Multimedia Tools and Applications (2022) 81:27309–27338
https://doi.org/10.1007/s11042-022-12815-1

Role of twitter user profile features in retweet prediction for big data streams

Saurabh Sharma [1] · Vishal Gupta [1]

Received: 19 February 2021 / Revised: 2 February 2022 / Accepted: 9 March 2022 /
Published online: 26 March 2022

Abstract

To study the various factors influencing the process of information sharing on Twitter is a very active research area. This paper aims to explore the impact of numerical features extracted from user profiles in retweet prediction from the real-time raw feed of tweets. The originality of this work comes from the fact that the proposed model is based on simple numerical features with the least computational complexity, which is a scalable solution for big data analysis. This research work proposes three new features from the tweet author profile to capture the unique behavioral pattern of the user, namely "Author total activity", "Author total activity per year", and "Author tweets per year". The features set is tested on a dataset of 100 million random tweets collected through Twitter API. The binary labels regression gave an accuracy of 0.98 for user-profile features and gave an accuracy of 0.99 when combined with tweet content features. The regression analysis to predict the retweet count gave an R-squared value of 0.98 with combined features. The multi-label classification gave an accuracy of 0.9 for combined features and 0.89 for user-profile features. The user profile features performed better than tweet content features and performed even better when combined. This model is suitable for near real-time analysis of live streaming data coming through Twitter API and provides a baseline pattern of user behavior based on numerical features available from user profiles only.

Keywords Twitter · Social media analysis · Retweet prediction · User behavior · User profiling · Big data analysis

✉ Vishal Gupta
vishal@pu.ac.in

Saurabh Sharma
saurabhsharma381@gmail.com

[1] University Institute of Engineering and Technology, Panjab University, Chandigarh, India

1 Introduction

Today we are living in a world, where people have an active participation in online platforms of social interaction. Some kind or other, online social networks are part of our daily lives. The various types of social media platforms provide different types of services ranging from sharing personal views, collaborating with others, spreading the information of interest, exploring new ideas, discussing real-life events, and participating in evolving communities. Every social media network has a unique purpose, for example, Facebook is primarily used to connect with family and friends, Linkedin is used to connect with people from the professional circle, Instagram is used to share multimedia content, Pinterest is used to explore interesting pins of others and Tumblr is used to find and follow blogs from various categories [4, 49].

In the last 10 years, social media analysis has shown a growth in research studies ranging from ROI for organizations, prediction of real-life changes influenced by social media, descriptive analysis of real-life events as discussed on online platforms [10], viral marketing, social issues, health issues, natural disasters, emergencies, online surveys, countering fake information, detecting cyber bullying and use of abusive language, e-learning, online monitoring, etc.

In the research area of social media analysis, Twitter is a very popular choice of researchers because of its simple method of accessing data using an API interface. The raw feed from Twitter API is very rich in information, in terms of tweet content features and user profile features. The real potential of getting data from API is that it can be used for real-time data analysis and also for batch processing of a huge amount of data [4, 37].

1.1 Motivation

To study the activities of online users and to understand the behavioral pattern of users in various research domains noteworthy efforts are being made in the past few years [29, 35, 36, 38, 44]. The online activities make every user unique from other users which will become visible as strong patterns in time. The behavior patterns or signature style of a user is very useful in authentication, identification, and access control applications [19].

User centric approach The proposed work is an attempt to predict retweets from the point of view of a single user. A Twitter user is the only identity who will take action using free will. As shown in the Fig. 1, a Twitter user receives a huge amount of information from various sources. This information overload has a very deep impact on user actions. A user can not consume the sheer amount of information at the same pace as it arrives. This will leads to a situation where a user may take an active action or a passive action on the current piece of information. All the actions where a user generates some new content fall under the active actions and all the actions where a user does not generate new content come under passive actions. The action of retweeting comes under the category of passive action because without adding any new information, a user let the existing information flow towards its followers in the network.

These actions form the basis of user behavior and all these actions get recorded in the user profile. For example, a user profile contains information about how many tweets have been posted by a user (active action), and how many tweets are marked as favorites (passive action). The number of tweets retweeted by a user is not recorded in the profile and hence, it is a

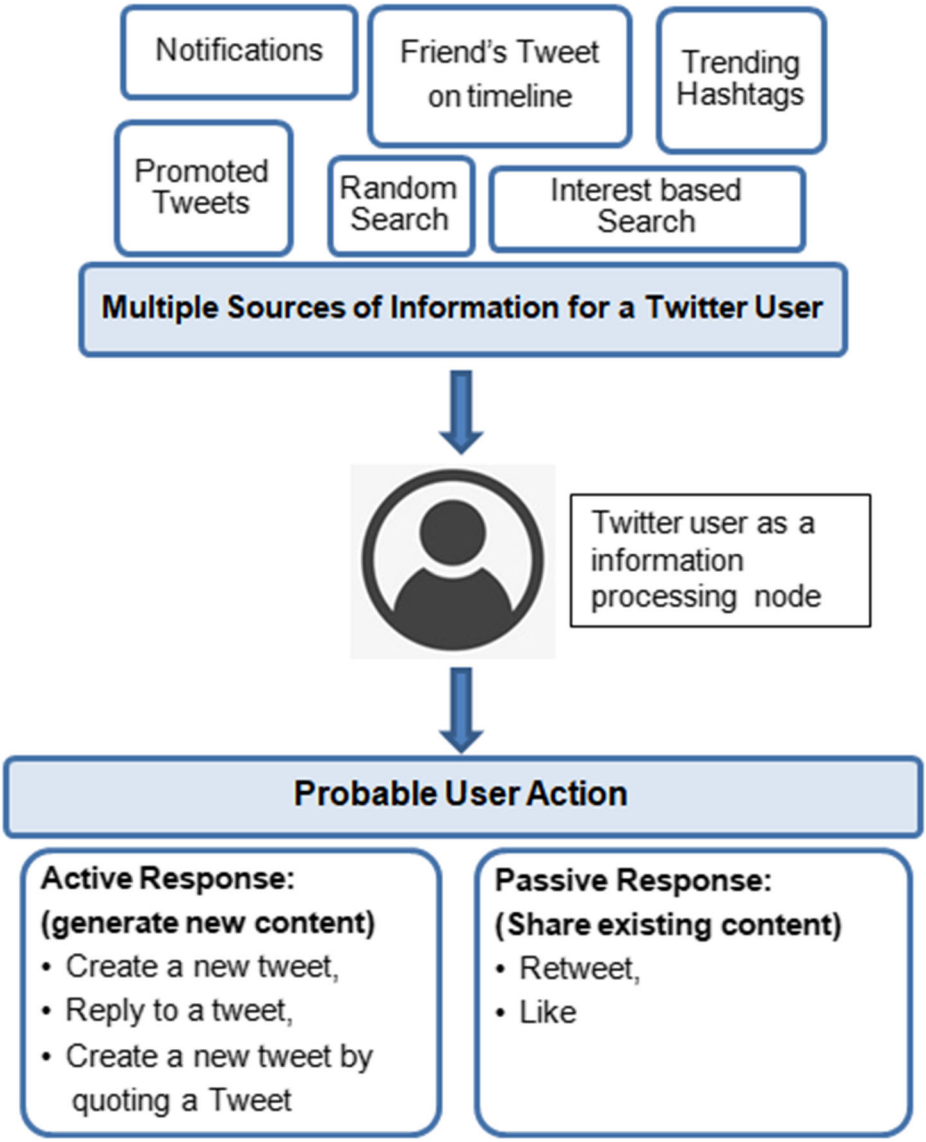


Fig. 1 Twitter user as information processing node

research problem of retweet prediction by analyzing the all other actions performed by the user from the time a user account has been created.

User data and twitter dataset The problem of reproducing a Twitter dataset is a major issue for user behavior analysis. The public datasets, release only tweet content features or sometimes just TweetIDs. The challenge of hydrating the dataset from TweetIDs after 4 years results in a loss of 30% dataset [48]. The terms and conditions of Twitter API do not allow fetching user profiles from TweetIDs. The proposed work is an attempt to provide an alternative way to handle this problem by using public Twitter archives [24, 40].

The Fig. 2 has shown three layers of features which can be used for the retweet prediction. The first layer consists of user features which are available with every tweet collected using API. The numerical features can be used as it is and some features can be computed with basic mathematical operations. The second layer i.e. tweets' content features are partially available in API and more features can be created using complex algorithms such as NLP features. The third layer of features is not directly available in the random feed of tweets. These features must be generated using various methods of data collection, complex algorithms and different assumptions about the structure of network. The recent studies have used various combinations of features from all three layers. However, those methods are not reproducible because user information cannot be shared publically.

User profiles are the most significant part of the user behavior analysis, and easily available with every tweet coming from random feed. Zubiaga et al. [48] found that the most common method of data collection from Twitter is using Twitter streaming API. The use of limited features available in Twitter API can be one of the solutions to generate domain independent, language independent and general purpose analysis on very large datasets. Recent studies [24, 48] have found that due to concerns of user privacy and restrictions imposed by social media companies on the distribution and sharing of dataset makes it very difficult to reproduce the same dataset for social media analysis [48].

A study [40] on the comparison of Twitter datasets and Twitter archives suggested that freely available archives should be used as an alternative way to reproduce and distribute datasets. The available archives are collections of the live feed of random tweets captured using Twitter API. Each tweet contains all data fields available in API as a JSON document. The significance of using archives is that it contains the full user profile along with tweet content features.

1.2 Significance of proposed work

Objectives:

- To provide a baseline pattern of retweet prediction (using 100 million random tweets) for domain-independent data feed with a minimum feature set and low computation requirement.

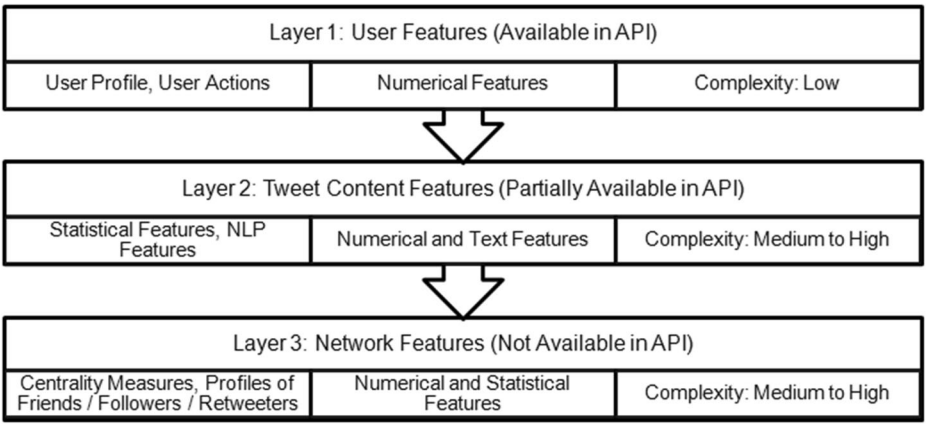


Fig. 2 Twitter API and availability of Features

- To propose a method for user behavior research that is reproducible, scalable, and using a public dataset without violating the terms and conditions of Twitter API.
- To reduce the complexity of social media analysis for big data streams using basic numerical features.
- To predict the retweets for every random user irrespective of the fact if a user is a normal user or a influencer/celebrity user.

Conditions:

- The dataset contains a random feed without any specific domain, topic, or other conditions.
- The proposed feature set is created from features available in Twitter streaming API only.
- The dataset, containing full user profiles, is freely available for research.
- The feature set includes only numerical values for fast processing and to reduce the computational complexity of text features.

Outcomes:

- The user profile features performed better than tweet content features for retweet prediction.
- The basic numerical features are very useful for real time user behavior analysis.
- No preprocessing requirement for proposed features set makes it fast and scalable for processing of big data streams.
- The proposed features set have shown promising results for regression and classification algorithms.
- The proposed work is able to predict for every user profile, influential or normal user.

In the following sections, the article is divided as follows. The related work on retweet prediction is given in section 2. In section 3 authors described the methodology of the study. The evaluation of the proposed work using Machine Learning Algorithms is presented in section 4. Section 5 comprises of Conclusions and the future scope of this study.

2 Related works

To understand the user behavior, one interesting research question is, why a user shares few tweets within network and not all of them. The probable reason can be due to information overload, it is practically not possible for a user to keep sharing every incoming tweet. Hemsley [25] found that approximately 47% tweets did not get retweets [14]. It presents an opportunity to study and analysis various factors of user actions to predict information sharing behavior.

Recent studies on information sharing proposed various methods to answer these questions. The studies focused on the content of tweets used sentiment analysis, location-based features, NLP techniques, use of hashtags (#), cashtags ($), URLs, and various text-based statistical features [10, 22, 26, 45]. The text-based approaches demand heavy computational resources and also in some cases all past tweets of the user [10, 23, 27, 43, 47]. The tradeoff between

accuracy and computational resources is the bottleneck to scale up for big data analysis and real-time analysis of live data streams.

The graph-based approaches are commonly limited to well-defined network boundaries and some static assumptions about the growth of the network [8]. In reality, to replicate these studies is a very big computation challenge and also very difficult to produce the same accuracy every time due to evolving network structure.

The retweet cascade techniques need data for first k retweets or the first 5–10 min window of temporal features for retweet prediction. The problem with this method is that the time stamp and user profile of each retweeter is needed to create a retweet cascade for every single tweet. These approaches are not useful for live feed data, because it is not possible to monitor every single tweet for its upcoming retweets before starting predicting [14, 18, 28, 31, 46, 47].

Retweet prediction is a very popular way of understanding the dynamics of information sharing on Twitter. In recent years, various combinations of features have been proposed for more accurate retweet prediction. The features range from simple statistical features to more complex features including language-specific NLP features, network structure and centrality-based features, temporal features consisting of first n retweets, etc. There are three main questions to understand information sharing on Twitter. The first question is which tweet will get retweets and why? The second question is, what is the significance of network structure and position of a user in the network for successful information diffusion? The third question is which user will retweet a tweet and why? To answer these questions, information required includes information about tweet content, network structure and user profiles of the author of the source tweet, and user profiles of users who will retweet it further.

Hemsley [25] used network structure features to predict the extent of information sharing for political messages and found that users with medium size network are more successful in spreading political information as compared to influential users with large network size. Dinh & Parulian [15] used cascade model for retweet, quote and reply tweets for COVID related tweets. They found that average cascade length for retweets is 4 h, for quote tweets is 3 days and for reply tweets is 2 days. This pattern indicates that active actions of users in form of quote and reply have more impact than passive action of retweet. Chen e.t. [10] studied the information sharing in the domain of disaster related tweets using NLP and network features and found that neutral and positive sentiment tweets had larger reach as compared to negative information. This finding is just opposite for political messages. Interestingly, they also found that if any negative information gets few retweets then it gets more responses than positive posts. The panic situation and worries about the disaster impact user behavior to share negative information more rapidly.

For handling big data streams, recent studies have proposed some very promising solutions. Murshed et al. [34] have proposed a model to calculate the overall accuracy of Twitter dataset using three different methods. Atish's measures outperformed other methods. They found that due to several language issues related to spelling, grammar and unstructured style of writing makes it very challenging to achieve higher level of accuracy. Singh e.t. [42] have proposed a framework for processing of big data using machine learning approach. The proposed framework showcased fast processing using distributed computing and ability to scale performance of machine learning algorithm. The clustering of incoming data stream is very difficult for standard machine learning algorithms. Arpaci et al. [5] have proposed evolutionary clustering for Twitter streams on COVID related tweets. They used 43 M+ tweets as a dataset. Duan et al. [16] proposed an algorithm SELM (Spark Extreme Learning Machine) for multi-

classification of big data using Apache Spark cluster. The proposed algorithm performed better and achieved highest speedup than traditional ELM (Extreme Learning Machine) algorithms.

The information sharing can be analyzed from three different points of view. The first view [10, 14, 20, 26] is to predict if a tweet will get a retweet or not? The second view [35, 36, 38, 44] is why tweets of some users get more retweets than other user's tweets? The third view [18, 31, 46, 47] is to predict which user will retweet a post and why? To answer these questions, many recent studies have proposed a large number of new features and claimed better results. However, every study is unique in terms of a dataset, domain, set of assumptions, manually coded features, and nature of findings. The replication of these studies is not suitable for domain-independent, standard features set, and real-time analysis.

A brief summary of related work categorized by feature set used is given in Table 1.

2.1 Challenges for retweet prediction in real time big data analysis

Based on the literature review, following issues are listed:

- NLP based approaches need language specific libraries and very hard to scale for language independent analysis.
- Network based approaches need huge amount of information about social circle of each user, which is not feasible for real time random data feed.
- Manually coded features do not support real time analysis of big data streams.
- User data is not available from recent studies for performance comparison.

The new features proposed in recent studies are given with the description and whether these features can be extracted using the free Twitter API service. The tweet content features are given in Table 2 and Table 3 shows the features based on the Author profile.

3 Methodology

Based on the challenges of retweet prediction for big data streams of random tweets, authors proposed a simple, fast and scalable machine learning approach using simple numeric features available in Twitter API. The category and list of features is shown in Fig. 3. The categorization of features is based on the information contained by a feature. The tweet content features have information about the tweet text and the count of user responses. The user profile features contains the information about the author of the tweet. It includes information about user social circle and user past actions/activities since user account created.

To understand the active and passive participation of a user, authors have proposed a new feature as "Author total activity". This feature is defined as the sum of all tweets posted by a user (active action) and the total tweets liked by a user (passive action). For a user, the total tweets posted and total activity shows very large values for old accounts and small values for new accounts. Therefore, the new features are introduced to calculate per year values for these features by dividing it from user account age counted in years.

$$Author\ Total\ Activity = Author\ Tweets\ Count + Author\ Favorites\ count \quad (1)$$

Table 1 Brief summary of related work

Research Work	Year	Dataset	Features Used	Topic
BPF A Unified Factorization model for predicting retweet behaviors [47]	2020	Sina weibo Dataset 1,60,02,390 microblogs 7,982,752 users	Network features, NLP features, Tweet cascades	Random
Composing tweets to increase retweets [26]	2019	Twitter Dataset a subset of a large corpus of about 1.77 million topic-author controlled tweets	NLP features, User profile, Tweet cascades	Random
COVID-19 pandemic and information diffusion analysis on Twitter [15]	2020	Twitter Dataset, 675,228 tweets	Network features	COVID-19
Crowd or Hubs information diffusion patterns in online social networks in disasters [18]	2020	Twitter Dataset 14 million tweets	NLP features, Tweet cascades	Hurricane Harvey
Followers Retweet The Influence of Middle-Level Gatekeepers on the Spread of Political Information on Twitter [25]	2019	Twitter Datasets 20,580 tweets, 755,957 tweets	Tweet cascades	Random
HawkesEye Detecting Fake Retweeters Using Hawkes Process and Topic Modeling [17]	2020	Twitter Dataset 30,000 tweet objects, 2, 508 retweeters	NLP Features, Manually coded User profile features	Random
Popularity Prediction for Single Tweet based on Heterogeneous Bass Model [22]	2020	Twitter Dataset 2,516,440 tweets 2,122,135 users.	NLP features User Profile features,	Random
Predicting Rumor Retweeting Behavior of Social Media Users in Public Emergencies [46]	2020	Sina weibo Datasets historical tweets 1: 284238 historical tweets 2: 203523	NLP features, Tweet cascades	Public emergencies
Predicting User Retweeting Behavior in Social Networks With a Novel Ensemble Learning Approach [7]	2020	Sina weibo Dataset 762,936 microblogs published by 68,817 users	NLP features Network features, User Profile features	COVID-19
Prediction of Likes and Retweets Using Text Information Retrieval [13]	2020	Twitter Dataset 2 million Tweets,	NLP features	Data science
R-Map A Map Metaphor for Visualizing Information Reposting Process in Social Media [9]	2019	Sina weibo Dataset	Network features, NLP features, Tweet cascades	Random
Temporal Sequence of Retweets Help to Detect Influential Nodes in Social Networks [6]	2019	Twitter Datasets 12,44,645 Tweets, 7,63,109 Tweets	Network features, Tweet cascades	Random
Uncovering sentiment and retweet patterns of disaster-related tweets from a spatiotemporal perspective – A case study of Hurricane Harvey [10]	2020	Twitter Dataset, 7,041,866 tweets	NLP features	Hurricane Harvey

$$Author\ \ Tweets\ per\ year = \frac{Author\ \ Tweets\ \ Count}{Account\ Age} \tag{2}$$

Table 2 List of Features based on Tweet Content

Sr. No.	Feature	Description	Whether the Feature can be computed from Twitter API	Recent Studies
1.	Total hashtag	Count of hashtags in a tweet	Yes	[29, 35, 36, 41, 43, 46]
2.	Total link	Count of link in a tweet	Yes	[29, 33, 35, 43, 44]
3.	Total mention	Count of users mentions in a tweet	Yes	[35, 36, 41]
4.	Total retweet	Count of retweets received by a tweet	Yes	[29, 33, 36, 43]
5.	Is marked Favourite	Check if the favorite count is zero or not	Yes	[10, 36]
6.	Publication time of Tweet	Timestamp in 24 h format as per local time zone of a user account	Yes	[35]
7.	Original tweet or retweet	Check if a tweet is an original post or a retweet	Yes	[36]
8.	Tweet text	The textual content of a tweet post in UTF-8 format	Yes	[36, 43]
9.	URLs	Hypertext of URL posted in a tweet	Yes	[36]
10.	Tweet ID	Unique ID of a tweet	Yes	[36]
11.	Word Count	Total number of words in the text of a tweet	Yes	[43, 44]
12.	Character Count	Total number of characters in the text of a tweet	Yes	[26]
13.	Symbols and acronyms Count	Total number of symbols and acronyms in the text of a tweet	Yes	[26]
14.	Punctuations	Total number of punctuations in a tweet	Yes	[26]
15.	Creation Time of Tweet	The timestamp of posting a tweet	Yes	[10, 26, 29, 43]
16.	Hashtag ratio	The ratio of all hashtags to all tweets	No	[48]
17.	Link ratio	The ratio of all links to all tweets.	No	[48]
18.	Mention ratio	The ratio of all mention to all tweets.	No	[48]
19.	Retweet ratio	The ratio of all retweets to all tweets.	No	[48]
20.	Total likes count	Count of all tweets liked	No	[1, 35]
21.	Tweet similarity	The similarity of tweet text using cosine similarity.	No	[48]
22.	Unique URL ratio	The ratio of unique URLs posted to total tweets.	No	[48]
23.	Duplicate tweet count	Count of tweets posted as duplicate	No	[48]
24.	Unique hashtag	Count of unique hashtags used in all tweets.	No	[41]
25.	Unique mention	Count of unique mentions in all tweets.	No	[41]
26.	Maximum frequency of hashtag	Hashtag with maximum frequency in all tweets.	No	[41]
27.	Average frequency of hashtag	Mean value of hashtags used in all tweets.	No	[41]
28.	Average frequency of mention	Mean value of mentions used in all tweets.	No	[1]
29.	Average frequency of URLs	Mean value of URLs posted in all tweets.	No	[41]
30.	Deviation of hashtag	Hashtags population deviation in all tweets.	No	[1]
31.	Deviation of link	Links population deviation in all tweets.	No	[1]
32.	Deviation of mention	Mentions population deviation in all tweets.	No	[1]
33.	Deviation of re-tweet	Retweets population deviation in all tweets.	No	[1]
34.	Deviation of tweet length	Tweet length population deviation in all tweets.	No	[1]

Table 2 (continued)

Sr. No.	Feature	Description	Whether the Feature can be computed from Twitter API	Recent Studies
35.	Deviation of hashtag position aggregate	Population deviation of hashtag position aggregate.	No	[1]
36.	Deviation of link position aggregate	Link position population deviation aggregate.	No	[1]
37.	Deviation of mention position aggregate	Mention position population deviation aggregate.	No	[1]
38.	Average daily tweet	The ratio of all tweets to count of days between first and last tweet.	No	[1]
39.	Average tweet length	Mean value of the lengths of all tweets.	No	[1]
40.	Average sentiment polarity	Mean value of the polarity of sentiment for every posted tweet.	No	[1, 12]
41.	Average sentiment subjectivity	Mean of sentiment subjectivity for every posted tweet.	No	[1, 2]
42.	Average TF-IDF score	Mean value of TF-IDF weight of the tweets.	No	[1]
43.	Popularity ratio	The ratio of the favourites count plus re-tweet count to the number of all tweet count.	No	[1]

$$Author\ Total\ Activity\ per\ year = \frac{Author\ Tweets\ Count + Author\ Favorites\ count}{Account\ Age} \quad (3)$$

The methodology is explained step by step in Fig. 4. The first requirement is to collect tweets from random feed of Twitter API. Then for each tweet, extract all features available and categorize them into two categories. After that, select only numerical features and compute new proposed features.

The proposed work is an attempt to predict retweets with the help of information available from a single tweet post without any prior information about the user, network structure, temporal features, and historical tweets. For each random tweet there are following questions for retweet prediction:

RQ1: How to predict whether a tweet will be retweeted or not?
RQ2: How to estimate the exact number of retweets a tweet will get?
RQ3: How to categorize tweets into different classes based on estimated ranges of retweet count?

The machine learning algorithms used in this study is regression algorithms and classification algorithms as shown in Fig. 4.

Table 3 List of Features based on User (Author) Profile

Sr. No.	Feature	Description	Whether the Feature can be computed from Twitter API	Recent Studies
1.	**Author total Activity**	Sum of all the tweet posted by a user and all the tweets liked by a user	**Yes**	**Proposed**
2.	**Author total Activity per year**	Sum of all the tweet posted by a user and all the tweets liked by a user divided by user account age in years	**Yes**	**Proposed**
3.	**Author tweets per year**	Sum of all the tweet posted by a user divided by user account age in years	**Yes**	**Proposed**
4.	Screen name length	Count of characters in the screen name of a user.	Yes	[30]
5.	User location	If user location is mentioned or not.	Yes	[1, 10, 36]
6.	Age in days (Creation date of User Account)	The number of days since User Account created.	Yes	[35, 36, 44, 49]
7.	Followers count	Followers count of the user.	Yes	[26, 35, 36, 43, 44, 48]
8.	Friends count	Friends count of the user.	Yes	[26, 33, 35, 36, 43, 44]
9.	Statuses count	Number of statuses posted by a user	Yes	[1, 26, 35, 43, 44, 48]
10.	Favorites count	Count of tweets a user has marked as favorite.	Yes	[33, 48]
11.	User description	Check If the user description is provided or left blank.	Yes	[2]
12.	Account verified	Check if the user account is marked as verified or not.	Yes	[11]
13.	Default profile image	Check if the profile image is default or changed by the user.	Yes	[3]
14.	Listed count	Count of lists where the user account is listed.	Yes	[33, 35]
15.	Account reputation	Normalized ratio of user followers to user friends.	Yes	[41]
16.	Follower following ratio	The ratio of the count of user followers to user friends.	Yes	[48]
17.	Following follower ratio	The ratio of the count of user friends to user followers.	Yes	[49]
18.	User ID	Unique ID of the User	Yes	[36]
19.	User Name	Display name of the user account	Yes	[36]
20.	Profile URL	Check if profile URL is provided or not	Yes	[1]
21.	Default profile	Check if the profile theme is default or changed by the user.	Yes	[1]
22.	User Time zone	Check if user time is present or not.	Yes	[1]
23.	Geo-enabled	Check if geotagging is enabled or not by the user.	Yes	[1]
24.	Tweet text of all past Tweets	Collection of Text of all posted tweets	No	[22]
25.	Sentiment Score	Sentiment score based on tweet text	No	[10, 22]

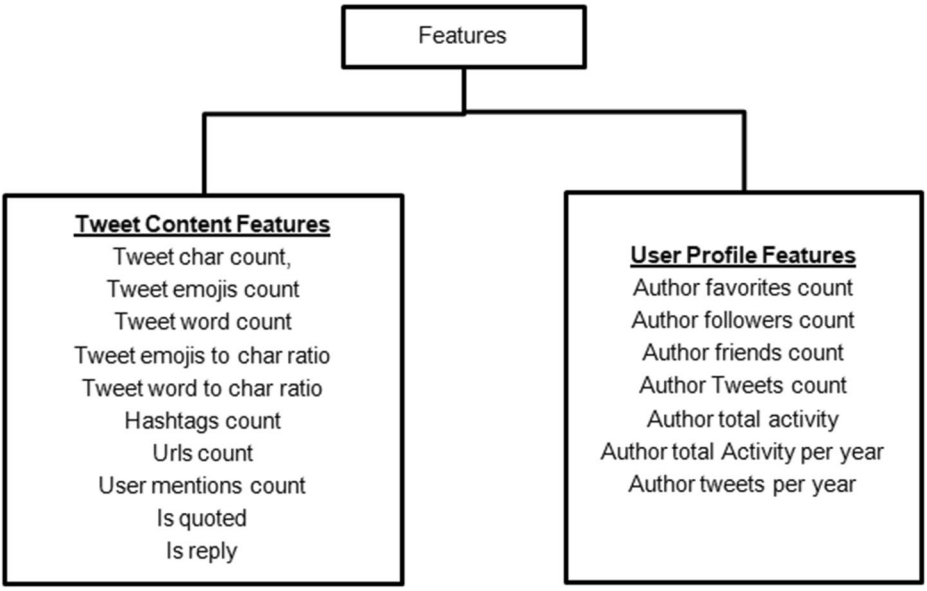


Fig. 3 Features used in the study

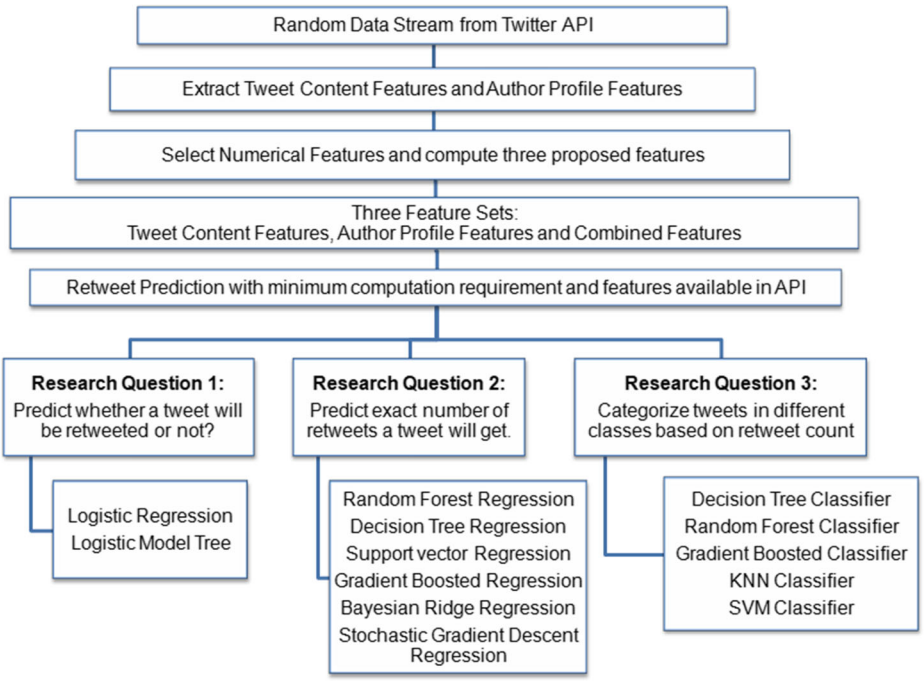


Fig. 4 Proposed Methodology for Retweet Prediction

Algorithm for the generation of Features sets from Twitter data stream.

```
1:Input: Raw tweets from Twitter stream API stored in MongoDB.
2: df_all_tweets ← Read MongoDB collection
3: new_columns[ ] ← ["Tweet char count, Tweet emojis count, Tweet word count, Tweet emojis to char ratio,
4:                    Tweet word to char ratio, Hashtags count, Urls count, User mentions count, Is quoted, Is
5:                    reply, Author favorites count, Author followers count, Author friends count, Author
6:                    Tweets count, Author total activity, Author total Activity per year, Author tweets per year"]
7: df_all_tweets ← Create new columns df_all_tweets[new_columns]
8: For each Tweet in df_all_tweets do
9:       For each col in new_columns do
10:            Calculate numeric value of col
11:            Set df_all_tweets [col] = numeric value
12:      End for
13: End for
14: df_all_features ← Select df_all_tweets[new_columns]
15: df_tweet_features [ ] ← Select df_all_features ["Tweet char count, Tweet emojis count, Tweet word count,
16:                          Tweet emojis to char ratio, Tweet word to char ratio, Hashtags count, Urls count, User
17:                          mentions count, Is quoted, Is reply"]
18: df_user_features [ ] ← Select df_all_features ["Author favorites count, Author followers count, Author friends
19:                         count, Author Tweets count, Author total activity, Author total Activity per year,
20:                         Author tweets per year"]
21: Output df_tweet_features [ ], df_user_features [ ]
```

4 Experimental evaluation and results

4.1 Dataset: The dataset, of 100 million random tweets, is created from the online twitter archive of august 2018 [39, 40]

The description of the dataset used in the study is given in Table 4. The skewness and kurtosis along with other statistical metrics will help to reproduce this dataset and will also help in comparing any other dataset with similar properties. The maximum value for "Tweet char count" and "Tweet emojis count" is very large because Twitter supports Unicode format for emojis in which single emojis can be a combination of multiple characters.

4.2 Experimental setup (Fig. 5)

The Twitter data collected using streaming API is available as archives online. The Twitter archives are in compressed file format. These compressed files are a collection of JSON files that contain the actual raw data as received from streaming API. The JSON file format is a very good option for unstructured and text data of variable length. The size of every tweet object can vary depending upon the number of fields. For example, a tweet object of a retweet contains information of tweet author and retweeter, however, an original tweet object has only tweet author information. The NoSQL databases are used for handling variable-length documents with a large number of missing data fields. The MongoDB NoSQL database is used in this study. The distributed computing on 100 million tweets for big data analysis is done on an 8 node Apache Spark cluster where each node had 16 GB RAM, Intel 4 core i5 CPU. The programming is done in python language using the pyspark interface of Apache Spark. The Jupyter notebook is used for IDE.

Table 4 Description of 100 Million random Tweets Dataset created from Twitter Archives

Features	Count	Mean	std	Min	Max	Skewness	Kurtosis
Tweet char count	101,681,675	85.24	44.51	1	494	−0.08	−1.40
Tweet emojis count	101,681,675	0.46	1.95	0	140	20.97	833.89
Tweet word count	101,681,675	11.32	7.97	1	70	0.54	−0.78
Tweet emojis to char ratio	101,681,675	0.01	0.05	0	1	17.84	362.05
Tweet word to char ratio	101,681,675	0.08	0.08	0	1	1.68	13.16
Hashtags count	101,681,675	0.33	1.03	0	46	5.01	35.18
Urls count	101,681,675	0.18	0.41	0	5	2.05	4.01
User mentions count	101,681,675	0.91	0.95	0	28	4.08	30.33
Is quoted	101,681,675	−0.85	0.52	0	1	3.24	8.52
Is reply	101,681,675	−0.63	0.77	0	1	1.63	0.67
Author favorites count	101,681,675	8170.77	28,736.30	0	2,792,266	6.23	57.57
Author followers count	101,681,675	322,862.39	2,803,970.00	0	106,873,281	16.92	347.31
Author friends count	101,681,675	3178.34	30,755.90	0	4,710,009	19.22	621.58
Author Tweets count	101,681,675	28,310.74	424,569.00	0	27,837,830	11.24	259.67
Author total activity	101,681,675	36,481.99	426,640.00	0	27,838,020	5.85	63.66
Author total Activity per year	101,681,675	6636.31	50,133.70	0	5,508,960	5.85	63.66
Author tweets per year	101,681,675	4705.38	49,218.70	0	5,508,960	11.24	259.67
Retweet Count (as Label)	101,681,675	2341.21	18,934.10	0	3,614,140	23.49	1473.16

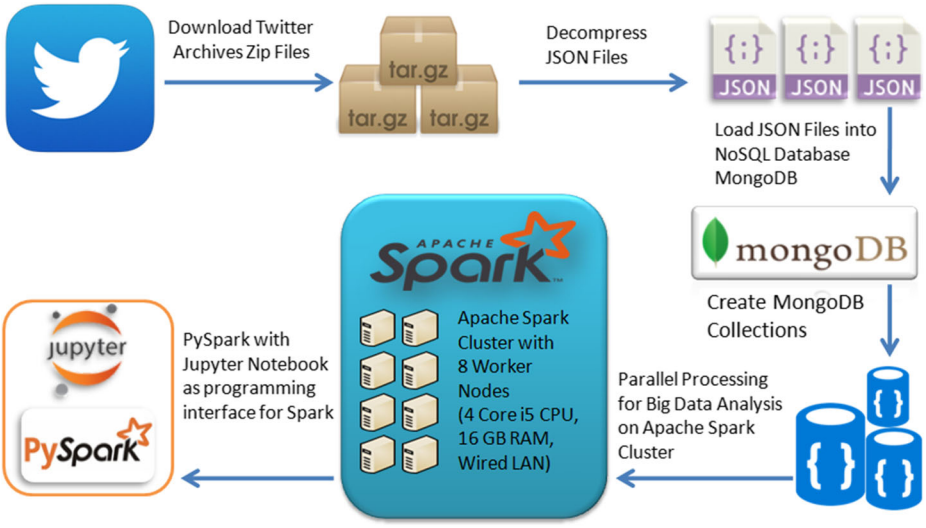


Fig. 5 Schematic representation of Experimental setup used for the study

4.3 Evaluation metrics

The evaluation metrics used in the study is given in Table 5.

$$Precision = \frac{TP\_retweet\_count}{TP\_retweet\_count + FP\_retweet\_count} \tag{4}$$

$$Recall = \frac{TP\_retweet\_count}{TP\_retweet\_count + FN\_retweet\_count} \tag{5}$$

$$F1\ Score = \frac{2*Precision\_retweet\_count*Recall\_retweet\_count}{(Precision\_retweet\_count + Recall\_retweet\_count)} \tag{6}$$

$$Accuracy = \frac{TP\_retweet\_count + TN\_retweet\_count}{TP\_retweet\_count + FP\_retweet\_count + FN\_retweet\_count + TN\_retweet\_count} \tag{7}$$

Where TP: True Positive, FP: False Positive, FN: False Negative

$$Log\ Loss = \frac{-1}{T}\sum_{i=1}^{T} rc_i.\log(P(rc_i)) + (1-rc_i).\log(1-p(rc_i)) \tag{8}$$

Where T: Number of Tweets, rc_i: observed retweet count

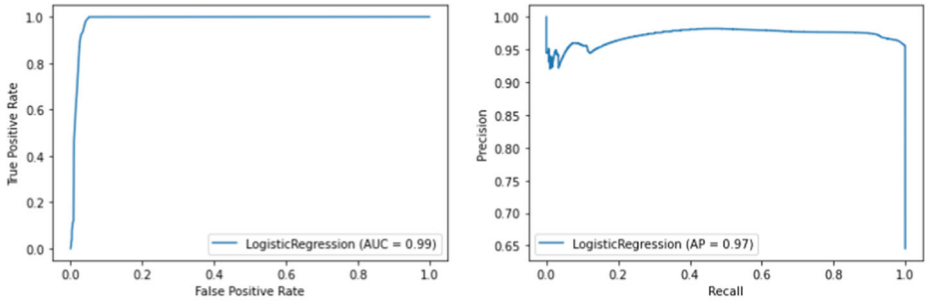
$$AUC = \int_i^j f(z).dz \tag{9}$$

Where i, j are limits of area, f(z) function of the curve

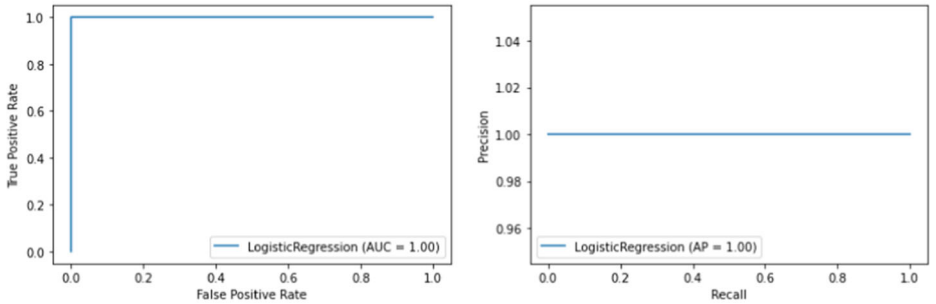
$$R^2 = 1-\frac{\sum\left(rc_i-\widehat{rc}_i\right)^2}{\sum\left(rc_i-\overline{rc}\right)^2} \tag{10}$$

Table 5 List of Evaluation metrics

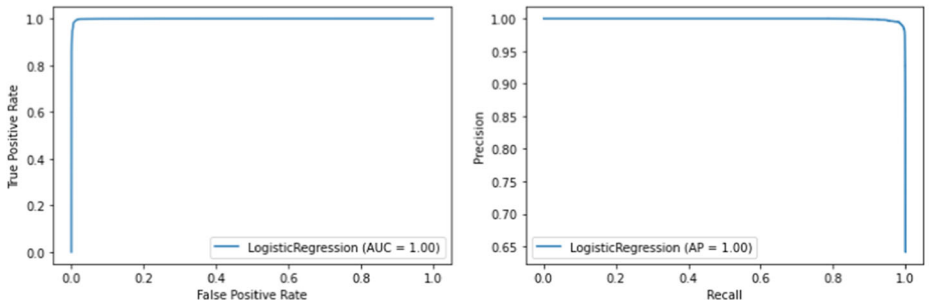
RQ 1: Binary Prediction	RQ 2: Regression Analysis	RQ 3: Classification
Precision	R-squared	Precision
Recall	Mean Square Error	Recall
F1-Measure	Root mean Square Error	F1-Measure
Log loss	Mean Absolute Error	
AUC	Median Absolute Error	
Accuracy		



(a)



(b)

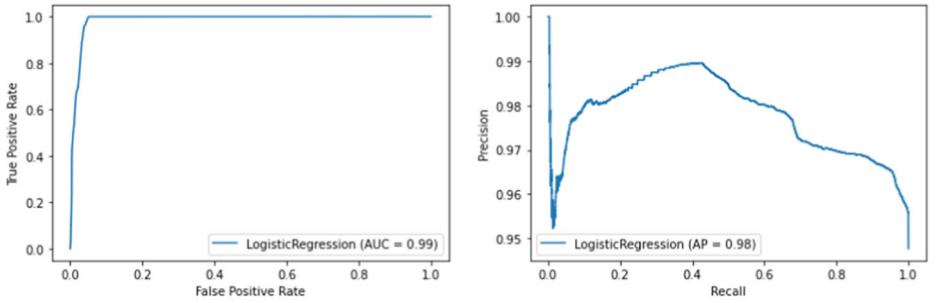


(c)

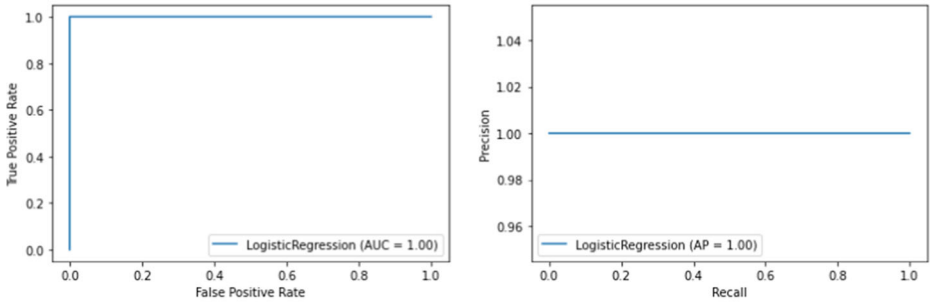
Fig. 6 AUC and PR Curve of Logistic Regression for RQ1. (**a**) LR: Tweet Content features. (**b**) LR: Author Profile features. (**c**) LR: Proposed Combined features

$$Mean\ Square\ Error = \frac{\sum_{i=1}^{T}\left(rc - \widehat{rc}_i\right)^2}{T} \tag{11}$$

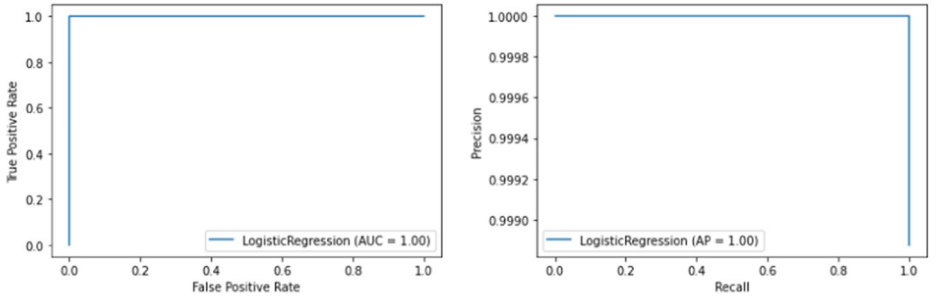
$$Root\ Mean\ Square\ Error = \sqrt{\frac{\sum_{i=1}^{T}\left(rc_i - \widehat{rc}_i\right)^2}{T}} \tag{12}$$



(a)



(b)



(c)

Fig. 7 AUC and PR Curve of Logistic Model Trees for RQ1. (**a**) LMT: Tweet Content features. (**b**) LMT: Author Profile features. (**c**) LMT: Proposed Combined features

$$MedAE\left(rc,\widehat{rc}\right) = median\left(\left|rc1-\widehat{rc}1\right|,\ldots\ldots,\left|rc_T-\widehat{rc}_T\right|\right) \tag{13}$$

Where rc_i: observed value, $\widehat{rc}_i$: predicted value, $\overline{rc}$: mean of all observed values

$$MAE = \frac{1}{T}\sum_{i=1}^{T}|rcp_i-rct_i| \tag{14}$$

Where rcp_i: retweet count predicted value, rct_i: retweet count true value.

4.4 Performance evaluation

To answer the three research questions, three feature sets were tested. The first set consists of only tweet content-based features, the second set consists of only author profile-based features and the third set is a proposed combination of both sets. The performance of each feature set is compared for each algorithm.

RQ1: Whether a tweet will be retweeted or not?

The RQ1 is a binary choice question. The reason for choosing binary labels is that in a random sample of tweets 45% to 50% tweets do not get any retweet [25]. The binary label will help to categories tweets into two classes which will reduce the total number of tweets for further analysis of predicting number of retweets a tweet can get. Two algorithms have been used for this task: logistic regression and logistic model trees. The results are given in Fig. 6, Fig. 7, Tables 6 and 7. All three feature sets were able to predict with very high accuracy. The small improvement is visible in result values starting from tweet content to author features to combined features. The answer to the first research question is yes, it is possible to predict with accuracy that whether a tweet will get a retweet or not.

Table 6 Performance Comparison part 1 for RQ1

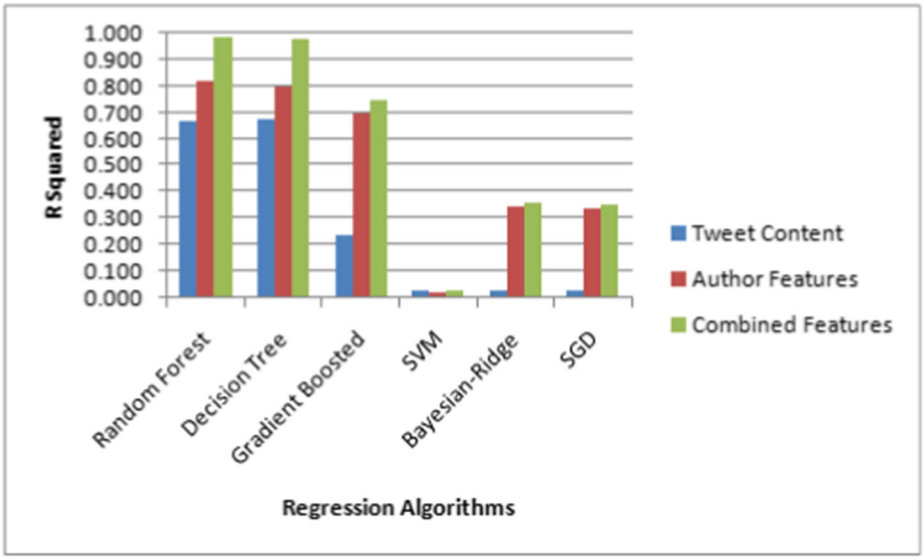
Logistic Regression			
	Tweet Feature	Author Feature	Proposed Combined Features
Log loss	0.13111	0.06666	0.03719
AUC	0.98531	1	0.99937
Accuracy	0.97	0.98	0.99
Logistic Model Tree			
	Tweet Feature	Author Feature	Proposed Combined Features
Log loss	0.12247	0.00045	0.00126
AUC	0.98579	1	1
Accuracy	0.97	1	1

Table 7 Performance Comparison part 2 for RQ1

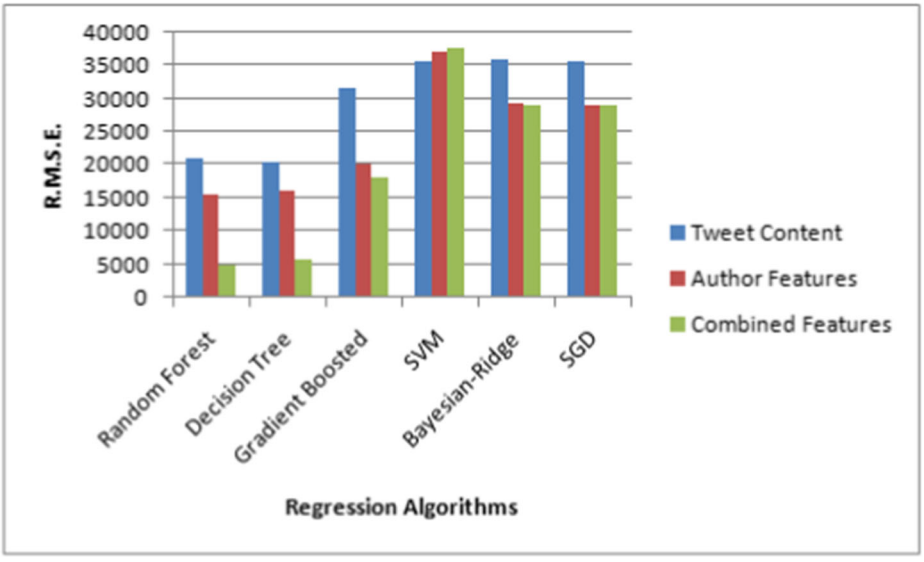
Logistic Regression									
	Tweet Feature			Author Feature			Proposed Combined Features		
	Precision (P)	Recall (R)	F1-score	Precision (P)	Recall (R)	F1-score	Precision (P)	Recall (R)	F1-score
FALSE	1	0.95	0.97	0.96	1	0.98	0.98	0.99	0.99
TRUE	0.96	1	0.98	1	0.96	0.98	0.99	0.98	0.99
Macro average	0.98	0.97	0.97	0.98	0.98	0.98	0.99	0.99	0.99
Weighted average	0.98	0.97	0.97	0.98	0.98	0.98	0.99	0.99	0.99
Logistic Model Tree									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
FALSE	1	0.95	0.97	1	1	1	1	1	1
TRUE	0.95	1	0.98	1	1	1	1	1	1
Macro average	0.98	0.97	0.97	1	1	1	1	1	1
Weighted average	0.98	0.97	0.97	1	1	1	1	1	1

RQ 2: Predict the accurate retweet count for a tweet.

The regression analysis is performed to determine the accurate retweet count for a random tweet. The results from regression algorithms are given in Fig. 8 and Table 8. The results from various regression algorithms indicates that author features performed better that tweet features and combined features gave the best performance as compared to both. The R-squared and



(a)



(a)

Fig 8 Regression Analysis for RQ 2. (**a**) R-Squared Value comparison, (**a**) RMSE value comparison

Table 8 Performance Comparison of Regression Analysis for RQ2

	Random Forest Regression		
	Tweet Content	Author Features	Proposed Combined Features
R-squared	0.6701	0.8162	0.9824
Mean Square Error	447,262,685.9598	240,524,382.1370	22,716,410.1733
Root mean Square Error	21,148.5859	15,508.8485	4766.1735
Mean Absolute Error	4540.3229	2121.2072	465.9543
Median Absolute Error	46.4760	8.2595	6.5000
	Decision Tree Regression		
	Tweet Content	Author Features	Proposed Combined Features
R-squared	0.6776	0.8005	0.9756
Mean Square Error	417,917,933.0894	256,786,293.8943	31,564,528.4161
Root mean Square Error	20,443.0412	16,024.5528	5618.2318
Mean Absolute Error	4468.0276	2139.7003	435.7641
Median Absolute Error	31.7551	1.0000	1.0000
	Gradient Boosted Regression		
	Tweet Content	Author Features	Proposed Combined Features
R-squared	0.239	0.699	0.745
Mean Square Error	1,000,375,856.76	405,327,941.74	326,527,452.02
Root mean Square Error	31,628.718	20,132.75	18,070.07
Mean Absolute Error	8533.43	4743.67	4582.83
Median Absolute Error	3261.74	834.79	859.04
	Support Vector Regression		
	Tweet Content	Author Features	Proposed Combined Features
R-squared	0.029	0.021	0.0257
Mean Square Error	1,271,439,879.27	1,386,167,002.3	1,417,650,832.75
Root mean Square Error	35,657.25	37,231.26	37,651.7
Mean Absolute Error	6137.07	6535.5	6584.76
Median Absolute Error	29.13	5.08	25.61
	Bayesian Ridge Regression		
	Tweet Content	Author Features	Proposed Combined Features
R-squared	0.024	0.346	0.36
Mean Square Error	1,290,683,865.92	860,205,619.56	850,860,029.43
Root mean Square Error	35,926.08	29,329.26	29,169.5
Mean Absolute Error	11,324.58	7611.37	7961.72
Median Absolute Error	6295.9	2529.78	3041.12
	Stochastic Gradient Descent Regression		
	Tweet Content	Author Features	Proposed Combined Features
R-squared	0.025	0.34	0.35
Mean Square Error	1,272,648,394.91	851,245,028.42	838,362,217.15
Root mean Square Error	35,674.19	29,176.1	28,954.48
Mean Absolute Error	11,309.67	7873.71	8166.94
Median Absolute Error	6297.14	3041.28	3142.62

RMSE value of every regression algorithm is plotted in Fig. 8. The Random Forest and Decision Tree classifiers performed best among all. All the algorithms produce poor results. It indicates that these features are not a good choice for answering this research question. Hence, the answer to the second research question is that prediction of the exact number of retweets is not possible. These features can be combined with some other features in future studies for exploratory analysis.

RQ3: Categorize tweets into multi-label classes.

To classify the tweets into various classes based on ranges of retweet count, different classification algorithms were used. The performance of three feature sets tested on the different number of bins. The criterion of binning is given in Table 8. The results are given in Tables 9, 10, 11, 12, 13 and 14.

Table 9 The binning criteria for classification

b1	b2	b3	b4	b5	b6	b7
rtc == 0	0<rtc<=10	10<rtc<=100	100<rtc<= 500	500<rtc<= 1000	1000<rtc<= 5000	rtc>5000

rtc: Retweet Count

The values of precision, recall, F1-score, and Accuracy measure are plotted. The performances of all three feature sets in terms of accuracy measure are above 0.8 score for number of classes less than 4. After that as the number of bins/classes increases, a steady decline in performances is visible. At the highest values of bins (bins = 7), tweet features performed less than 0.6 accuracy score, whereas, author features and proposed combined features performed more than 0.6 accuracy score.

The F1-score is plotted for all three classification algorithms and for all values of bins. The results are shown in Fig. 9. The best performing algorithm is Random

Table 10 Performance Metrics for Decision Tree Classification

Number of Bins=2									
	Tweet Feature			Author Feature			Proposed Combined Features		
	Precision (P)	Recall (R)	F1-score	Precision (P)	Recall (R)	F1-score	Precision (P)	Recall (R)	F1-score
Accuracy			0.98			1			1
macro average	0.98	0.98	0.98	1	1	1	1	1	1
weighted average	0.98	0.98	0.98	1	1	1	1	1	1
Number of Bins=3									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.87			0.89			0.9
macro average	0.77	0.67	0.66	0.6	0.67	0.63	0.84	0.71	0.71
weighted average	0.85	0.87	0.83	0.8	0.89	0.84	0.89	0.9	0.87
Number of Bins=4									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.77			0.79			0.8
macro average	0.53	0.51	0.46	0.52	0.56	0.53	0.57	0.54	0.51
weighted average	0.71	0.77	0.7	0.72	0.79	0.75	0.74	0.8	0.74
Number of Bins=5									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.68			0.71			0.72
macro average	0.37	0.43	0.37	0.38	0.48	0.42	0.39	0.45	0.4
weighted average	0.61	0.68	0.62	0.62	0.71	0.66	0.63	0.72	0.65
Number of Bins=6									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.65			0.67			0.68
macro average	0.29	0.36	0.3	0.35	0.39	0.34	0.31	0.38	0.32
weighted average	0.58	0.65	0.59	0.62	0.67	0.62	0.59	0.68	0.61
Number of Bins=7									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.57			0.64			0.63
macro average	0.22	0.3	0.23	0.38	0.38	0.34	0.37	0.37	0.33
weighted average	0.53	0.57	0.53	0.63	0.64	0.61	0.63	0.63	0.6

Table 11 Performance Metrics for Random Forest Classification

	Tweet Feature			Author Feature			Proposed Combined Features		
Number of Bins=2									
	Precision (P)	Recall (R)	F1-score	Precision (P)	Recall (R)	F1-score	Precision (P)	Recall (R)	F1-score
Accuracy			0.98			1			1
macro average	0.98	0.98	0.98	1	1	1	1	1	1
weighted average	0.98	0.98	0.98	1	1	1	1	1	1
Number of Bins=3									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.87			0.9			0.9
macro average	0.75	0.71	0.72	0.8	0.77	0.78	0.82	0.79	0.8
weighted average	0.85	0.87	0.86	0.89	0.9	0.89	0.9	0.9	0.9
Number of Bins=4									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.78			0.82			0.83
macro average	0.6	0.57	0.57	0.67	0.67	0.67	0.68	0.68	0.68
weighted average	0.75	0.78	0.76	0.81	0.82	0.81	0.82	0.83	0.82
Number of Bins=5									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.71			0.77			0.79
macro average	0.51	0.49	0.49	0.61	0.61	0.61	0.63	0.63	0.63
weighted average	0.69	0.71	0.69	0.77	0.77	0.77	0.78	0.79	0.78
Number of Bins=6									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.68			0.76			0.77
macro average	0.46	0.44	0.43	0.58	0.57	0.57	0.61	0.59	0.6
weighted average	0.66	0.68	0.66	0.76	0.76	0.76	0.77	0.77	0.77
Number of Bins=7									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.65			0.76			0.79
macro average	0.44	0.43	0.43	0.61	0.6	0.61	0.66	0.64	0.65
weighted average	0.65	0.65	0.65	0.76	0.76	0.76	0.79	0.79	0.79

Forest with R1- score value always greater than 0.7 for author features and combined features. After that Gradient Boosted Tree performed better as compared to Decision Tree classifier.

An interesting observation is that as the number of classes increases, author features perform very close to combined features. This pattern can be interpreted as for large number of classes/bins, the author features can be used instead of combined features which will help in reducing number of total feature required and also reduce the complexity of the system. The results from classification algorithms have shown promising results. The answer to the third research question is that it is possible to categorize tweets in different classes. However, a tradeoff between accuracy and the number of classes should be considered as shown in eq. 1.

$$Accuracy \propto \frac{1}{Number\ of\ Bins(Classes)} \tag{15}$$

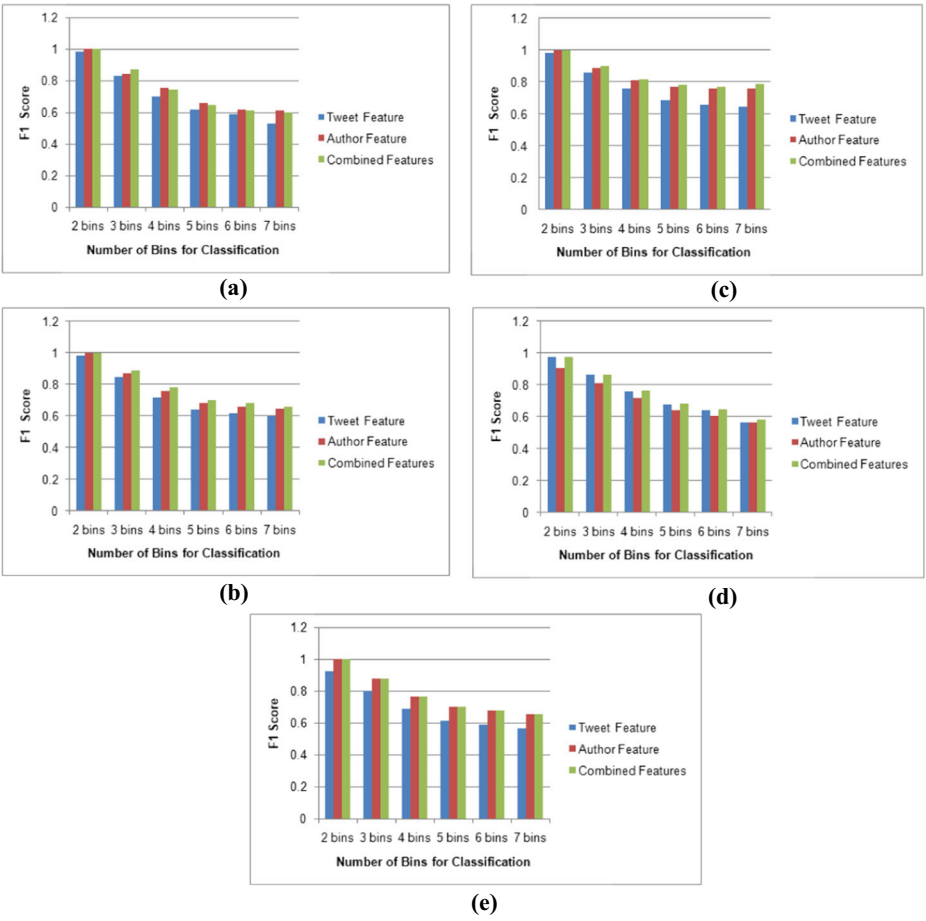


Fig. 9 Performance comparison of Classification algorithms for RQ3. **(a)** Decision Tree. **(b)** Gradient Boosted Tree. **(c)** Random Forest. **(d)** SVM. **(e)** KNN

4.5 Comparison with other works

The comparison of proposed work is given in Table 15. The highlight of the proposed work is feature set proposed have low complexity in implementation.

5 Conclusions and future work

In this paper, an attempt is made to understand the point of view of a user as information processing node and the role of user profiles on Twitter to predict retweets. The criteria of using only Twitter API as the data source and less number of features provides a unique way of looking at the problem of retweet prediction. The Twitter API is the most common method for data collection from Twitter which makes it a natural choice for creating reproducible research work.

Table 12 Performance Metrics for Gradient Boosted Tree Classification

	Tweet Feature			Author Feature			Proposed Combined Features		
Number of Bins=2									
	Precision (P)	Recall (R)	F1-score	Precision (P)	Recall (R)	F1-score	Precision (P)	Recall (R)	F1-score
Accuracy			0.98			1			1
macro average	0.98	0.98	0.98	1	1	1	1	1	1
weighted average	0.98	0.98	0.98	1	1	1	1	1	1
Number of Bins=3									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.88			0.89			0.91
macro average	0.8	0.69	0.68	0.8	0.71	0.71	0.85	0.76	0.77
weighted average	0.86	0.88	0.85	0.87	0.89	0.87	0.9	0.91	0.89
Number of Bins=4									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.78			0.8			0.82
macro average	0.61	0.53	0.5	0.65	0.58	0.56	0.68	0.6	0.59
weighted average	0.75	0.78	0.72	0.78	0.8	0.76	0.79	0.82	0.78
Number of Bins=5									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.7			0.72			0.74
macro average	0.5	0.44	0.4	0.55	0.5	0.45	0.57	0.52	0.49
weighted average	0.67	0.7	0.64	0.71	0.72	0.68	0.72	0.74	0.7
Number of Bins=6									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.67			0.69			0.71
macro average	0.5	0.39	0.35	0.49	0.43	0.4	0.51	0.45	0.43
weighted average	0.66	0.67	0.62	0.68	0.69	0.66	0.69	0.71	0.68
Number of Bins=7									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.61			0.66			0.67
macro average	0.42	0.36	0.34	0.48	0.42	0.41	0.48	0.44	0.43
weighted average	0.62	0.61	0.6	0.68	0.66	0.65	0.68	0.67	0.66

The manually coded features or creating new features using complex algorithms reduces the chances of scaling up and replication for other scenarios. In a recent study [10], it is found that a positive sentiment result in more retweets during natural disasters. However, previous studies [29] have found that negative sentiment increased retweets in the election campaign. In two different domains, same feature resulted in different outcomes. This is an example that some complex features are not good for domain-independent, very large scale fast data processing.

The contribution of this paper is the effort of reducing complexity and the computational requirement for big data analysis of social media data. The ability to use only numerical features is a very fast, scalable, and feasible solution. Two out of three types of features related to retweet prediction are available in Twitter API, from which author features proved to be more significant than tweet content features. The combination of both features produced the best results.

Table 13 Performance Metrics for SVM Classification

Number of Bins=2									
	Tweet Feature			Author Feature			Proposed Combined Features		
	Precision (P)	Recall (R)	F1-score	Precision (P)	Recall (R)	F1-score	Precision (P)	Recall (R)	F1-score
Accuracy			0.977			0.903			0.977
macro average	0.98	0.98	0.98	0.91	0.91	0.9	0.98	0.98	0.98
weighted average	0.98	0.98	0.98	0.92	0.9	0.9	0.98	0.98	0.98
Number of Bins=3									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.863			0.81			0.863
macro average	0.79	0.66	0.62	0.54	0.61	0.57	0.79	0.66	0.62
weighted average	0.85	0.86	0.82	0.72	0.81	0.76	0.85	0.86	0.82
Number of Bins=4									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.76			0.717			0.766
macro average	0.52	0.51	0.46	0.35	0.45	0.39	0.51	0.51	0.47
weighted average	0.7	0.76	0.7	0.56	0.72	0.63	0.69	0.77	0.7
Number of Bins=5									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.674			0.641			0.68
macro average	0.41	0.43	0.37	0.24	0.36	0.29	0.4	0.43	0.38
weighted average	0.63	0.67	0.62	0.46	0.64	0.53	0.61	0.68	0.62
Number of Bins=6									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.639			0.607			0.647
macro average	0.33	0.36	0.31	0.19	0.3	0.23	0.34	0.36	0.31
weighted average	0.6	0.64	0.59	0.43	0.61	0.5	0.59	0.65	0.58
Number of Bins=7									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.567			0.562			0.584
macro average	0.25	0.31	0.26	0.38	0.27	0.22	0.33	0.31	0.27
weighted average	0.54	0.57	0.54	0.54	0.56	0.48	0.58	0.58	0.55

Three new features "author total Activity", "author total activity per year" and "author tweets per year" are easy to compute, useful in capturing the active and passive participation of a user. The ability to scale down any spikes of total activity value is achieved by dividing the number of years of user account age. The same method is used to scale down the spikes in the count of tweets posted by a user. This averaging of total activity count and total tweet count by the number of years of account age is very useful for those users who are not regularly active. This provides an ability to predict for a random user who is not an influential user or a celebrity. Most of the state of the art research works give more importance to influential users. In real time data analysis, every tweet is important and every user profile is useful for accurate prediction of retweets. The proposed features provide better results for every type of users. These features provide an important insight for categorization of users as trustworthy and less trustworthy account. It will form basis for highlighting genuine users from non-genuine looking accounts.

Table 14 Performance Metrics for KNN Classification

Number of Bins=2									
	Tweet Feature			Author Feature			Proposed Combined Features		
	Precision (P)	Recall (R)	F1-score	Precision (P)	Recall (R)	F1-score	Precision (P)	Recall (R)	F1-score
Accuracy			0.923			1			1
macro average	0.93	0.92	0.923	1	1	1	1	1	1
weighted average	0.93	0.92	0.92	1	1	1	1	1	1
Number of Bins=3									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.802			0.875			0.875
macro average	0.67	0.65	0.65	0.75	0.72	0.73	0.75	0.72	0.73
weighted average	0.78	0.8	0.79	0.86	0.88	0.86	0.86	0.87	0.86
Number of Bins=4									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.688			0.766			0.766
macro average	0.51	0.5	0.5	0.6	0.59	0.59	0.6	0.59	0.59
weighted average	0.67	0.69	0.68	0.76	0.77	0.76	0.75	0.77	0.76
Number of Bins=5									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.614			0.7			0.699
macro average	0.42	0.42	0.42	0.51	0.51	0.51	0.51	0.51	0.51
weighted average	0.61	0.61	0.61	0.7	0.7	0.7	0.69	0.7	0.7
Number of Bins=6									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.589			0.677			0.677
macro average	0.36	0.36	0.35	0.45	0.45	0.45	0.45	0.45	0.45
weighted average	0.57	0.59	0.58	0.67	0.68	0.67	0.67	0.68	0.67
Number of Bins=7									
	Tweet Feature			Author Feature			Proposed Combined Features		
	P	R	F1	P	R	F1	P	R	F1
Accuracy			0.564			0.654			0.655
macro average	0.34	0.33	0.33	0.43	0.42	0.42	0.43	0.42	0.42
weighted average	0.55	0.56	0.56	0.65	0.65	0.65	0.65	0.66	0.65

The proposed method of retweet prediction can easily predict whether a tweet will be retweeted or not. The ability to predict the exact number of retweets is not achievable with these assumptions and feature sets, but it can be used with some other features to reduce the margin of error. The classification of tweets based on retweet count is possible, however, it is difficult to predict accurately with a large number of classes. The fine grain classes come with the drawback of poor accuracy and the small number of classes results in high accuracy but a very large range as one label which is practically not useful for multiclass classification.

In future work, the proposed feature sets will be applied for the categorization of user accounts based on activities, user account role as hub or crowd [17], and the impact of information overload on social media users. For categorization of fake and genuine accounts [32] based on their user profile features, the proposed three features will be used. The proposed

Table 15 Comparison with recent works on information sharing techniques

Research Work	Year	Domain Independent	Independent of Network features	Independent of NLP features	Handling Big Data Streams	User Profile Features included	Independent of Historical Tweets or Cascade Retweets	Complexity of features used
Ensemble learning approach [7]	2020	Yes	No	No	No	Yes	Yes	High
A media synchronicity theory for effective communication during disasters [42]	2020	Yes	Yes	No	No	No	No	High
Covid 19 information diffusion model [15]	2020	No	No	Yes	No	No	No	High
Popularity prediction using heterogeneous Bass model [21]	2020	Yes	Yes	No	No	Yes	Yes	High
Hawkes process and topic modeling [17]	2020	No	Yes	No	No	No	No	High
Network analysis for predicting influence of user nodes [22]	2019	No	Yes	Yes	No	No	No	Low
Temporal features for influential node detection [6]	2019	Yes	No	Yes	No	No	No	Medium
Visualizing reposting process [9]	2019	Yes	No	No	No	No	No	High
Spatiotemporal features for sentiment and retweet analysis [10]	2020	No	Yes	No	No	No	Yes	High
Text information retrieval for retweet and like prediction [13]	2020	No	Yes	No	No	No	Yes	High
Analysis of user nodes for information diffusion patterns [18]	2020	No	Yes	No	No	No	No	High
unified factorization model for retweet prediction [43]	2020	Yes	No	No	Yes	Yes	No	High
Tweet content analysis for increased retweets [24]	2019	No	Yes	No	No	Yes	No	High
A user retweet prediction method for hot topics [14]	2021	No	No	No	No	No	No	High
COVID-19 pandemic machine learning to measure twitter users' perceptions [23]	2021	No	Yes	No	No	No	No	High
Retweet Prediction based on Topic, Emotion and Personality [21]	2021	Yes	No	No	No	Yes	No	Medium
Value-Based Retweet Prediction on Twitter [29]	2021	Yes	Yes	No	No	Yes	No	Medium
Proposed Work	2022	Yes	Yes	Yes	Yes	Yes	Yes	Low

profile features will also be used for opinion mining, sentiment analysis and fake account detection.

Acknowledgments The author wishes to thank the Design and Innovation Center, Chandigarh, UIET, Panjab University for providing computational resources for big data analysis on the Apache Spark cluster.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose. The authors have no conflicts of interest to declare that are relevant to the content of this article. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no financial or proprietary interests in any material discussed in this article.

References

1. Adewole KS, Anuar NB, Kamsin A, Sangaiah AK (2019) SMSAD: a framework for spam message and spam account detection. Multimed Tools Appl 78(4):3925–3960
2. Aggarwal, A., Rajadesingan, A., & Kumaraguru, P. (2012). PhishAri: automatic realtime phishing detection on twitter. In 2012 eCrime researchers summit :1-12IEEE.
3. Alsaleh M, Alarifi A, Al-Salman AM, Alfayez M, & Almuhaysin A (2014). Tsd: detecting sybil accounts in twitter. 13th international conference on machine learning and applications :463-469.IEEE.
4. Antonakaki D, Fragopoulou P, Ioannidis S (2021) A survey of twitter research: data model, graph structure, sentiment analysis and attacks. Expert Syst Appl 164:114006
5. Arpaci I, Alshehabi S, Al-Emran M, Khasawneh M, Mahariq I, Abdeljawad T, Hassanien AE (2020) Analysis of twitter data using evolutionary clustering during the COVID-19 pandemic. Comput Mater Contin 65(1):193–203
6. Bhowmick AK, Gueuning M, Delvenne JC, Lambiotte R, Mitra B (2019) Temporal sequence of retweets help to detect influential nodes in social networks. IEEE Trans Comput Soc Syst 6(3):441–455
7. Chen L, Deng H (2020) Predicting user retweeting behavior in social networks with a novel ensemble learning approach. IEEE Access 8:148250–148263
8. Chen G, Kong Q, Xu N, Mao W (2019) NPP: a neural popularity prediction model for social media content. Neurocomputing 333:221–230
9. Chen S, Li S, Chen S, Yuan X (2019) R-map: a map metaphor for visualizing information reposting process in social media. IEEE Trans Vis Comput Graph 26(1):1204–1214
10. Chen S, Mao J, Li G, Ma C, Cao Y (2020) Uncovering sentiment and retweet patterns of disaster-related tweets from a spatiotemporal perspective–a case study of hurricane Harvey. Telematics Inform 47:101326
11. Chu Z, Gianvecchio S, Wang H, Jajodia S (2012) Detecting automation of twitter accounts: are you a human, bot, or cyborg? IEEE Trans Dependable Secure Comput 9(6):811–824
12. Chung W, Toraman C, Huang Y, Vora M, & Liu J (2019). A Deep Learning Approach to Modeling Temporal Social Networks on Reddit. In 2019 IEEE International Conference on Intelligence and Security Informatics (ISI) :68–73. IEEE.
13. Daga I, Gupta A, Vardhan R, Mukherjee P (2020) Prediction of likes and retweets using text information retrieval. Procedia Comput Sci 168:123–128
14. Dinh L, Parulian N (2020) COVID-19 pandemic and information diffusion analysis on twitter. Proc Assoc Inf Sci Technol 57(1):e252
15. Duan M, Li K, Liao X, Li K (2017) A parallel multiclassification algorithm for big data using an extreme learning machine. IEEE Trans Neural Netw Learn Syst 29(6):2337–2351
16. Dutta HS, Dutta VR, Adhikary A, Chakraborty T (2020) HawkesEye: detecting fake retweeters using Hawkes process and topic modeling. IEEE Transactions on Information Forensics and Security 15:2667–2678
17. Fan C, Jiang Y, Yang Y, Zhang C, Mostafavi A (2020) Crowd or hubs: information diffusion patterns in online social networks in disasters. Int J Disaster Risk Reduct 46:101498
18. Firdaus SN, Ding C, Sadeghian A (2018) Retweet: a popular information diffusion mechanism–a survey paper. Online Soc Netw Media 6:26–40

19. Firdaus SN, Ding C, Sadeghian A (2019) Topic specific emotion detection for retweet prediction. Int J Mach Learn Cybern 10(8):2071–2083
20. Gao X, Zheng Z, Chu Q, Tang S, Chen G, Deng Q (2019) Popularity prediction for single tweet based on heterogeneous bass model. IEEE Trans Knowl Data Eng:1
21. Hemphill L, Hedstrom ML, Leonard SH (2021) Saving social media data: understanding data management practices among social media researchers and their implications for archives. J Assoc Inf Sci Technol 72(1): 97–109
22. Hemsley J (2019) Followers retweet! The influence of middle-level gatekeepers on the spread of political information on twitter. Policy Internet 11(3):280–304
23. Jain DK, Kumar A, Sharma V (2020) Tweet recommender model using adaptive neuro-fuzzy inference system. Futur Gener Comput Syst 112:996–1009
24. Jalali NY, Papatla P (2019) Composing tweets to increase retweets. Int J Res Mark 36(4):647–668
25. Jung AK, Ross B, Stieglitz S (2020) Caution: rumors ahead—a case study on the debunking of false information on twitter. Big Data Soc 7(2):2053951720980127
26. Lee S, & Kim J (2014) Early filtering of ephemeral malicious accounts on Twitter.Computer communications 54:48-57.
27. Lee J, Xu W (2018) The more attacks, the more retweets: Trump's and Clinton's agenda setting on twitter. Public Relat Rev 44(2):201–213
28. Lymperopoulos IN (2021) RC-tweet: modeling and predicting the popularity of tweets through the dynamics of a capacitor. Expert Syst Appl 163:113785
29. Miller Z, Dickinson B, Deitrick W, Hu W, Wang AH (2014) Twitter spammer detection using data stream clustering. Inf Sci 260:64–73
30. Murshed BAH, Al-Ariki HDE, Mallappa S (2020) Semantic analysis techniques using twitter datasets on big data: comparative analysis study. Comput Syst Sci Eng 35(6):495–512
31. Nesi P, Pantaleo G, Paoli I, Zaza I (2018) Assessing the reTweet proneness of tweets: predictive models for retweeting. Multimed Tools Appl 77(20):26371–26396
32. PV, S., & Bhanu, S. (2020) UbCadet: detection of compromised accounts in twitter based on user behavioural profiling. Multimed Tools Appl 79:1–37
33. Rousidis D, Koukaras P, Tjortjis C (2020) Social media prediction: a literature review. Multimed Tools Appl 79(9):6279–6311
34. Safari RM, Rahmani AM, Alizadeh SH (2019) User behavior mining on social media: a systematic literature review. Multimed Tools Appl 78(23):33747–33804
35. Scott, Jason, and Sketch the Cow. "Archiveteam-Twitter-Stream-2018-08 : Free Download, Borrow, and Streaming." Internet Archive, Archive Team: The Twitter Stream Grab, 6 Dec. 2012, 01:03:03, archive.org/details/archiveteam-twitter-stream-2018-08.
36. Sequiera R, & Lin J (2017) Finally, a downloadable test collection of tweets. In proceedings of the 40th international ACM SIGIR conference on Research and Development in information retrieval :1225-1228.
37. Shyni CE, Sundar AD, Ebby GSE (2016) Spam profile detection in online social network using statistical approach. Asian J Inf Technol 15(7):1253–1262
38. Singh SK, Cha J, Kim TW, Park JH (2021) Machine learning based distributed big data analysis framework for next generation web in IoT. Comput Sci Inf Syst 18(2):597–618
39. Son J, Lee HK, Jin S, Lee J (2019) Content features of tweets for effective communication during disasters: a media synchronicity theory perspective. Int J Inf Manag 45:56–68
40. Son J, Lee J, Oh O, Lee HK, Woo J (2020) Using a heuristic-systematic model to assess the twitter user profile's impact on disaster tweet credibility. Int J Inf Manag 54:102176
41. Tardelli S, Avvenuti M, Tesconi M, Cresci S (2020) Characterizing social bots spreading financial disinformation. In: International conference on human-computer interaction :pp. Springer, Cham, pp 376–392
42. Tian Y, Fan R, Ding X, Zhang X, Gan T (2020) Predicting rumor retweeting behavior of social media users in public emergencies. IEEE Access 8:87121–87132
43. Wang S, Li C, Wang Z, Chen H, Zheng K (2020) BPF++: a unified factorization model for predicting retweet behaviors. Inf Sci 515:218–232
44. Yang C, Harkreader R, Gu G (2013) Empirical evaluation and new design for fighting evolving twitter spammers. IEEE Trans Inf Forensics Secur 8(8):1280–1293
45. Zheng X, Zeng Z, Chen Z, Yu Y, Rong C (2015) Detecting spammers on social networks. Neurocomputing 159:27–34
46. Zhou F, Xu X, Trajcevski G, Zhang K (2021) A survey of information cascade analysis: models, predictions, and recent advances. ACM Comput Surv 54(2):1–36
47. Zola P, Cortez P, Carpita M (2019) Twitter user geolocation using web country noun searches. Decis Support Syst 120:50–59

48. Zubiaga A (2018) A longitudinal assessment of the persistence of twitter datasets. J Assoc Inf Sci Technol 69(8):974–984
49. Zubiaga A, Aker A, Bontcheva K, Liakata M, Procter R (2018) Detection and resolution of rumours in social media: a survey. ACM Comput Surv 51(2):1–36