] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[34] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*. ACM, 851–860.

[35] Markus Schedl and Dominik Schnitzer. 2013. Hybrid retrieval approaches to geospatial music recommendation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 793–796.

[36] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603* (2016).

[37] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social computing (socialcom), 2010 ieee second international conference on*. IEEE, 177–184.

[38] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-To-End Memory Networks. *Computer Science* (2015).

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[40] Beidou Wang, Can Wang, Jiajun Bu, Chun Chen, Wei Vivian Zhang, Deng Cai, and Xiaofei He. 2013. Whom to mention: expand the diffusion of tweets by@ recommendation on micro-blogging systems. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 1331–1340.

[41] Can Wang, Qiudan Li, Lei Wang, and Daniel Dajun Zeng. 2017. Incorporating message embedding into co-factor matrix factorization for retweeting prediction. In *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 1265–1272.

[42] Mengmeng Wang, Wanli Zuo, and Ying Wang. 2015. A multidimensional nonnegative matrix factorization model for retweeting behavior prediction. *Mathematical Problems in Engineering* 2015 (2015).

[43] Kumanan Wilson and John S Brownstein. 2009. Early detection of disease outbreaks using the Internet. *Canadian Medical Association Journal* 180, 8 (2009), 829–831.

[44] Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604* (2016).

[45] Zhiheng Xu and Qing Yang. 2012. Analyzing user retweet behavior on twitter. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*. IEEE, 46–50.

[46] Zhiheng Xu, Yang Zhang, Yao Wu, and Qing Yang. 2012. Modeling user posting behavior on social media. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 545–554.

[47] Jing Zhang, Jie Tang, Juanzi Li, Yang Liu, and Chunxiao Xing. 2015. Who influenced you? predicting retweet via social influence locality. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 9, 3 (2015), 25.

[48] Jing Zhang, Jie Tang, Yuanyi Zhong, Yuchen Mo, Juanzi Li, Guojie Song, Wendy Hall, and Jimeng Sun. 2017. StructInf: Mining Structural Influence from Social Streams.. In *AAAI*. 73–80.

[49] Qi Zhang, Yeyun Gong, Ya Guo, and Xuanjing Huang. 2015. Retweet Behavior Prediction Using Hierarchical Dirichlet Process.. In *AAAI*. 403–409.

[50] Qi Zhang, Yeyun Gong, Jindou Wu, Haoran Huang, and Xuanjing Huang. 2016. Retweet prediction with attention-based deep neural network. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 75–84.

[51] Wei Zhang, Jianyong Wang, and Wei Feng. 2013. Combining latent factor model with location features for event-based group recommendation. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 910–918.

[52] Xue Zhang, Hauke Fuehres, and Peter A Gloor. 2011. Predicting stock market indicators through twitter "I hope it is not as bad as I fear". *Procedia-Social and Behavioral Sciences* 26 (2011), 55–62.

[53] Zhou Zhao, Lingtao Meng, Jun Xiao, Min Yang, Fei Wu, Deng Cai, Xiaofei He, and Yueting Zhuang. 2018. Attentional Image Retweet Modeling via Multi-Faceted Ranking Network Learning.. In *IJCAI*. 3184–3190.

[54] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018. End-to-End Dense Video Captioning with Masked Transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8739–8748.