Online Social Networks and Media 6 (2018) 26–40

Contents lists available at ScienceDirect

Online Social Networks and Media

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

Retweet: A popular information diffusion mechanism – A survey paper

Syeda Nadia Firdaus*, Chen Ding, Alireza Sadeghian

Department of Computer Science, Ryerson University, 350 Victoria Street, Toronto, Ontario M5B 2K3, Canada

ARTICLE INFO

Article history:
Received 1 October 2017
Revised 12 February 2018
Accepted 4 April 2018

Keywords:
Retweet
Features
Prediction model
User behavior

ABSTRACT

Retweeting or reposting a message is considered as an easily available information diffusion mechanism provided by Twitter or any other social network sites. By finding out why a user retweets a tweet, or predicting whether a tweet will be retweeted by a user, we can not only understand user's behavior or interest better, but also understand how information is diffused on the online social network. In this survey paper, we have explored various research works related to retweet prediction and retweeting behavior analysis, investigated the underlying reasons for spreading information in forms of retweets, and discussed the challenges of retweet related research. The purpose of the paper is to provide an overview of researches in retweet prediction area, which can be used as an introductory guide for future researchers in this field.

1. Introduction

Human beings tend to be social in nature. Being social not only means to live in a society but also means to exchange views and information with the members of the society. Traditional social network represents a society in a specific geographical location, or serving a specific purpose (e.g., common interest). But the boon of the Internet helps us form society including people from all over the world and serving all kinds of purposes. In simple words, online social network is the network of people formed over the Internet. It carries a great amount of data which reflect its users' interest, behavior, and activities. It has the ability to spread information all around the world in the least amount of time. The data collected from these social networks have the potential to make effective contribution in many different areas of research, such as marketing, business analysis, human psychology analysis, etc. Among them, one important area is to study their use as a mechanism for information diffusion.

Twitter is one of the most popular and widely used online social network sites. Twitter allows its users to create profile, publish messages, and share information with others. On Twitter, users' posts are known as tweets which are not more than 140 characters long. These tweets may contain URLs, hashtags (keywords followed by "#" symbol used to categorize the tweet), mentions (other users' usernames followed by "@" symbol), and emoticons. Users can also include photos and videos in their tweets. In Twitter, there exists follower–followee relationship between users. When a user wants to subscribe to other users' posts, he can follow them. For example, if user *A* follows user *B*, *A* is known as follower of *B* and *B* is known as followee of *A*. Twitter allows its users to maintain one-way relationship; that means user *A* can follow any other user whereas it is not mandatory that other users have to follow *A*. In this type of relationship, it is very typical to follow celebrities and famous people to get continuous update from them. When user *A* follows *B*, *B*'s posts will appear in *A*'s Twitter main page. *A* can also like and repost *B*'s tweets. These reposts are known as retweets. Retweets are reposted tweets which look like original tweets with keyword "RT" and author's username (followed by "@" symbol) at the beginning of the text. Retweeters are also allowed to add their own comments with the original text. In this case, the retweeter's text is placed at the beginning of the retweet.

The research on information diffusion in online social networks is mainly focused on the following three general research questions [21]: (i) Which piece of information is being popular and diffused vastly? (ii) How information is diffused in a certain path in the network and why? and (iii) Which members in the network are playing important role in information diffusion? These three questions are common for any social network related information diffusion research. They can be applied to research on information diffusion in Twitter too.

Since retweet is a feature provided by Twitter as an easily accessible and popular information diffusion mechanism, there are also specific research questions that are closely related to Twitter's own distinctive characteristics, and most of the diffusion related research works are focused on user's retweet activity. Twitter network is not based on two-way relation and personal connection, rather it allows users to follow topic or people who they find

* Corresponding author.
E-mail addresses: sfirdaus@ryerson.ca (S.N. Firdaus), cding@ryerson.ca (C. Ding), asadeghi@ryerson.ca (A. Sadeghian).

https://doi.org/10.1016/j.osnem.2018.04.001
2468-6964/

interesting. So, the diffusion is fast and can engage a large number of users in short time. In Twitter, user is the main actor in retweeting activity, so their interest and intention are mainly explored by the researchers. Twitter also provides hashtag feature which allows users to track and follow topic and information according to their interest and need. Retweeting tweets with hashtags help to spread topic-related information quickly to large audience. The broad spectrum of retweet-related research includes research on prediction of retweets, retweeters, retweet counts, popularity of tweets as well as tweet recommendation. Some research papers explore, analyze, and predict user's retweet activity, some papers are focused on finding out potential retweeters, whereas other papers investigate the underlying reasons of why some tweets get more retweets or are spreading more virally. These research papers can be categorized based on their focus and research question that they try to answer in their work. Based on the general questions listed earlier, in this paper, we define three main Twitter-specific information diffusion related research questions 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?

In these papers, users' retweeting activity is mainly investigated from two perspectives: local and global. In case of local perspective, retweeting activity is explored from individual user's point of view. Every user's profile and interest are investigated to explore his retweet decision. The first research and second research question are focused on retweet activity from local perspective. In case of global perspective, tweets' general characteristics are investigated to find their retweetability. These types of research papers are focused on the third research question.

Retweeting activity is mainly dependent on three factors: user/reader of the tweet, author of the tweet, and content of the tweet. User represents the target user who gets the tweets in his timeline and decides the retweet action; author is the publisher of the target tweet; and content represents the target tweet itself including the terms, their meanings as well as the overall information carried by the tweet. Every factor can be described by multiple features. The relation between user and author is one type of feature that is associated with both of them. We can consider it either as an author factor or a user factor. To make our discussion unambiguous, in the rest of this paper, we treat it as a user factor.

In this paper, we would like to give an in-depth review on research works related to retweet. In particular, we focus on retweeting behavior analysis and retweet prediction because most of the research works presented here are analyzing retweet behavior and proposing retweet prediction models. We have also included some papers on tweet recommendation and retweeter prediction because these research works are quite related to retweet prediction. Tweet recommender system predicts retweets to build recommendation model considering retweet as an indicator of user's preference. Retweeter prediction explores user's interest on tweets to find potential retweeters. The paper will give a detailed description of retweet prediction process, including the major categories of features used for prediction, the common prediction models, evaluation metrics, datasets, as well as a discussion on challenges and open research issues.

The rest of the paper is organized as follows. Section 2 describes the categorization of papers based on aforementioned three Twitter-specific research questions. Section 3 provides the analyses of user's retweeting behavior. Section 4 explains the detailed steps for retweet prediction. Section 5 briefly describes works on information diffusion using retweet. Section 6 gives a discussion on research challenges and open issues related to retweet prediction and Section 7 concludes the paper.

2. Categorization of research papers

In Table 1, we have categorized the research papers based on the research questions they try to answer in their work. Though many research works have been done on retweet, their primary objectives can be different. In this section, we have described the categorization of retweet related research papers based on their primary objective to solve one of the three Twitter-specific research questions through their work.

The first research question is focused on investigating and predicting the tweets which will be retweeted by the user. These research papers can be further categorized based on their primary objectives. In the first sub-category, the primary objective is to analyze and investigate the factors that have influence on users' retweet activity. In these papers, researchers list all features that might have impact on users' retweeting activity and then they analyzed the effects of these features on users' retweeting behavior to identify the most influencing features. Comarela et al. [14] explored some important features such as user's prior interaction with author, author's tweeting rate, content of tweet on user's retweeting behavior. This research revealed some interesting behavioral details behind a user's retweet decision. Sun et al. [52] studied the influence of serendipitous information on user's retweet behavior and showed that users like to propagate tweets containing serendipitous information.

In the second sub-category, the objective is to not only explore and analyze the features influencing user's retweeting behavior but also propose retweet prediction models based on their investigated features. Research papers in this spectrum, investigate and predict retweet behavior from the perspective of individual users. Peng et al. [42] 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. Zhang et al. [64] explored the influence of friends from user's ego-network on his retweeting activity and then proposed retweet prediction model using only their explored features based on social influence locality. Zhang et al. [66] explored influence of author, network structure, content of tweet, and temporal information on user's retweeting probability and then proposed Hierarchical Dirichlet Process based retweet prediction model incorporating these features. Xu and Yang [59] analyzed different features to develop retweet prediction model from the perspective of individual users. Their purpose was to investigate the importance of different author-based, social-relationship based, and content-based features on user's retweet decision. They explored the effectiveness of individual feature by developing and comparing the performance of retweet prediction models with different features. Yang et al. [62] also analyzed different features related to user interest, content of tweet, and time on user's retweeting behavior and then proposed factor-graph-based retweet prediction model. Xu et al. [60] analyzed the influence of social friends and breaking news on user's retweeting behavior and incorporated these influences in their proposed mixture latent topic retweet prediction model. Hoang and Lim [23] analyzed three behavioral factors: topic virality, user virality and user susceptibility on users' retweet decision and proposed a tensor factorization retweet prediction model which represents retweets as three-dimensional tensors based on the mentioned factors. We can see that author influence, social influence/friend's influence, and content of the messages are some common factors which had been explored by many researchers. These research works made remarkable contribution to the field because they worked with different datasets and used different mechanisms to describe as well as analyze the effects of these factors to build efficient retweet prediction models.

Table 1
Paper categorization based on the research questions.

Research question	Primary objective	Title	Reference
Which tweets will be retweeted by user?	Analyze features influencing retweet activity	Understanding factors that affect response rates in Twitter	[14]
		Unexpected relevance: an empirical study of serendipity in retweets	[52]
	Analyze features influencing retweet activity and build retweet prediction model based on those features as well	Retweet modeling using conditional random fields	[42]
		Social influence locality for modeling retweeting behaviors.	[64]
		Retweet behavior prediction using hierarchical Dirichlet process	[66]
		Analyzing user retweet behavior on twitter	[59]
		Understanding retweeting behaviors in social networks	[62]
		Modeling user posting behavior on social media	[60]
		Retweeting: an act of viral users, susceptible users, or viral topics?	[23]
	Design effective retweet prediction model	Retweet behavior prediction in twitter	[25]
		Why do people retweet? Anti-homophily win the day!	[39]
		A multidimensional nonnegative matrix factorization model for retweeting behavior prediction	[57]
		Message clustering based matrix factorization model for retweeting behavior prediction	[27]
		Identifying retweetable tweets with a personalized global classifier	[56]
		Retweet prediction with attention-based deep neural network	[67]
	Design effective tweet recommendation model	User oriented tweet ranking: a filtering approach to microblogs	[55]
		Twitter user modeling and tweets recommendation based on Wikipedia concept graph	[36]
		Collaborative personalized tweet recommendation	[13]
Who will retweet the target tweet?	Finding out retweeters	Who will retweet me? Finding retweeters in Twitter	[37]
		Who will retweet this? Detecting strangers from Twitter to retweet information	[32]
Why do some tweets get more retweets?	Finding out the reasons behind spreading of information by retweet activity	RT to Win! Predicting message propagation in Twitter	[44]
		Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network	[51]
		Predicting retweet count using visual cues	[10]
		Modeling and predicting retweeting dynamics on microblogging platforms	[19]
		Bad news travel fast: a content-based analysis of interestingness on twitter	[41]
		Analyzing and predicting viral tweets	[26]
		Emotional divergence influences information spreading in Twitter	[45]
		Political communication and influence through microblogging – an empirical analysis of sentiment in Twitter messages and retweet behavior	[50]
		Role of sentiment in message propagation: reply vs. retweet behavior in political communication	[30]

In the third sub-category, the primary objective is to develop retweet prediction models based on already known features. The focus for these research papers is on the design of effective prediction models. They use different machine learning methods to build novel and accurate retweet prediction models. Huang et al. [25] proposed a novel methodology based on Bayes model to find user's interest in different categories and predicts user's retweet decision depending on the interest measurement. Macskassy and Michelson [39] developed different retweet prediction models: general/random decision based, recent communication based, on-topic based, and homophily based to have detailed understanding on users' retweet decision. Wang et al. [57] and Jiang et al. [27] proposed matrix factorization retweet prediction models. Wang et al. [57] used user-based and content-based features to incorporate user similarity, activity, interest, and content's influence on his retweeting activity and developed nonnegative matrix factorization retweet prediction model. Jiang et al. [27] tried to avoid the complexity of finding user similarity in a large network. So, they only utilized the impact of message similarity on user's retweeting behavior and proposed message-clustering-based retweet prediction model using matrix factorization technique. Zhang et al. [67] designed retweet prediction model using attention based deep neural network incorporating user's interests and user/author information. The capability of deep neural network to learn optimal features automatically helped them build a state-of-the-art prediction model without the complex task of feature engineering. Vougiouk et al. [56] investigated the effectiveness of different feature sets based on user, author, and content of the tweet, built logistic-regression-based personalized retweet prediction model and proposed a state-of-the-art model with only 10 features. These research works are mainly focused on the design of prediction models and they did not intend to analyze the influencing features on user's retweet behavior, rather they explored different machine learning techniques to design one competent model.

In the fourth sub-category, researchers worked on the detection of retweetable tweets in order to design tweet recommender system considering retweetable tweets as user's preferred item. Uysal and Croft [55] explored different user, author, and content-based features to define tweet's interestingness and used learning to rank strategy to define retweet-likelihood-based tweet ranking. Lu et al. [36] considered retweets as tweets relevant to user's interest and ranked tweets based on their similarity with user profile developed using Wikipedia concept graph. Assuming user's retweeting action as their personal preferences based on usefulness and informativeness of the tweets, Chen et al. [13] developed a personalized tweet recommender system using collaborative ranking method. Though primary objective of these research papers is to design personalized tweet recommender system, they did investigate on the behaviors of the retweetable tweets because retweetability is a good indication of being a good recommendation candidate.

The second research question is related to finding potential retweeters, or identifying who is more likely to retweet a tweet among all the followers of the author of the tweet. Since retweet is a significant mechanism for information diffusion, finding out proper target users is an important task in order to spread the information efficiently. Luo et al. [37] and Lee et al. [32] both were focused on prediction of potential retweeters for target tweet though their approaches towards the problem are different. Lee et al. [32] aimed to find out retweeters among users who are requested to retweet the tweets on a specific topic. Their purpose was to find out potential retweeters to spread information during an emergency case. Luo et al. [37] used learning to rank model to rank followers based on their retweet probability for a target tweet. In case of finding out potential retweeters, researchers mainly put emphasis on feature sets that define follower's intention and activity for the task of retweeting.

It was observed that some tweets have the potential to be retweeted more by the users. Researchers focusing on the third research question explored the underlying reasons that caused the virality of tweets. These research papers did not predict or study retweet behavior from the perspective of individual users, rather they explored the retweetability of a tweet from global perspective. Suh et al. [51] explored a large number of content-based and contextual features to find their underlying association with tweet's retweetability. The objective of Petrovic et al. [44] was similar to the work of Suh et al. [51], but they explored relatively small number of tweet's content-based features and social features related to author to predict retweetability of streaming tweets. Finding out effective features from tweets is a challenging task due to their length restriction. To overcome this limitation, Can et al. [10] used visual cues from the image linked to the tweet to find its retweetability and showed that visual cues served as a competent added factor to find a tweet's retweet count. Gao et al. [19] included the impact of tweet's age and users' time-dependent activity to find the popularity of tweets. They showed that not only the interestingness of tweets but their posting times also have effect on popularity of tweets. Researchers also investigated the impact of sentiments on tweets' retweet probability [26,30,41,45,50]. With different objectives, different settings, and different datasets, all these research papers found that tweets reflecting negative sentiments have higher probability to be retweeted by users.

3. Analysis of retweeting behavior

Users are the main actors in online social networks. They create, initiate and propagate information. So, users' retweeting behavior has been investigated and analyzed broadly in this area. Though user's retweeting decision is subjective, results from these analyses can help us gain a better understanding on why they make these decisions. Comarela et al. [14] showed that newer tweets, tweets from previously retweeted authors, and authors with lower posting rate have higher probability to be retweeted by the users. The study also showed that users like to retweet shorter tweets. The reason can be that, in case of shorter tweets, users might get room to add their own text. Zhang et al. [64] investigated neighbors' influence on user's retweet activity. The experiment showed that a user's retweeting probability was positively correlated with number of his active friends whereas the probability was negatively correlated with the number of connected circles formed by those active friends. The reason can be that a user might not be interested to retweet a message which is already known by many of his neighbors. Zhang et al. [66] and Petrovic et al. [44] showed that, author of tweet has good influence on user's retweet activity. When same microblog was posted by two different authors at different time slot, many users repost the microblog posted at earlier time even though another same post appeared first in their main page, which clearly indicates the influence of author on user's retweeting behavior [66]. It was found that author's authority such as number of followees, number of times the author was listed, and inclusion of teen-related topics increases the retweetability of a tweet [44].

A large-scaled analysis has been done to find features that have good impact on tweet's retweetability [51]. Suh et al. [51] used Principal Component Analysis (PCA) to explore influencing features and built Generalized Linear Model to explain the influence of these features on finding the retweet probability. According to their result, the number of followers and followees and age of the account of the author have positive influence on the retweetability of a tweet. On the other hand, there is no strong correlation between an author's total number of past tweets and his retweet rate. As per their analysis, hashtags and URLs have strong correlation with retweetability of a tweet and in case of URLs, the retweet rate varies in different domains.

Sun et al. [52] made an interesting finding that users like to diffuse serendipitous information. They defined serendipity as unexpected tweet from source (author) which is useful or relevant for receiver (user). They developed method using Likelihood Ratio Test to check unexpectedness and relevance of tweets. The unexpectedness test was eventually a test to find out whether the tweet can be explained by the perceived model (based on the received information from the source) of the source (developed by the receiver) or can be explained by the mixture model of multiple contexts. They also developed a preference model of the receiver based on his postings. Then they checked if the tweet is relevant to the user's posting. If a tweet is unexpected from a source as well as relevant to the user, it is serendipitous. From this work, the researchers found that 27% of retweets in Twitter and 30% of retweets in Weibo contain serendipitous information.

Lee et al. [32] built models based on user's personality traits, social behavior, social relation, and content of the tweet to see the willingness of the user to propagate information when they were asked to do so during the emergency case. In this research, a good number of features have been used to define a user as a potential retweeter or non-retweeter. User's activity, personality, readiness (to retweet), and past retweeting behavior related features showed strong impact on classifying a user as retweeter. Researchers also confirmed that aging of a message has impact on its popularity to be retweeted. As per Gao et al. [19], popularity of a message to be retweeted follows power law distribution with its aging process.

Many researchers investigated the impact of sentiments of tweets on user's retweeting behavior [26,30,41,45,50]. Based on their findings, in general, tweets with negative sentiment were retweeted more by the users. Jenders et al. [26] showed that tweets with excessive negative sentiments do not have the potential to be viral. Stieglitz and Linh [50] investigated that in case of political information diffusion, messages containing positive or negative sentiment had higher probability to be retweeted by others. In this case as well, massages with negative sentiment were retweeted more than messages with positive sentiment. As a measure of sentiment, researchers mainly considered positive, negative, and neutral sentimental score of the tweet. Some researchers [26,45] also calculated emotional divergence of a tweet which is basically the normalized absolute difference between the positive and negative sentiment score of the tweet. Pfitzner et al. [45] showed that highly emotionally diverse tweets had five times higher chance to be retweeted by the users. Naveed et al. [41] used dictionary-based approach [29] to find sentiments of the tweets. As a measure of sentiment, they used valence, dominance and arousal score of a tweet. Researchers also used LIWC (Linguistic Inquiry and Word Count) program [43] to find the sentiments of tweets based on the number of words in the tweet which belong to the following two LIWC categories: "Positive emotion" and "Negative emotion" [50]. According to an experiment by Berger [5], an expert in viral marketing and social influence, people in high arousal state (after running or jogging) tend to spread information more than people in low arousal state (sitting still). Berger also showed that arousal always increases social transmission no matter it is positive (amusement) or negative (anxiety). Results of Burger's experiment somehow correlate with user behavior analysis research for retweeting, as it is found that users usually like to spread information containing non-neutral sentiment, especially negative sentiment.

4. Retweet prediction

The research in retweet prediction is mainly conducted in four steps. In the first step, researchers collect Twitter dataset and then in the second step, various features belonging to three factors (author, user, content) are extracted from the dataset. The third step includes design of retweet prediction model using the extracted features. The final step includes evaluation of the proposed model. In this section, we discuss each of these steps and how they are implemented in different research works, as well as how the three tweet-related factors and their corresponding features are utilized for retweet prediction. Here we have included research on tweet recommendation because these works are focused on the same research question as retweet prediction (see Table 1) and follow the similar steps as the work on retweet prediction. We have also included a few research papers on retweeter prediction as they explore some important user-author relations and tweet content features to find users who might have interest to retweet the tweet.

4.1. Collection of data

Data collection is an important step of retweet prediction research. Some research used publicly available dataset [27,41,57] whereas some collected data from Twitter on their own. Twitter provides APIs to the developers to get access to Twitter social network data. Researchers use these APIs to gather required information regarding users, their networks, and tweets ([42,55,36,10,32,37,44,51,52,60,67]). Twitter offers two types of APIs: REST and Streaming. These APIs provide different methods to get data such as user status information, user's tweets, user's follower/followee information etc. REST API allows developers to get information based on specific parameters whereas Streaming API delivers live tweet data based on query. These APIs are available through Outh-based authorization system. There are also some limitations of using these APIs. REST API allows to get at most 3200 latest tweets from a user.[1] In case of searching tweets based on a query, standard version of Twitter REST API returns a sampling of recent tweets published in the last 7 days.[2] For getting real-time tweets using streaming API, standard version of API allows to track at most 400 keywords, 5000 users, and 25 locations.[3] Twitter provides enterprise versions of these APIs which allow the users to get elevated access to their data. But enterprise versions are not free while the standard ones are. Twitter also puts rate limit per request on getting their data. All necessary information of using Twitter APIs is available in their developer platform website.[4]

Responses from these APIs are in JSON format which can then be parsed to get the required data. Third party libraries such as Twitter4J,[5] tweepy,[6] twitter-python[7] can also be used to collect and process data from Twitter. Snowball sampling method can be used to get information of large connected network [38,39]. In this method, researchers select some seed users and then collect data from users who are connected to the seed users (through retweet/mention) and this process continues until an adequate amount of data is obtained.

4.2. Feature extraction

The accuracy of retweet prediction greatly depends on which features are used and whether they are effective in terms of predicting retweet. The past research has shown that author, user, and content of the tweet have great impact on user's retweeting decision. These factors could capture or reflect the impact of author's influence, author and user's social relation, user's interest, and content of the tweet on retweeting activity. Different features based on these three factors and their objectives are given in Table 2. Below is the description of features based on these factors.

4.2.1. Author of the tweet

Intuitively it can be said that author of a tweet has good impact on its retweetability. Findings from the past research also support this intuitive observation. According to the study conducted by Cha et al. [11], if a tweet is from content aggregation service or news media, or from a popular and most mentioned user such as celebrities, it will get more retweets. Number of followers and followees of the author, age of the account, number of tweets from the author, tweet frequency (per day) of the author, number of tweets favored by others, language of the author, ratios of retweeted tweets, ratios of tweets receiving replies, and whether the author is a verified user or local elite, are good features that can be used to measure author's influence on the retweet decision ([13,26,44,51,55,59]).

4.2.2. User of the tweet

One of the basic questions in retweet related research is "Which tweets will be retweeted by user?". From this research question, it is evident that user is the primary actor in retweeting activity. Since retweet is a personal decision, it is hard to find any definite answer to this question as the reasons for retweeting could be purely subjective and thus varied from user to user. The most common reasons could be listed as follows: the user wants to spread the information; the user finds it interesting enough to share with others; the user finds the tweet helpful for others; the user's relation with the author of the tweet influences him; the user is influenced by his neighbors in the social network.

[1] https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user_timeline

[2] https://developer.twitter.com/en/docs/tweets/search/overview

[3] https://developer.twitter.com/en/docs/tweets/filter-realtime/overview

[4] https://developer.twitter.com/en.html

[5] http://twitter4j.org/en/

[6] http://www.tweepy.org/

[7] https://github.com/bear/python-twitter/

Table 2
Features and their objectives based on different factors.

Factors	Objective	Related features
Author of the tweet	To define author's (global) influence on his tweet's retweetability	Author's number of followers and followees, age of author's account, number of tweets from the author, tweet frequency (per day) of the author, author's number of tweets favored by others, language of the author, ratios of author's tweets retweeted by others, ratios of author's tweets received replies, and whether the author is a verified user or local elite
User of the tweet	To include user-author relation on a tweet's retweetability	User's recent communication with the author, user's social relation with the author, user mentioned in the tweet, author's influence on user (friend's influence)
	To define user-author interest similarity on a tweet's retweetability	User's interest similarity with author's interest.
	To include impact of user's activity on his retweeting probability	User's average retweet per day, tweeting likelihood of the day (hour), tweeting steadiness, and number of status messages, tweeting likelihood of the hour (to find retweeters), URLs/hashtags/mentions per day in postings
	To include the impact of user's personality on his retweeting activity	User's Big 5 and their 30-sub dimensional personality scores based on his postings
Content of the tweet	To include the impact of term distribution in tweet on its retweetability.	Presence/absence of hashtag, URLs, emoticons, positive/negative words, punctuation marks, username, first person pronoun, second person pronoun, third person pronoun, language, length of tweet.
	To define the impact of tweet's topic on its retweetability	Topic of the tweet, novelty and virality of tweet's topic
	To define the impact of tweet's terms on its retweetability	Importance of terms in the tweet (using TF–IDF scores), hashtags and URLs in the tweets
	To include tweet's emotion/sentiment on its retweetability	Emotion/sentiment reflected by the tweet, emotional divergence indicated by the tweet

Many features related to users have been explored for the purpose of retweet prediction, retweeter prediction, and tweet recommendation. Features that are used to measure the user-author relation include user's recent communication with the author, interest similarity with the author, social relation with the author, and whether user is mentioned in the tweet [13,39,55]. A user's interest profile can be derived from his past postings. Commonly used profiles include bag-of-word profile using the Term Frequency–Inverse Document Frequency (TF–IDF) weights of the words [37,59,60], hashtag-based profile [59], and entity-based profile [59]. TF–IDF technique finds scores/weights for terms in user's tweets based on their importance in distinguishing the user from others. Thus, TF–IDF based profile has the capability to represent users uniquely. In case of hashtag and entity-based profiles, only distribution of hashtags and entities might not give much information because many users might use the same hashtag and entity. So, the preferred method is to check their weights (frequency) while creating user profile because frequency of using hashtags and entities might give more information about a user's preference/interest. Another constraint of using entity-based profile is to select efficient method to extract entities from the tweets. Performance of entity-based profile is quite dependent on the efficiency and accuracy of entity extraction methods. Researchers mainly used AlchemyAPI[8] to extract entities from user's tweets. Some research works use third-party knowledge base to create the user interest profile. Macskassy and Michelson [39] used Wikipedia's knowledge base to create user's topic of interest profile. They identified the entities from user's tweets and categorized them based on their category in Wikipedia page. They also matched the content of entities in this process to solve the problem of ambiguity. The categories of mentioned entities were used to define user interest profile.

According to the research ([57,62]), oftentimes, user's retweet decision is consistent with the similarity degree between the user profile and the content of the tweet. Different similarity measures such as cosine similarity and Jaccard similarity have been used to calculate the similarity between user profile and content of the tweet. The calculated similarity scores are then used as potential features for retweet prediction [13,56,59,60,62]. In case of cosine similarity measure, user profile and target tweet profile are defined by the vectors represented by the distribution of topics/terms/hashtags in user profile and in the target tweet respectively. Cosine similarity between user and target tweet profile can be found by Eq. (1). According to this similarity measure, user profile and target tweet profile are similar if these vectors share similar orientation and the angle between them is small.

$$Cosine\ Similarity = \frac{A \cdot B}{\|A\|\,\|B\|} \tag{1}$$

where, A is user profile vector and B is target tweet profile vector.

Jaccard similarity measure was used when the researchers define user interest as a set of terms derived from his past postings and target tweet was defined by the set of terms consisting the target tweet. The Jaccard similarity between the two sets is defined by Eq. (2).

$$Jaccard\ Similarity = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

where A represents the user profile and B represents the target tweet profile.

Since cosine similarity finds the match by computing the angle between user's and tweet's vector representation but Jaccard similarity only checks the presence of common terms in user profile and target tweet profile, it can be said that cosine similarity measure gives more refined result than Jaccard similarity measure.

Friends' influence on user can also be used as predictive features. Zhang et al. [64] used data from Weibo micro-blogging service.[9] They defined social influence locality as a function to measure how a user's retweet decision is influenced by his active neighbors (users who have already retweeted the target tweet). The designed social influence locality function was based on pairwise influence and structure influence. In case of pair-wise influence, they used Random Walk with Restart (RWS) method to calculate random walk probability for each active neighbor of the given

[8] http://www.alchemyapi.com/

[9] Chinese micro-blogging service, allows its users to repost the tweets similar to Twitter's retweeting service.

user to reach the given user following the network connection. In case of structural influence, they used a linear combination of the number of connected circles formed by the active neighbors.

In case of finding retweeters, influencing features are mainly related to user's activity, intention, and interest. Lee et al. [32] explored a large number of features in this regard. They defined user's activeness by features such as average retweets per day, tweeting likelihood of the day (hour), tweeting steadiness, number of status messages [32]. They also included personality scores derived from user's postings to describe the impact of user's personality on his retweeting activity.

4.2.3. Content of the tweet

Both explicit and implicit features related to the content of the tweet are used in retweet prediction models. Explicit content-based features are directly measurable such as presence/absence of hashtag, URLs, emoticons, positive/negative words, punctuation marks, username, first person pronoun, second person pronoun, third person pronoun [30,41,55]. Researchers also used language and length of the tweet as features for prediction model [13,44,55]. Implicit content-based features are not directly measurable. Tools or algorithms are needed to extract these features. Popular implicit content-based features related to tweet include topics, terms with their TF–IDF scores, topic novelty, topic virality, sentiment, emotional divergence, etc. [13,23,26,41,44,45,55,59,62].

Some of the algorithms and tools used to extract implicit features include Latent Dirichlet allocation (LDA) [7], Term Frequency–Inverse Document Frequency (TF–IDF) [33], Linguistic Inquiry and Word Count (LIWC) [43], SentiStrength [54], and AlchemyAPI. LDA is used to determine user's topics of interest or topics of tweets. It is a generative statistical method that considers each document as a collection of topics and finds the latent topics in the document [6,7]. The topic assignment of the document is an iterative process which checks and updates the topic assignment of each word in every document based on the following two criteria: how frequent the word occurs across topics, and how frequent the topic occurs in the document. Finally, the most appropriate topics are chosen for the document. There are a few tools and packages implementing LDA technique that can be used to find topics in a document, for example, Stanford Topic Modeling toolbox[10] and Mallet[11] (Machine Learning for Language Toolkit) topic modeling tool. When LDA is used to create topic-based user profile, a user's all past tweets are considered as a single document and LDA is used to find the topic distribution of that document [59]. When LDA is used to identify topic of tweet, the single tweet is considered as a document and LDA finds the distribution of topic for that document [64]. Since the original LDA model was developed for long document and might not work properly for short document like tweets, Zhao et al. [68] proposed Twitter-LDA which is an extension of original LDA. This extended version determines a single topic for a tweet. It is reasonable to assume that tweets are focusing on a single topic because of their short length. It is assumed that, there are T topics in Twitter where every topic is represented by a word distribution. There is also a word distribution for background model and topic distribution for every user. Since each tweet is generated by single topic and background model, in case of tweet generation process, a user first picks a topic based on topic distribution for user. Then words for the tweet are chosen one by one based on the selected topic or background model. This word selection process is directed by Bernoulli distribution.

TF–IDF is a statistical measure used to find out the importance of a word for a document in a collection of documents. In this method, the importance (or weight) of a word increases proportionally by its number of use in the document but is counterbalanced by its use in the whole corpus. This measure has been used by the researchers to generate user's bag-of-word profile which consists of his preferred words based on their TF–IDF weights. This profile can represent user's content-based interest.

Researchers use LIWC[12] technique to find different text-based features. It finds the percentage of words in a document/text which belong to more than 70 different categories. These categories include simple linguistic factors such as Word Count, first person pronoun, as well as factors which indicate affect and emotion such as positive or negative emotion. AlchemyAPI uses machine learning technique to perform text analysis tasks. It is used to find entities in a user's tweets [59,60]. Sentiments of tweets can be determined by SentiStrength method. SentiStrength is a lexicon-based approach which uses linguistic rules to find the positive and negative sentiment score of a tweet. Researchers also use LIWC [50] and dictionary-based approach [41] to find sentiments of tweets. Affective Norms of English Words (ANEW) is a dictionary [8] that gives numerical values of 1030 words for three attributes indicating emotions: valence, dominance, and arousal. Valence refers to the degree of goodness/pleasantness (from displeasure to pleasure) invoked by the word, dominance refers to the extent of dominance (from weakness to strength) denoted by the word, and arousal refers to the degree of arousal (from calmness to excitement) evoked by the word. The total values of these three attributes for a tweet were the summation of these values for each word in the tweet.

Though different approaches have been used to define sentiment of tweets, accurate detection of sentiments in tweets can be a little tricky because of the presence of informal words in tweets. Dictionary that is developed particularly for Twitter can potentially solve this problem. For example, SentiStrength is considered as a better detector than other general dictionary-based methods because it was developed to find sentiments from short informal text. Bravo-Marquez et al. [9] also extended general word-emotion lexicon developed by Mohammad and Turney [40] to include informal words used in Twitter and Firdaus et al. [18] used this lexicon to extract sentiment from tweets and achieved good results.

4.3. Prediction model

Retweet prediction models are normally developed using different machine learning techniques. The fundamental task is to utilize the extracted features to develop effective retweet prediction model. It could be considered as a typical classification task. So, the basic retweet prediction model consists of feature extraction step followed by machine learning step to classify a tweet as "to be retweeted" or "not to be retweeted" based on the extracted features. It is important to find right features that are useful for the prediction task and to find right learning algorithms to make accurate predictions. In this section, we discuss some strategies to build retweet prediction models.

Many retweet prediction models have been developed based on the aforementioned two-step process and usually the original machine learning algorithms are used as they are. Zhang et al. [64] defined functions to model social influence locality features. Social influence locality implies that user's retweeting behavior is influenced by close friends in the ego-network. They developed logistic regression classifier to build their prediction model using social influence locality features. Xu and Yang [59] developed a retweet prediction model where they created TF–IDF, LDA, hashtags, and entity-based user profile. Cosine similarities between

[10] http://nlp.stanford.edu/software/tmt/tmt-0.4/
[11] http://mallet.cs.umass.edu/topics.php
[12] http://liwc.wpengine.com/

user profile and the target tweets were used as content-based features of their prediction model. Using some author-user relation features, content-based features, and author-based features, they developed three different prediction models using three machine learning techniques: decision tree, support vector machine, and logistic regression. Vougioukas et al. [56] explored a wide range of author-based, user-based, and content-based features and used logistic regression method to build retweet prediction model. Through experiment, they identified 10 most effective features to create the prediction model. Can et al. [10] proposed retweet count model based on visual cues of an image that is linked to the tweet. Along with two low-level features such as color histograms (distribution of color intensities in the image) and GIST (set of perceptual dimensions), the researchers used object-based feature [35]. Object-based feature was a set of object detectors to detect 177 objects in the image. They used 3 different regression methods: linear, SVM, and Random Forest to built their retweet count model using different features.

Some research works design retweet prediction models based on their own prediction strategies and novel models are proposed. Macskassy and Michelson [39] built four models to find the probability of a tweet to be retweeted by a specific user. The first model is called general model in which a user will randomly retweet a tweet with higher probability on more recently seen tweet. The second model is recent communication model in which a user will retweet tweets from authors with whom he has recent communications. The third model is called on-topic model in which the probability of a tweet to be retweeted is high if its profile is similar to user's interest. And the last model is homophily model in which a user will retweet tweets from authors with similar taste. The objective of this research was to find out the most effective model which can predict a user's probability to retweet a tweet. On a dataset with 79k tweets, the proposed homophily model showed best performance followed by recency model, on-topic model, and finally general model. They also found that a user's retweet behavior is better predicted by multiple models instead of one.The retweet prediction model proposed by Huang et al. [25] measures a user's interest in following categories: technology, politics, life, sports, entertainment, health, travel, and finance. Then for retweet prediction, they computed the probability of the target tweet belonging to a final category; if this probability is greater than user's interest in that category, they predicted that the user is going to retweet the target tweet.

Researchers have also adapted and modified existing machine learning methods to make them more fit as retweet prediction models. Petrovic et al. [44] used different author-based and content-based features to design their model. They used Passive-Aggressive algorithm (PA) [15] based machine learning approach to design a model to predict streaming retweets. They customized the original prediction rule of PA algorithm to adapt the time-sensitive rules for retweeting (e.g., tweets containing a specific word might have more probability to be retweeted in the morning than in the evening). Zhang et al. [66] adapted Hierarchical Dirichlet Process (HDP) [53] to design a nonparametric statistical method for retweet prediction. They incorporated structural, textual, and temporal information in their proposed HDP model. First, they extended HDP to model author, structure and content information. In the model, for each followee, the probability of retweeting his postings was subjected to binomial distribution with beta error. Structural influence (influence from neighbours) for users was also modeled by Beta distribution. Content influence was modeled by hidden topics and HDP-based generative process finds the topic assignment of microblogs. In the retweet prediction phase, the weights of recent topics were increased to incorporate temporal information and the retweeting probability of microblogs was then calculated. Peng et al. [42] proposed Condition Random Field (CRF) based retweet prediction model. Assuming user's retweet decision is influenced by local and network factors, they chose conditional random field to find the retweet probability conditioned on features related to the target tweet and target user. The researchers were concerned about the conditional distribution of user decision given the new tweet and the user. In their proposed method, they modeled tweet's content influence as well as network influence on user's retweet decision. For content influence, they included similarity between tweet's content and user's interest, similarity between tweet's content and user's friends' or followees' interest, and similarity between global interest (determined based on all tweets and retweets in the dataset) and tweet. They also included URLs, hashtags, and mention-based features to model tweet's content influence on user. To define network influence, they used author-based features such as author's number of followers/followees, author's number of tweets/retweets; and author-user relationship based on common followers, followees, mentions, and retweets. They utilized retweet network's "small world" [4,58] nature to design an efficient graph partitioning algorithm to make their method suitable for large, complicated network. In case of small world network, the network graph is highly clustered, average path length (APL) between all pairs of nodes is small, and an individual is mainly influenced by a small number of his connections. Retweet network can be considered as small world network because retweets are spreading through the connections of users and these connections normally show clustering property. In Twitter, the APL between pairs of nodes is small, and retweet network can be defined by fraction of edges (portion of connections) which make the clustering structure of the network. Xu et al. [60] proposed a mixture latent topic model to explore user's retweet behavior. Assuming user's posting behavior is influenced by breaking news, posts from friends, and his intrinsic interest; the researchers extended the widely used author-topic model [46] to include the mentioned factors to build their proposed mixture topic model. Yang et al. [62] proposed a semi-supervised factor graph model to predict users' retweeting behavior based on factors such as user, message, and time.

Matrix factorization is an effective technique used by the researchers to design retweet prediction model. The fundamental task of matrix factorization technique is to factorize the observed user-message retweeting matrix $R \in \mathrm{R}^{M \times N}$ for M users and N messages into two low dimensional matrices $P \in M \times k$ and $Q \in N \times k$ such that product of P and Q approximates R. The main objective is to find the latent features k which defines the latent relationship between user and message. Jiang et al. [27] proposed message-clustering-based matrix-factorization models assuming that if messages are similar in observed space then they are similar in latent space as well. So, they extended the basic matrix factorization model by using clustering-based regularization term. Different content-based features were used to find the similarity between messages which was then used to define cluster of messages. Wang et al. [57] proposed two matrix factorization based retweet prediction models. They used strength of social relationship between users to generate objective function for user-based prediction model. Another prediction model was developed using content-based features. Finally, they fused both models based on their error rates. Hoang and Lim [23] represented retweets as three-dimensional tensors of authors, their followers, and tweets themselves. Then they proposed a tensor factorization model to derive three behavioral factors – topic-specific user virality, topic-specific user susceptibility, and topic virality. These factors were then used as features to predict user retweet actions.

Nowadays deep learning methods become popular for their efficiency and ability to learn optimal features automatically. Zhang et al. [67] proposed a retweet prediction model using attention-based deep neural network. In this model, they used convolutional neural network to encode content of the tweet and attention-based neural network to encode the attention interest of the user. Similarity between user's attention interest and tweet was also computed. They encoded each user and author with continuous vector. Finally, a concatenation layer was used to produce a hidden state using these vectors and a fully connected Softmax function was used for retweet prediction.

Researchers working with tweet recommender systems consider retweet as a mechanism to identify user's preference. These research works predict retweets to check which tweets are retweeted or preferred by the user. Uysal and Croft [55] explored user's retweeting behavior to filter tweets for individual users. They used author-based, user-based, and content-based features to develop a decision-tree-based classifier to classify tweets as retweetable or not for a specific user. They used learning to rank method to rank incoming tweets to develop tweet recommendation list for a user. Lu et al. [36] built a tweet recommender system by ranking incoming tweets based on their similarity with user profile. In this research, user's retweets are considered as relevant tweets for recommendation. Their novel approach to create user profile using Wikipedia concept graph showed better performance for tweet recommendation compared to models with TF–IDF based profile. Chen et al. [13] developed a personalized tweet recommendation method assuming retweets as a measure of user's interest and authors of retweets as a measure of social relationship. They included topic level user interest and user-author relation features to build collaborative-ranking based tweet recommender system.

To design model to predict retweeters, Luo et al. [37] used SVM^{Rank} method to rank potential retweeters based on their probability to retweet. Lee et al. [32] explored a wide range of features and built different prediction models to compare their results using the following machine learning techniques: Random Forest, Naïve Bayes, Logistic, SMO, and AdaBoostM1. They found that Random Forest based model performed best in predicting potential retweeters.

Retweet prediction would become more accurate if data from active users are used to learn and train the model. Social networks have many active as well as inactive users. Inclusion of inactive users' data might not have accurate contribution in prediction model. So, finding out active users is an important step when collecting the data. On the other hand, most of the time, retweet prediction model needs both positive and negative examples. The positive examples are a person's retweets. Negative examples are the tweets which are posted by user's followees and appear in user's timeline, but not retweeted by the user. The reasons for which a user is retweeting a target tweet is somehow understandable and derivable. However, the reasons for which a target tweet is not retweeted by the user is tricky to find out and they often are decided by many unseen features. We cannot just say that the user did not like the tweet or the user is not interested in the topic of the tweet. A user might not retweet a target tweet for some concealed reasons such as he might not see the tweet, he might not be active during the posting time of the tweet, he might not be in the frame of mind to spread any information. Researchers handled these issues in different manners. Some of the approaches to choose negative examples are described below:

- Zhang et al. [64] predefined 6 timestamps to define negative instances. If a tweet is not retweeted by the user within any of the mentioned timestamp (selected randomly), then they considered it as negative instance.
- Zaman et al. [63] used one-hour time window to see if a tweet is retweeted by the user within one hour of its posting time. If the tweet is not retweeted by the target user within one hour, then that tweet was considered as negative instance.
- Uysal and Croft [55] selected active users based on the following three criteria: he has 10–1000 friends/followers, tweets 1–200 times a week, and tweets more than 10 times. By considering only the active users they eliminated the uncertainty to some extent about the user not seeing the non-retweeted tweet.
- Xu and Yang [59] first selected seed/active users who have 100–3000 followers and followees, are listed 1–50 times and have 10–200 tweets per week. Then they considered a non-retweet as a negative example if the author (of that tweet) was seed user's social friend, if it was published within 3 h before the seed user retweeted any other tweet and if that tweet was not retweeted by the seed user.
- Li et al. [34] did research on predicting personality traits of Weibo (Chinese version of Twitter) users and they selected active users using the following steps: (i) Included users having more than 532 micro-blogs after registration, (ii) Included users having average count of micro-blog update per day between 2.84 and 40, (iii) excluded users who had no update in last three months, and (iv) excluded users who only updated in the first month after registration.
- User activity can vary with time. Generally speaking, users are more active during daytime than night. A tweet posted during day time might get more retweet (spread broadly) than a tweet posted during night. Neglecting this phenomenon might give wrong information regarding the popularity and interestingness of tweets. Gao et al. [19] introduced a new time notation *Weibo time* and a new time mapping process to handle the effect of activity variation with time in retweet prediction. For *Weibo time,* they measured the time by the number of users' posting on Weibo instead of wall time in seconds. So, wall time was mapped to corresponding *Weibo time*.
- Sometimes retweets happen just because of the retweet request from friends (followee).This type of retweets is not very helpful to find out the latent feature causing a tweet being retweeted. Yang et al. [61] proposed some measures which can be used to handle this type of retweets from dataset.

4.4. Evaluation

The last important step of retweet prediction or tweet recommendation model is to evaluate the performance of the model. In case of prediction, dataset is divided into training and testing set; the model is trained using training dataset and tested using testing dataset. Machine learning techniques analyze the training data (instances with observed outcomes) and learn reasoning to find the outcome for the instance. Testing dataset (instances with unknown outcome) is used to find performance of model on unseen data. Standard approach is to use 70−90% data as training samples and the rest for testing. Another popular approach to evaluate the learning model is k-fold cross validation. In this technique, the dataset is divided into k equal subsets then $k-1$ subsets are used as training data and the remaining subset is used as testing data. This process is repeated k times such that every subset is used exactly once as test data. Finally, the results from all iterations are averaged to get the final result

Performance of the learning model is evaluated using different metrics. Many different evaluation metrics are available; researchers pick the one suitable for their work. In case of predictive model, researchers usually pick the following metrics: accuracy, precision, recall, and $F1$-score. Accuracy refers to the fraction of correctly classified instances to the total number of instances as defined in Eq. (3). Precision refers to the fraction of the classified positive instances that are true positives as defined in Eq. (4). Recall refers to the fraction of the positive instances that are correctly identified as positive as defined in Eq. (5). $F1$-score is the weighted average of precision and recall as defined in Eq. (6).

$$Accuracy = \frac{t_p + t_n}{t_p + f_p + t_n + f_n} \tag{3}$$

$$Precision = \frac{t_p}{t_p + f_p} \tag{4}$$

$$Recall = \frac{t_p}{t_p + f_n} \tag{5}$$

$$F1 - score = 2 * \frac{precision * recall}{precision + recall} \tag{6}$$

where, $t_p =$ number of positive instances classified as positive

$t_n =$ number of negative instances classified as negative
$f_p =$ number of negative instances classified as positive
$f_n =$ number of positive instances classified as negative

Sometimes a single metric is not enough to define the performance of a predictive model depending on the type of the problem. For example, accuracy only might not be able to show the true performance of a model when the dataset is vastly imbalanced. In this case, a model might fail to classify any of the instances from a class with only a small number of instances but would still show high accuracy overall. For retweet prediction problem, when the number of positive instances (retweets) and negative instances (non-retweets) are hugely unequal, researchers preferred precision, recall, or $F1$-score either by themselves or in addition to accuracy. In most of the cases for retweet prediction problem, researchers used precision to show model's performance in predicting retweets (true positive instances) correctly out of all its positive predictions and recall to define model's performance in predicting retweets correctly out of all the retweets (true positive instances). Precision only may not give any good idea about model's performance in predicting all retweets correctly; on the other hand, recall only may not give a good idea about model's behavior in predicting non-retweets as retweets. So, researchers use both metrics to show model's overall performance. Researchers also use $F1$-score to define performance of the model because with the harmonic mean of precision and recall, it shows the balance between them and helps to select a standard model.

Another performance measure, Mean Average Precision (MAP) is used to evaluate the performance of ranking. MAP is a preferred metric when not only the prediction or recommendation of relevant item but also their rank is important. MAP is calculated as shown in Eq. (7) [13].

$$MAP = \frac{\sum_{n=1}^{N} AvgP(n)}{N} \tag{7}$$

where N is the number of users, and $AvgP(n)$ is the average precision for user n. $AvgP(n)$ is calculated using Eq. (8) [13].

$$AvgP = \frac{\sum_{k=1}^{K} (P(k) \times rel(k))}{R} \tag{8}$$

where R is the number of retweeted tweets for the given user; K is the number of recommended tweets; $P(k)$ is the precision at rank k calculated by considering only top k results; $rel(k) = 1$ if the tweet at rank k is retweeted by the user, 0 otherwise.

Root Mean Square Error (RMSE) is a measure which finds the difference between actual value and predicted value, and has been used by Can et al. [10]. In this case, errors are squared and averaged before taking their square root. It gives more weight to the large errors [12]. RMSE can be calculated using the following equation:

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N} (A_n - P_n)^2}{N}} \tag{9}$$

where N is the number of instances in the dataset; A_n is the actual and P_n is the predicted value for nth instance.

To evaluate the performance of ranked recommendation list, Lu et al. [36] used the following measures: recall-at-k, precision-at-k, and average hit-rank. For a user u, these metrics can be calculated using the below equations.

$$recall - at - k = \frac{hit}{n_T(u)} \tag{10}$$

$$precision - at - k = \frac{hit}{k} \tag{11}$$

$$average\ hit - rank = \frac{1}{n_T(u)} \sum_{i=1}^{hit} rank_i \tag{12}$$

where hit is the total number of relevant tweets in the top k tweets in the recommendation list, $n_T(u)$ is the total number of relevant tweets in the test dataset, and $rank_i$ is the position of the relevant tweet in the recommendation list.

Area Under Curve (AUC), Area Under Precision Recall Curve (AUPRC), and Receiver Operating Characteristic (ROC) curve are the evaluation techniques we can use to visualize the performance of models. ROC curve plots true positive rate against false positive rate to show how the number of correctly classified positive instances varies with the number of incorrectly classified negative instances [16]. The goal for a model is to be at the upper-left corner of the plotting space which indicates lower false positive and higher true positive rate. AUC is the area under the ROC curve. Lee et al. [32] used AUC as their evaluation metric because AUC can indicate the model's performance in predicting both positive and negative instances in spite of imbalance class distribution in the dataset [17]. AUPRC is the area under precision–recall curve where precision–recall curve is plotted using precision against recall. AUPRC is the single number giving information about precision and recall where higher value defines better performance. Precision–recall curve is an alternative of ROC curve which should be used when there is huge skew in the class distribution [16].

In this survey paper, we are not planning to compare the results of different research papers on retweet prediction. Since every paper used a different dataset with data collected in different time frame, it would not be appropriate to compare their results directly. In Table 3, we have listed the metrics used in different research works for evaluating their models.

4.5. Performance of different prediction approaches

In this section, we discuss about the performance of different prediction approaches. As we mentioned earlier, due to the use of different datasets, direct comparison of these approaches in terms of their reported performance values will not be accurate. Therefore, we just discuss their individual performance, comparison of their performance with baselines as they reported in terms of evaluation metrics and their conclusion from the experiment. Here, if the paper listed the actual values of the used evaluation metrics, we report the values; but if the paper only showed its performance comparison in a chart or figure, we explain their relative performance because it is hard to find actual values from figures. The objective of this section is to shed light on the performance of different prediction models to help new researchers gain insights on how these models perform comparing with each other and identify a suitable model or promising new direction for their particular problem.

Peng et al. [42] proposed a graph partitioning method to capture network relations that are then used in the retweet prediction model. They compared their model with an un-partitioned one and also explored the model in following different settings: (i)

Table 3
Evaluation metrics used indifferent research papers.

Research	Precision	Recall	*F*1-score	Accu-racy	AUC	MAP	AUP-RC	Recall-at-*k*	Precision-at-*k*	Avg. hit rank	RM-SE	ROC
[42]	x	x	x									
[64]	x	x	x	x								
[66]	x	x	x									
[59]	x	x	x									
[62]	x	x	x									
[60]	x	x										
[23]							x					
[25]	x											
[57]	x	x	x									
[27]	x	x	x	x								
[56]	x	x	x									
[67]	x	x	x									
[55]	x	x	x									
[36]								x	x	x		
[13]						x						
[37]						x						
[32]			x		x							
[44]			x									
[10]											x	
[41]												x

with no edge and random partition (*N-Rd*), (ii) with explicit edges formed by users' following relationship and random partition (*E-Rd*), and (iii) network partition using their proposed greedy iterative partition method (*E-Min*). They showed that partitioned settings always give better prediction accuracy in terms of precision, recall, and *F*1-score than the baseline method (no partition) and prediction model with *E-Min* setting outperforms the other settings. Zhang et al. [64] showed that their proposed retweet prediction model (*LRC-Q*) based on social influence locality with precision 0.619, recall 0.927, *F*1-score 0.742, and accuracy 0.678 performed better (+0.184 in terms of recall and +0.031 in terms of *F*1-measure) than baseline model (*LRC-B)* using traditional features including gender, verification status, number of followers, following relationships, historical tweets, instantaneity, and topic propensity. Model (*LRC-BQ*) defined by the combination of influence locality function and traditional features gives a little improvement over the proposed model. So, it can be said that their social influence locality function alone gives better predictive power than other traditional user and content features.

Zhang et al. [66] incorporated author, social, and content information in Hierarchical Dirichlet Process to build their retweet prediction model *ASC-HDP*. They compared ASC-HDP with some previous models such as *LRC-BQ* with *F*1-score 0.589 [64], Naïve Bayes with *F*1-score 0.456, and SVM$^{\text{rank}}$ with *F*1-score 0.467. They also compared the proposed *ASC-HDP* with HDP models incorporating only content (*C-HDP*) or social-content (*SC-HDP*) or author-content information (*AC-HDP*), and *ASC-LDA* (Used LDA instead of HDP to find topics of tweets). They found that *AC-HDP* with precision 0.809 and *SC-HDP* with recall 0.727 performed better than *ASC-HDP* in terms of precision and recall, respectively. The proposed *ASC-HDP* model outperforms all other models in terms of *F*1-score (0.730). Xu and Yang [59] built three retweet prediction models based on user-author relationship, content, and author features using three machine learning techniques: Decision Tree (C4.5), Support Vector Machine (SVM) and logistic regression. The decision tree model performed better (*F*1-score: 0.832) than SVM and logistic regression models when using the same features. In case of predicting pair-wise retweeting behavior, factor-graph-based model proposed by Yang et al. [62] did not outperform the baseline models (Linear SVM, and L1-regularized logistic regression) because their proposed method discriminates instances based on least sum of square error factor but the baselines represent the samples in feature space more accurately which made them more discriminable. The proposed method performed better (*F*1-score: 0.3252) than the baselines in predicting spread of messages (range of message propagation) because it considers the entire graph as a whole and captures important information from users' followers as well as followees. In terms of recall, SVM performed better than the proposed method to predict message propagation because their method predicts that messages with short augmented retweeting tree will not be propagated. For predicting retweet, Petrovic et al. [44] achieved *F*1-score 0.466 which is significantly higher than the baselines Random (*F*1-score: 0.119) and Majority (*F*1-score: 0.127) where Random considers tweets will be retweeted randomly and Majority considers everything will be retweeted. For predicting retweet count, baseline along with visual cues related features proposed by Can et al. [10] gave RMSE 1.703, 1.559, and 1.297 in log-scale (5.489, 4.753, and 3.659 in linear scale) for linear, SVM, and random forest regression model. The reported error is smaller than baseline when compared with model using only basic features such as presence of hashtags, followers count, friends count, age of account, status count, favorite count, etc.

Xu et al. [60] proposed mixture topic retweet prediction model and showed that the proposed one with precision 0.172 performed better than the baseline models (TF–IDF based, entity based, and author-topic based) in terms of precision. In terms of recall, the proposed model alone did not perform well in predicting retweets when compared with baseline model using previously used features, but inclusion of these previously used features together with the proposed model increased its recall to 0.412 from 0.378. Tensor factorization retweet prediction model proposed by Hoang and Lim [23] performed better than the baseline models (developed based on user specific retweetable topics, global retweetable topics, combination on user specific and global retweetable topics, and collaborative topic regression) in terms of AUPRC. Huang et al. [25] showed that their model based on user interest matrix showed more stable and precise result than baseline models (text-similarity algorithm and co-terms algorithm) in terms of precision. Wang et al. [57] compared their proposed nonnegative matrix factorization retweet prediction model with a few baseline models (SVM, Naïve Bayes, Back-Propagation Neural Network, Random Forest, Decision Tress) and showed that the proposed method with *F*1-score 0.79 performed better than the baselines in terms of performance and efficiency (shorter execution time). Jiang et al. [27] compared their model with multi-dimensional nonnegative matrix factorization model (MNMFRP) proposed by Wang et al. [57] and social influence locality model (LRC-BQ) proposed by Zhang et al. [64] and showed that performance of their proposed model ($F1$-score: 0.82) is better than performances of the past models ($F1$-score: 0.79 for MNMFRP, 0.73 for LRC-BQ) for predicting retweets. Vougioukas et al. [56] showed that their proposed model has $F1$-score around 0.9 using only 10 features. These features include similarity to tweets retweeted by the user in previous week, similarity to tweets previously posted by the author, influence of author, number of tweets posted by the author, similarity to tweets retweeted by the user, number of times the tweet has been retweeted by neighbors, whether the author is a neighbor of the user, whether the user retweeted the author before, similarity to user's posted tweets, and whether the author ever mentioned the user.

Zhang et al. [67] compared their deep neural network retweet prediction model (SUA-CNN) with SVM models, random model, HDP model [66], Convolutional neural network (CNN) model, CNN with user embedding (U-CNN) and CNN with user-author embedding (UA-CNN) models. They showed that user embedding, user-author embedding, and embedding of similarity score and user interest each can improve the prediction performance significantly when compared to baseline models, but the best performance was obtained by the model which integrates all the information. In terms of $F1$-score, the proposed model (SUA-CNN) with $F1$-score 0.721 showed better performance than HDP model with $F1$-score 0.656.

For tweet recommendation, Uysal and Croft [55] predict retweetability of tweets to create tweet recommendation list for the user. They showed that the proposed method performed the best ($F1$-score: 0.724) with all features when compared to methods using user, content, and author-based features separately. Lu et al. [36] compared their tweet recommendation model using Wikipedia concept graph with TF–IDF model. They showed that the proposed model gave better performance than TF–IDF model because it used Wikipedia's vast knowledgebase to create effective user profile. Chen et al. [13] compared their method with some baseline methods (Chronological, Retweeted times, Profiling, LDA, RankSVM, and JointMF models) and showed that MAP (0.7627) of the proposed method is significantly higher than the baseline ones.

Luo et al. [37] and Lee et al. [32] worked on finding out retweeters. Method proposed by Luo et al. [37] showed better performance with MAP 0.087 than random baseline with MAP 0.0217 and PRT baseline with MAP 0.069 (ranking followers by the number of times they retweeted the user before). Lee et al. [32] showed that the proposed model had $F1$ scores 0.692 (public safety dataset) and 0.785 (bird flu dataset); AUC 0.954 (public safety dataset) and 0.815 (bird flu dataset)for prediction of retweeter.

From the performances of the discussed prediction models, we can see that, this research area still has a lot of room for improvement in terms of accuracy. Sometimes, the proposed prediction model or strategy of feature extraction might not be the better one compared to some existing ones, but it points out a new potential research direction and it might be worthwhile to have some further investigation along the line. For example, social influence locality retweet prediction model proposed by Zhang et al. [64] showed that the proposed method to capture social influence has good predictive power than many traditional features. With combination of traditional features, it shows a little improvement. There is still scope to improve it further by exploring other effective traditional features with the proposed social influence locality function. The proposed strategy to design social influence shows a new path to apply network influence in prediction retweet decision. Wang et al. [57] and Jiang et al. [27] both used matrix factorization technique to build their retweet prediction model. But the model proposed by Jiang et al. [27] was more cost effective (computationally feasible) and outperformed the model proposed by Wang et al. [57]. In this case, although one model outperformed the other according to their experiments, it would be useful to keep both as potential prediction model to use because there could be different scenarios where one of these will be better suited. In the past years, many retweet prediction approaches have been proposed by the researchers, but selecting the best one among them is a tricky task because of the frequent changing nature of retweet network. Design of an accurate retweet prediction model is quite dependent on network as well as its users' characteristics. Sometimes, it is always good to try some simple models first. For example, with some simple features, model used in Vougioukas et al. [56] achieved highly accurate results.

5. Retweet for information diffusion

Data from social networks is a great source of information. This information can be more useful when it can reach to appropriate users. In case of Twitter, its retweeting feature provides an important mechanism for information diffusion. Retweet has been used to determine trend and popularity of event [22,24,65]. It was assumed that if an event gets relatively more tweets than retweets then the event might not last long on Twitter and in turns might not become popular. Gupta et al. [22] checked the ratios of retweets at consecutive hours to capture the changes in popularity of the event over time. Zhang et al. [65] used retweet to predict Twitter trend and Hong et al. [24] explored retweets as a measure to find popular messages. Research showed that retweets are used vastly by the users to spread disaster related useful information during emergency [31,48]. Kogan et al. [31] explored the retweeting pattern of geo-vulnerable users during hurricane Sandy in year 2012. After checking the retweeting activity of geographically vulnerable users during four different time frames (before, during, short-after, and long after the disaster), they found that the size (based on nodes and edges) of retweeting network during the disaster is bigger than the size before and after the disaster. They also determined the important nodes of each time-sliced retweet network using PageRank method and found that local government authorities and media are the most important nodes (most retweeted) in Geo-During network (formed by the retweets of geographically vulnerable users). Starbird and Palen [48] also explored the use of retweet during two emergency situations – "Red River Flooding (USA), 2009" and "Oklahoma Fires (USA), 2009". This research also indicated that during emergency, tweets of local users, media and service organizations as well as tweets containing emergency related terms were retweeted more.

Contribution of retweet to engage remote individuals in 2011 Egyptian political uprising has been explored by Starbird and Palen [49]. During this event, protesters, journalist, media on the ground used to post movement-related information which were vastly retweeted by others to spread the information. This study showed that some tweets were not authored by people from Cairo but got high number of retweets and revolution-related metaphors were highly propagated in the Twitter. These findings clearly indicate the use of retweets by the users from Twitter in support of revolution. Sanjari and Khazraee [47] explored information diffusion using Twitter during 2013 Iranian Presidential election. They showed that Iranian Twitter celebrities are most influential during election based on their retweet network. On the other hand, discussion about Iran in English was dominated by journalists and official media. Stieglitz and Dang-Xuan [50] identified that political discussion took place in Twitter through retweet and direct message functionality and few highly active users are most influential whose tweets were retweeted vastly. In this study they found that leftists are the most influential users (got highest retweets). Since positive or negative sentimental tweets have high retweetability, tweets containing political sentiments are retweeted more and thus influence political decision.

Retweets are not just a method of information diffusion; they can be considered as a measure of trust between author and user. Trust is an important factor in social network to assess the credibility of information as well as to understand the flow of information in social network. A user retweets an author's tweet when he has trust on that author. Adali et al. [3] showed that user explicitly retweeting an author's tweets is a reliable measure of trust between two users.

Retweet is an excellent medium of information diffusion, especially during the time of emergency. However, because of its easy availability, during crisis time, along with important information some rumours can also be spreading through this mechanism. Abdullah et al. [1] did a research to explore user's actions and decision-making behavior on retweeting which helped them explain the reasons for spreading rumours at crisis time. According to the survey conducted by the authors, when users just retweet a message finding it important or marking it as favorite, there is greater chance to spread inaccurate information. But, when users search for further information regarding the tweet (the current situation); there is less chance to spread rumours at the time of disaster. Acar and Muraki [2] also suggested that, use of official hashtags and provision to trace the originality of information would be effective solution to handle misinformation as well as might increase the reliability of information.

6. Discussion and challenges

In this survey paper, we have focused on retweet feature provided by Twitter as an information diffusion mechanism. Retweet prediction is an important area of research due to its importance in understanding user's intention and approach in dispersing information. We have described the basic steps of retweet prediction research, included the strategies used by the researchers to build and evaluate effective retweet prediction models, and discussed some challenges and respective measures used by the researchers to handle those challenges. A prior knowledge about these issues would be helpful for the future research. Our paper could serve as a guide for the people who want to conduct research in this area. We have included the analysis on user's retweeting behavior to understand user's retweeting activity. It is not easy to find out the actual reasons behind a person's actions. But past researches have successfully discovered some reasons which trigger user's retweeting decision. As an easily available and efficient information diffusion mechanism, retweet is used widely to spread information during disaster or emergency cases. We have also discussed some of these works in the paper.

Although there have been a lot of research efforts on retweet prediction, the prediction accuracy is far from perfect. Some of the major challenges in this area are explained as follows. First, retweets are very short (140 characters). Retweet itself carries only a small amount of explicit information due to its short length. Therefore, the challenge is to extract useful latent information from retweet. Second, retweet network is vast. Since neighbors (followees/friends) have good impact on user's retweeting behavior, analyzing the retweet network would be helpful. But the size and complexity of this network often make the task difficult. Third, user's fraudulent behavior is a challenging factor when considering retweeter as information spreader; especially during emergency situation it can create panic. Retweet fraud can also create false product advertisement which might lead to wrong product review based on Twitter data. Researchers are working on this issue [20,28], but research papers we have included here did not deal with user's fraudulent behavior in case of exploring retweet as information diffusion method. Fourth challenge is recommending tweet for inactive users. Most of the past research did not include inactive user data because there is not enough data available for inactive user to find his preference on tweeting/retweeting decision. Researchers also assume that inclusion of inactive user's data might lower model's accuracy. However, inactive user can be potential retweeter. A user not posting anything does not really imply that he is not checking Twitter. It might be possible that he is not interested in posting anything or he is not finding anything interesting to tweet/retweet. An effective tweet recommendation according to his interest might be able to make him post tweets in Twitter. Inactive user's activeness can be checked from his changing friend list. If he is not posting anything but adding new friends, then it indicates that he is following friends in Twitter. Data from his friends can be good source of information about an inactive user's preference which can be used to recommend tweets for him. Good recommendation has the ability to transform a user from inactive to active. Data from third party such as other social networking sites or from his online activity might also be helpful to create inactive user's interest profile.

The last but not the least challenge is to identify the latent attitude or behavior that controls user's retweet decision. This is probably the most challenging task because human behavior is a complex puzzle that cannot be solved entirely. Researchers have successfully identified some factors that influence or trigger user's retweeting activity, such as personality, emotion, sentiment, interest similarity. But there are still many unexplored latent factors such as values, beliefs, views on topics. Also, some users' interest is static whereas for other users, it changes frequently. Finding out the dynamics inside these changing interests, their evolution patterns, and reasons behind the change are important for designing the personalized retweet prediction model. Although Twitter data provides information about user's online activities, it could be impossible or unrealistic to find out all the situational and psychological factors affecting user's retweet decision without the knowledge of the real people behind their online identities and their real day-to-day lives.

7. Conclusion

In today's world, information is regarded as the weapon for solving many complex problems. But, conveying proper information to its potential users is a complicated task. Online social networks provide good opportunity to spread information more easily and reach out to a wider range of users. The main challenge for researchers is to find the reasons and methods to diffuse the information in an effective manner through these online social networks. In this survey paper, we have included research papers that worked on finding the underlying reasons behind user's intention in spreading information as well as proposed models to predict retweetable tweets and target/interested users. We have also discussed some open challenges in this area; exploring these unsolved issues would lead the researchers to build more useful and effective information diffusion model using retweets.

Acknowledgment

This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca)

Funding

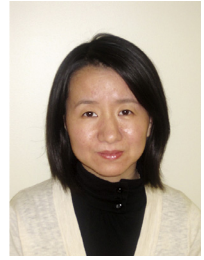
This work is partially sponsored by Natural Science and Engineering Research Council of Canada (grant 2015-05555).

References

[1] N.A. Abdullah, D. Nishioka, Y. Tanaka, Y. Murayama, User's action and decision making of retweet messages towards reducing misinformation spread during disaster, J. Inf. Process. 23 (1) (2015) 31–40.
[2] A. Acar, Y. Muraki, Twitter for crisis communication: lessons learned from Japan's tsunami disaster, Int. J. Web Based Commun. 7 (3) (2011) 392–402.
[3] S. Adali, R. Escriva, M.K. Goldberg, M. Hayvanovych, M. Magdon-Ismail, B.K. Szymanski, W.A. Wallace, G. Williams, Measuring behavioral trust in social networks, in: Proceedings of the 2010 IEEE International Conference on Intelligence and Security Informatics, IEEE, 2010, pp. 150–152.
[4] L.A. Adamic, The small world web, in: Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries, Springer-Verlag, 1999, pp. 443–452.
[5] J. Berger, Arousal increases social transmission of information, Psychol. Sci. 22 (7) (2011) 891–893.
[6] D.M. Blei, Probabilistic topic models, Commun. ACM 55 (4) (2012) 77–84.
[7] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022.
[8] M.M. Bradley, P.J. Lang, Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings, The Center for Research in Psychophysiology, University of Florida, 1999 Technical Report C-11999.
[9] F. Bravo-Marquez, E. Frank, S.M. Mohammad, B. Pfahringer, Determining word-emotion associations from tweets by multi-label classification, in: Proceedings of the 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), IEEE, 2016, pp. 536–539.
[10] E.F. Can, H. Oktay, R. Manmatha, Predicting retweet count using visual cues, in: Proceedings of the Twenty-Second ACM International Conference on Conference on Information & Knowledge Management, ACM, 2013, pp. 1481–1484.
[11] M. Cha, H. Haddadi, F. Benevenuto, K.P. Gummadi, Measuring user influence in Twitter: the million follower fallacy, in: Proceedings of the 2010 International AAAI Conference on Weblogs and Social, ICWSM'10, 2010.
[12] T. Chai, R.R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, Geosci. Model Dev. 7 (3) (2014) 1247–1250.
[13] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, Y. Yu, Collaborative personalized tweet recommendation, in: Proceedings of the Thirty-Fifth International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2012, pp. 661–670.
[14] G. Comarela, M. Crovella, V. Almeida, F. Benevenuto, Understanding factors that affect response rates in twitter, in: Proceedings of the Twenty-Third ACM Conference on Hypertext and Social Media, ACM, 2012, pp. 123–132.
[15] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, Y. Singer, Online passive-aggressive algorithms, J. Mach. Learn. Res. 7 (Mar) (2006) 551–585.
[16] J. Davis, M. Goadrich, The relationship between precision–recall and ROC curves, in: Proceedings of the Twenty-Third International Conference on Machine Learning, ACM, 2006, pp. 233–240.
[17] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (8) (2006) 861–874.
[18] S.N. Firdaus, C. Ding, A. Sadeghian, Topic specific emotion detection for retweet prediction, Int. J. Mach. Learn. Cybern. (2018) 1–13 https://doi.org/10.1007/s13042-018-0798-5.
[19] S. Gao, J. Ma, Z. Chen, Modeling and predicting retweeting dynamics on microblogging platforms, in: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, ACM, 2015, pp. 107–116.
[20] M. Giatsoglou, D. Chatzakou, N. Shah, C. Faloutsos, A. Vakali, Retweeting activity on Twitter: signs of deception, in: Proceedings of the 2015 Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, Cham, 2015, pp. 122–134.
[21] A. Guille, H. Hacid, C. Favre, D.A. Zighed, Information diffusion in online social networks: a survey, ACM SIGMOD Record 42 (2) (2013) 17–28.
[22] M. Gupta, J. Gao, C. Zhai, J. Han, Predicting future popularity trend of events in microblogging platforms, in: Proceedings of the 2012 American Society for Information Science and Technology, 49, 2012, pp. 1–10.
[23] T.A. Hoang, E.P. Lim, Retweeting: an act of viral users, susceptible users, or viral topics? in: Proceedings of the 2013 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, 2013, pp. 569–577.
[24] L. Hong, O. Dan, B.D. Davison, Predicting popular messages in twitter, in: Proceedings of the Twentieth International Conference Companion on World Wide Web, ACM, 2011, pp. 57–58.
[25] D. Huang, J. Zhou, D. Mu, F. Yang, Retweet behavior prediction in Twitter, in: Proceedings of the Seventh International Symposium on Computational Intelligence and Design (ISCID), 2, IEEE, 2014, pp. 30–33.
[26] M. Jenders, G. Kasneci, F. Naumann, Analyzing and predicting viral tweets, in: Proceedings of the Twenty-Second International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, 2013, pp. 657–664.
[27] B. Jiang, J. Liang, Y. Sha, L. Wang, Message clustering based matrix factorization model for retweeting behavior prediction, in: Proceedings of the Twenty-Fourth ACM International on Conference on Information and Knowledge Management, ACM, 2015, pp. 1843–1846.
[28] M. Jiang, P. Cui, C. Faloutsos, Suspicious behavior detection: current trends and future directions, IEEE Intell. Syst. 31 (1) (2016) 31–39.
[29] E. Kim, S. Gilbert, M.J. Edwards, E. Graeff, Detecting sadness in 140 characters: sentiment analysis of mourning Michael Jackson on Twitter, Web Ecol. 3 (2009) 1–5.
[30] J. Kim, J. Yoo, Role of sentiment in message propagation: reply vs. retweet behavior in political communication, in: Proceedings of the 2012 International Conference on Social Informatics (Social Informatics), IEEE, 2012, pp. 131–136.
[31] M. Kogan, L. Palen, K.M. Anderson, Think local, retweet global: retweeting by the geographically-vulnerable during hurricane Sandy, in: Proceedings of the Eighteenth ACM Conference on Computer Supported Cooperative Work & Social Computing, ACM, 2015, pp. 981–993.
[32] K. Lee, J. Mahmud, J. Chen, M. Zhou, J. Nichols, Who will retweet this? Detecting strangers from twitter to retweet information, ACM Trans. Intell. Syst. Technol. (TIST) 6 (3) (2015) 31.
[33] J. Leskovec, A. Rajaraman, J.D. Ullman, Mining of Massive Datasets, Cambridge University Press, 2014.
[34] L. Li, A. Li, B. Hao, Z. Guan, T. Zhu, Predicting active users' personality based on micro-blogging behaviors, PloS One 9 (1) (2014) e84997.
[35] L.J. Li, H. Su, F.F Li, E.P. Xing, Object bank: a high-level image representation for scene classification and semantic feature sparsification, in: Proceedings of the 2016 Advances in Neural Information Processing Systems, 2010, pp. 1378–1386.
[36] C. Lu, W. Lam, Y. Zhang, Twitter user modeling and tweets recommendation based on Wikipedia concept graph, in: Proceedings of the Twenty-Sixth Workshops AAAI Conference on Artificial Intelligence, 2012, pp. 33–38.
[37] Z. Luo, M. Osborne, J. Tang, T. Wang, Who will retweet me?: Finding retweeters in twitter, in: Proceedings of the Thirty-Sixth International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2013, pp. 869–872.
[38] S.A. Macskassy, On the study of social interactions in Twitter, in: Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, 2012.
[39] S.A. Macskassy, M. Michelson, Why do people retweet? Anti-homophily wins the day!, in: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011.
[40] S.M. Mohammad, P.D. Turney, Crowdsourcing a word–emotion association lexicon, Comput. Intell. 29 (3) (2013) 436–465.
[41] N. Naveed, T. Gottron, J. Kunegis, A.C. Alhadi, Bad news travel fast: a content-based analysis of interestingness on twitter, in: Proceedings of the Third International Web Science Conference, ACM, 2011.
[42] H.K. Peng, J. Zhu, D. Piao, R. Yan, Y. Zhang, Retweet modeling using conditional random fields, in: Proceedings of the Eleventh IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, 2011, pp. 336–343.
[43] J.W. Pennebaker, M.E. Francis, R.J. Booth, Linguistic Inquiry and Word Count: LIWC 2001, 71, Lawrence Erlbaum Associates, Mahway, 2001, p. 2001.
[44] S. Petrovic, M. Osborne, V. Lavrenko, RT to win! Predicting message propagation in Twitter, in: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011.
[45] R. Pfitzner, A. Garas, F. Schweitzer, Emotional divergence influences information spreading in Twitter, in: Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, 2012.
[46] M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth, The author-topic model for authors and documents, in: Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2004, pp. 487–494.
[47] A. Sanjari, E. Khazraee, Information diffusion on twitter: the case of the 2013 Iranian presidential election, in: Proceedings of the 2014 ACM Conference on Web Science, ACM, 2014, pp. 277–278.
[48] K. Starbird, L. Palen, Pass it on? Retweeting in mass emergency, in: Proceedings of the Seventh International ISCRAM Conference2010, May, vol. 1, Seattle, 2010.
[49] K. Starbird, L. Palen, (How) will the revolution be retweeted? Information diffusion and the 2011 Egyptian uprising, in: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, ACM, 2012, pp. 7–16.
[50] S. Stieglitz, L. Dang-Xuan, Political communication and influence through microblogging – an empirical analysis of sentiment in Twitter messages and retweet behavior, in: Proceedings of the Forty-Fifth Hawaii International Conference on System Science (HICSS), IEEE, 2012, pp. 3500–3509.
[51] B. Suh, L. Hong, P. Pirolli, E.H. Chi, Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network, in: Proceedings of the 2010 IEEE Second International Conference on Social Computing (SocialCom), IEEE, 2010, pp. 177–184.
[52] T. Sun, M. Zhang, Q. Mei, Unexpected relevance: an empirical study of serendipity in retweets, in: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, 2013.
[53] Y.W. Teh, M.I. Jordan, M.J. Beal, D.M. Blei, Hierarchical Dirichlet processes, J. Am. Stat. Assoc. 101 (476) (2012) 1566–1581.
[54] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, A. Kappas, Sentiment strength detection in short informal text, J. Assoc. Inf. Sci. Technol. 61 (12) (2010) 2544–2558.
[55] I. Uysal, W.B. Croft, User oriented tweet ranking: a filtering approach to microblogs, in: Proceedings of the Twentieth ACM International Conference on Information and Knowledge Management, ACM, 2011, pp. 2261–2264.
[56] Vougioukas M., Androutsopoulos I., Paliouras G. (2017) Identifying Retweetable Tweets with a Personalized Global Classifier. arXiv:1709.06518. 2017 Aug 21.
[57] M. Wang, W. Zuo, Y. Wang, A multidimensional nonnegative matrix factorization model for retweeting behavior prediction, Math. Probl. Eng. 2015 (2015) Article ID 936397, 10 pages. http://dx.doi.org/10.1155/2015/936397.
[58] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (6684) (1998) 440.
[59] Z. Xu, Q. Yang, Analyzing user retweet behavior on twitter, in: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), IEEE Computer Society, 2012, pp. 46–50.

[60] Z. Xu, Y. Zhang, Y. Wu, Q. Yang, Modeling user posting behavior on social media, in: Proceedings of the Thirty-Fifth International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2012, pp. 545–554.
[61] M.C. Yang, J.T. Lee, H.C. Rim, Using link analysis to discover interesting messages spread across twitter, in: Proceedings of TextGraphs-7: the Workshop on Graph-based Methods for Natural Language Processing, Association for Computational Linguistics, 2012, pp. 15–19.
[62] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, Z. Su, Understanding retweeting behaviors in social networks, in: Proceedings of the Nineteenth ACM International Conference on Information and Knowledge Management, ACM, 2010, pp. 1633–1636.
[63] T.R. Zaman, R. Herbrich, J. Van Gael, D. Stern, Predicting information spreading in twitter, in: Proceedings of the 2010 Workshop on Computational Social Science and the Wisdom of Crowds, NIPS, 104, 2010, pp. 17599–17601.
[64] J. Zhang, B. Liu, J. Tang, T. Chen, J. Li, Social influence locality for modeling retweeting behaviors, in: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, AAAI Press, 2013, pp. 2761–2767.
[65] P. Zhang, X. Wang, B. Li, On predicting Twitter trend: important factors and models, in: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2013, pp. 1427–1429.
[66] Q. Zhang, Y. Gong, Y. Guo, X. Huang, Retweet behavior prediction using Hierarchical Dirichlet process, in: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015, pp. 403–409.
[67] Q. Zhang, Y. Gong, J. Wu, H. Huang, X. Huang, Retweet prediction with attention-based deep neural network, in: Proceedings of the Twenty-Fifth ACM International on Conference on Information and Knowledge Management, ACM, 2016, pp. 75–84.
[68] W.X. Zhao, J. Jiang, J. Weng, J. He, E.P. Lim, H. Yan, X. Li, Comparing twitter and traditional media using topic models, in: Proceedings of the 2011 European Conference on Information Retrieval, Springer, Berlin, Heidelberg, 2011, pp. 338–349.

Syeda Nadia Firdaus is a PhD candidate in the Department of Computer Science at Ryerson University, Canada. She received MSc in Computer Science from Ryerson University, Canada in the year 2013. Her research interests include social network data analytics, prediction models, and data mining. She has publications in international journals and conferences such as International Journal of Machine Learning and Cybernetics, International Conference on Social Networks Analysis and Mining (ASONAM), and International C* Conference on Computer Science and Software Engineering. She also worked in projects funded by Natural Science and Engineering Research Council of Canada.



Chen Ding is a Professor in the Department of Computer Science at Ryerson University, Canada. She received her PhD in Computer Science from National University of Singapore in 2003. Her research interests include web service selection, recommender systems, social data analytics, and web mining. She has publications in various international journals and conferences such as IEEE Transactions on Services Computing (TSC), Future Generation Computer Systems (FGCS), IEEE International Conference on Web Services (ICWS), IEEE International Conference on Cloud Computing (Cloud), etc. She also serves various international conferences as a program committee member.

Dr. Alireza Sadeghian has been with the Department of Computer Science at Ryerson University since 1999, where he holds the position of the Professor. He is also an Affiliate Scientist at the Li Ka Shing Knowledge Institute, St. Michael's Hospital, and serves as the research Theme Lead on Healthcare Analytics and Applications at the Institute for Biomedical Engineering, Science, and Technology (iBEST). Dr. Sadeghian is the founding Director of the Computational Intelligence Initiative (CI2) and Ubiquitous and Pervasive Computing Laboratories (UPCL), and has extensive expertise in the areas of modeling of complex dynamical systems, and applied machine learning particularly related to industrial and medical applications. He has published over 140 journal manuscripts, refereed conference papers, and book chapters, as well as two edited books. Dr. Sadeghian has been actively involved with a number of professional and academic boards, including IEEE Education Activity Board and NAFIPS board (North American Fuzzy Information Processing Society). Presently, he is the Chair of IEEE Computational Intelligence Technical Society Chapter, Toronto Section. Dr. Sadeghian is also on the Editorial Board of Applied Soft Computing Journal and serve as an Associate Editor of IEEE Access, Information Sciences, and Expert Systems Journal.