Multimedia Tools and Applications (2022) 81:7669–7683
https://doi.org/10.1007/s11042-022-12221-7

# Semi-supervised labeling: a proposed methodology for labeling the twitter datasets

Tabassum Gull Jan[1] · Surinder Singh Khurana[1] · Munish Kumar[2]

Received: 21 May 2021 / Revised: 1 December 2021 / Accepted: 10 January 2022 /
Published online: 28 January 2022


**Abstract**
Twitter has nowadays become a trending microblogging and social media platform for news and discussions. Since the dramatic increase in its platform has additionally set off a dramatic increase in spam utilization in this platform. For Supervised machine learning, one always finds a need to have a labeled dataset of Twitter. It is desirable to design a semi-supervised labeling technique for labeling newly prepared recent datasets. To prepare the labeled dataset lot of human affords are required. This issue has motivated us to propose an efficient approach for preparing labeled datasets so that time can be saved and human errors can be avoided. Our proposed approach relies on readily available features in real-time for better performance and wider applicability. This work aims at collecting the most recent tweets of a user using Twitter streaming and prepare a recent dataset of Twitter. Finally, a semi-supervised machine learning algorithm based on the self-training technique was designed for labeling the tweets. Semi-supervised support vector machine and semi-supervised decision tree classifiers were used as base classifiers in the self-training technique. Further, the authors have applied K means clustering algorithm to the tweets based on the tweet content. The principled novel approach is an ensemble of semi-supervised and unsupervised learning wherein it was found that semi-supervised algorithms are more accurate in prediction than unsupervised ones. To effectively assign the labels to the tweets, authors have implemented the concept of voting in this novel approach and the label pre-directed by the majority voting classifier is the actual label assigned to the tweet dataset. Maximum accuracy of 99.0% has been reported in this paper using a majority voting classifier for spam labeling.

**Keywords** Twitter · Spam labeling · Clustering · Spam detection · Tweets

✉ Munish Kumar
munishcse@gmail.com

Extended author information available on the last page of the article

## 1 Introduction

Twitter is an online social communication website that empowers clients to send and share information as 140 characters in length messages called tweets. These tweets can contain words, photographs, feelings, exceptional character recordings, links, etc. The popularity of Twitter is increasing rapidly having 192 million daily active users that almost post trillions of tweets per day. In 2020, the global ranking of Twitter is 17 among all social networking sites with 10% of worldwide social media user base. Among top Twitter using countries, India ranks on third position[1]. Based on the client traffic, the worldwide positioning of twitter is 12 among every one of the sites accessible on Twitter. Twitter spams have long been a critical issue that needs to be addressed. Twitter Spam detection has been one of the most important topics in the field of social network security. A series of approaches have been used by most of the pioneer researchers for exploring spams in OSNs. The main concern of the recent works carried has been application of machine learning techniques for Twitter spam detection, Categorization of Spammers in OSNs, Types of Spam profiles in Twitter and so on. The problem pertaining to the existing approaches is that they are based on supervised machine learning which solely rely on a labeled dataset. To label the data manually is quite difficult as it needs a lot of manpower and time. Also, manual labeling leads to inter observer variability. This is because for a person to label a particular tweet as spam or non-spam is highly dependent on the nature of interest and behavior of the observer. As for one person it can be spam and for other it won't be. Also, the available datasets are too old and small in size, and for a machine learning model to achieve higher accuracy, the size of training dataset should be sufficiently large enough so that the learning model can learn the patterns in the dataset well and later on is quite accurate in its prediction. Hence there exists a need to prepare a recent dataset of Twitter that is either based on unsupervised learning or semi-supervised learning. Many researchers worked in the field of analyzing tweets but to prepare a recent dataset of twitter and then assign labels to the dataset based on semi-supervised learning has not yet been done.

In this paper, we will first of all discuss the various approaches used for the collection raw dataset of tweets from Twitter using Streaming APIs. Then we explained the various methods used for extraction of tweets and lastly, we have discussed how to apply the semi-supervised approach based on the self-training technique to label the tweets. Proposed method is quite a new approach quite different from existing methods. Initially after downloading the dataset, proper understanding of information available in the tweet is done, understanding the significance of various features made available by the Twitter platform, and how their values are effective in spam labeling. Extensive preprocessing of tweets has been done to eliminate the features having lower weights concerning spam labeling. It has been observed that tweets ensure high intraclass similarity and low inter-class similarity. Further, to label the similar tweets belonging to a particular spam category, we have proposed a quite new technique for labeling tweets using a semi-supervised algorithm based on self-training technique.

The noteworthy contributions of this novel semi-supervised approach used for labelling the unlabeled tweets are summarized as:

[1] https://backlinko.com/twitter-users

- We have prepared the recent dataset from Twitter using streaming APIs by extracting all maximum possible features from the tweets, which can later be used for different purposes in the future.
- We have represented the preprocessed tweets using proposed method.
- We have introduced a semi-supervised approach for labeling the large datasets of tweets using three machine learning algorithms, namely, semi-supervised Support Vector Machine classifier, Decision Tree classifier, and K-Means Clustering algorithm.
- We have introduced an ensemble of semi-supervised and unsupervised learning by applying the voting classifier-based approach to assign final label to the tweet dataset.

So, the main purpose of the paper is to prepare a recent dataset of Twitter and then label this dataset using semi-supervised approaches. The process of labeling is aimed at assigning labels to tweets (whether spam and non-spam tweets) by incorporating the new features provided by Twitter platform nowadays. In the next section, we demonstrate why semi- supervised labeling is needed and various Twitter spam detection approaches used.

## 2 Why semi-supervised labeling?

Having tremendous amounts of unlabeled data often poses a problem lack of specified accuracy ranges. An ideal and best solution to the problem is having a large labeled dataset, which in turn requires a lot of man power and time. Semi-supervised labeling is seeking a balance between the both by tackling both accuracy issues and need to have an extensively labeled large dataset. The semi-supervised learning is employed by making the machine to use the small amount of labeled dataset to predict or categorize the values of unlabeled data with better accuracy. The typical characteristics of SSL dataset are

- Percentage of unlabeled data should be large enough.
- Input –Output Proximity: SSL aims to predict the label of unlabeled data based on labeled data in its proximity.

## 3 Twitter spam detection approaches and related works

With recent advancements in information processing technologies, techniques and procedures used to detect spammers in social networks like Facebook, twitter have also got matured enough to evade the detection process as per Twitter rules, some common tactics followed by spam accounts that post spam tweets include:

- Posting harmful and malicious links (including links to phishing or malware sites).
- Abusing the answer or referencing somebody to present undesirable messages on different records.
- Aggressive after conduct that is mass after and mass unfollowing for looking for consideration.
- Creating multiple accounts either manually or by automated tools to hide the real identity of the account holder.
- Posting links with unrelated tweets.

- Repeatedly posting duplicate tweets.

But in reality, such spammers are becoming more intelligent and over smart. All this has been possible only by developing and using more robust and highly secure mechanisms to avoid detection. To address such security related problems researchers proposed different approaches from time to time based on different set of features like some of them rely on tweet content, some on user profile, some on graph-based features (like connectivity and distance) and some on URLs embedded in tweet content. Benevenuto *et al.* [4] analyzed the tweets based on account and content-based features for spam detection using Support Vector Machine. Further the same procedure in platforms also employed like Facebook, Instagram and MySpace by training Random Forest Classifier [4, 14].

Eshraqi *et al.* [6] proposed a method for detecting spam tweets in Twitter using a data stream clustering algorithm. These authors have keenly observed and analyzed various statistical and analytical features of tweets for tweet spam detection process. Labeled dataset has been collected from Spam accounts as spam and legitimate accounts as non-spam. For preprocessing the dataset, software called "RapidMiner" has been used. The output tweets were then given input to DenStream Algorithm for clustering spam tweets. Experimental evaluation has shown that when the algorithm was set properly, accuracy and precision was supposed to improve in comparison to past works done using classification algorithms.

Liu and Wang [11] explored the weakness and research gaps of existing works and are nowadays working on issues related to the Twitter spam detection techniques. New classification methods have been proposed that addresses these issues based on deep learning algorithms. Deep learning was applied on this dataset and it has been observed that performance evaluation parameters like accuracy and precision have values higher than 90%. Al-Zoubi *et al.* [2] created spam profile detection models that depend on a lot of basic and publicly available features on Twitter. It was concluded that promising outcomes can be acquired utilizing the Naive Bayes and Decision Tree Classifier. The outcomes uncovered those suspicious words and the repeated words in tweet content impact the exactness of the recognition procedure, irrespective of tweet language.

Abu-liash and Fazil [1] presented a hybrid approach for detecting automated spammers by amalgamating community-based features with other feature categories, namely metadata, content, and interaction-based features. They used nineteen different features, including six newly defined features and two redefined features for learning three classifiers, namely, random forest, decision tree, and Bayesian networks, on a real dataset that comprises non-spam users and spammers. They additionally examined the discriminating power of different feature categories. They inferred that interaction and community-based features are the most effective for spam detection, whereas metadata-based features are demonstrated to be the least successful. Peikari *et al.* [12] have proposed a technique for clustering then label a semi-supervised learning approach for pathology image classification. The proposed method has been the first and foremost method used in the field of labeling the images using semi-supervised approach. The idea behind this unique proposed method was to first of all cluster whole data space of images into clusters and then assign labels to these clustered images using a semi-supervised labeling. The weakest point of this novel approach was dependence on inter-observer variability for labeling the images for training phase. Also, an insufficient number of labeled data points that is scare data available often leads to failure of the clustering process. Sedhai and Sun [13] proposed a semi-supervised spam detection technique for twitter stream by taking into account the features of the tweet level. The proposed model has 2 modules

namely "the spam detection module" and the other module "the model update". The former was operating in real-time while later one was operating in batch mode.

Sun *et al.* [15] developed a near real time system-based spam detection model using machine learning. The empirical study was performed using nine machine learning algorithms with large amounts of datasets. Scalability of all algorithms was examined using different number of CPU cores. The system employed parallel computing technique and hence the speed of detection has dramatically been increased. Further the system can combat huge number of intelligent spammers. The proposed system can be used as a tweet collection tool thereby allowing researchers to analyze the performance of trained classifiers in realistic scenarios. [3] proposed a hybrid classification approach for Twitter spam detection in real-time Twitter datasets using SMOTE (Synthetic Minority Over Sampling Technique) and DE (Differential Evolution) strategies. The dataset (Twitter Spam) used in this research has been prepared by NSCLab in 2014 and it contains only 13 attributes. SMOTE tackled the imbalanced class distribution and DE was used to tune the hyper parameters of Random Forest classifier. The classification accuracy of optimized random forest classifier was quite high with excellent F1-Score of 98.97% which itself explains the high efficiency of the proposed method.

## 4 Proposed method

The proposed method is quite new approach for labeling the Twitter dataset using semi-supervised approaches. In this approach, we are first of all going to download the recent data set from Twitter using streaming APIs. From the collected dataset, we have performed extensive preprocessing to extract all the maximum possible features that in one way or the other way are related to spam detection and are directly available from Twitter data. Further, we calculated some features from the directly available features of Twitter that have a positive weightage regarding spam detection as per existing works. Using this recent dataset, we will examine the patterns, trends, tactics and examples followed by the spammers and how spammers turn out to be more intelligent with time. Using an existing dataset to get the labeled spammer and legitimate user Ids of twitter and then fetch recent data corresponding to those Ids and prepare the latest recent dataset accordingly. After dataset design, we are going to apply self-training based semi-supervised learning algorithm for labeling. We have used Support Vector Classifier and Decision Tree Classifier as the base classifier for self-training SSL approach. Further we have applied Semi-supervised labeling with K means clustering algorithm by initializing the K cluster centers equal to 2 (one cluster for spam and other for non-spam). All data points are assigned to closest centers and thus updating centroids till cluster centers stop varying. The labels predicted by the three algorithms (semi-supervised and unsupervised) are ensembled used majority voting scheme and that label is treated as the final label to the dataset.

## 5 Methodology

The overall flow of this novel predictive labeling approach used for labeling the spam and non-spam tweets is depicted in Fig 1. The noteworthy steps followed to carry out the methodology are as follows:

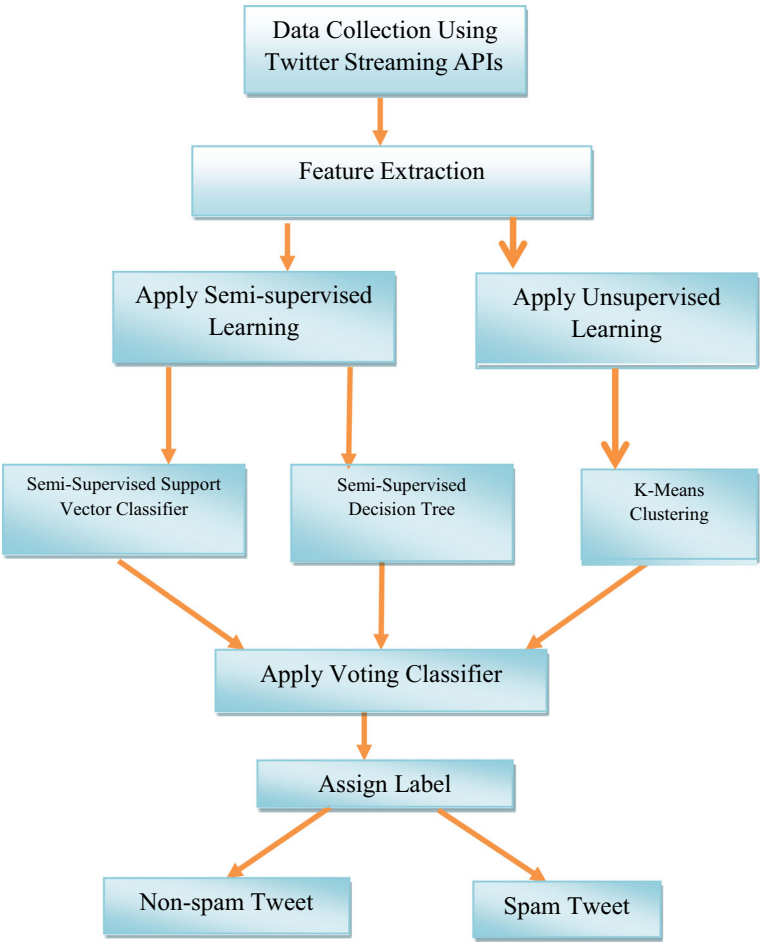


**Fig. 1** Block diagram of the proposed work

- Download and prepare recent dataset of tweets using Twitter Streaming APIs.
- Preprocess the data and extract maximum possible features from the dataset.
- Design a semi-supervised machine learning model for labeling of the spam tweets.
- Apply K means clustering to label tweets in unsupervised fashion.
- Combine semi-supervised and unsupervised approaches using Majority Voting Classifier.
- Analyze the performance of the results obtained.

### 5.1 Data collection and preparation

In this phase, we have downloaded and prepared the recent dataset of Twitter using streaming APIs. To fetch the recent dataset from Twitter using streaming APIs, the primary thing required is Twitter developer access. Once access is granted, the user will be given the access

token and secret keys. Using these access tokens and keys we can fetch recent data related to the articular tweet handle. Programs have been written in the python programming language to use the Twitter APIs. After downloading the dataset from Twitter that is available in .json format, it needs to be further processed to extract maximum possible features and save the extracted features in the .csv format. A large dataset containing 50 features has been prepared from the downloaded data. The dataset contains 7,64,416 tweets out of which 6,49,982 belong to legitimate user tweets and 1,14,434 tweets belong to spam user tweets. The sample dataset was taken using a random sampling method in which we selected a total of 11674 tweets out of which 5272 tweets belong to non-spam users and 6402 tweets have a place with real clients. Since collected dataset was large in size and it would lead to more computational cost. That is why small sample dataset was used for the design of semi supervised method for labeling tweets.

### 5.2 Feature extraction

In this phase, from the prepared dataset in (comma separated values) csv format, we select and extract the features used for spam detection in Twitter. Out of 50 features in the dataset, only 22 features are significantly related to spam detection in tweets. All features that have been used in the related works in the literature are extracted and a dataset is prepared for semi-supervised machine learning algorithms. Table 1 shows the various features that were used for labeling the Twitter dataset.

From the related work and the papers reviewed about Twitter spamming, the features to observe spam tweets from non-spam tweets are about the client, their practices, and the content within the tweets themselves. The following subsections describe all the features that are used in detail related to spam tweets, all according to the Twitter rules.

A. ***Tweet Based Features***

- *Number of @mentions per tweet:* This element encourages us to decide the mention usage in a tweet. As spammers usually send a lot of spam messages with @ client name as a promotion technique to flood the client timetable with such tweets. This check is less for authentic tweets and more for spam tweets [5].

**Table 1** Features used for Labeling the Twitter Dataset

| Tweet based features | User/ Account based features |
|---|---|
| Number of @ mentions per tweet | Statuses Count |
| Number of Hashtags per tweet | Listed count |
| Retweet count | Follower Count |
| Tweet time | Friends Count |
| Number of URLs per tweet | Following Count |
| Text | Favourites Count |
| Length of tweet | Verified |
| Digit count per tweet | Follower/Friends Ratio |
| Capital words per tweet | Username |
| Spam words per tweet | |
| Repeated words per tweet | |
| Favorited | |
| Tweet id | |

- *Number of Hashtags per tweet:* This feature is used to determine the concentration of hashtags in the tweet. Since spammers will in general send plenty of spam messages by making the tweets slanting by utilizing most extreme hashtags. So, as to highlight a particular by seeking the attention of more in case of an advertisement strategy. This count is higher for spam tweets as compared to normal tweets [5].
- *Retweet count:* Retweets are tweets of other users preceded by the word "RT". Spam tweets are not normally retweeted so they have a less retweet consider contrasted with non-spam tweets [10]. So, we use this feature to distinguish between spam and non-spam tweets.
- *Tweet time:* This element will be useful to decide the tweet time example, pursued by a spammer in his tweets. Spammers will in general post tweets all the more regularly by following a specific example [2].
- *Number of URLs per tweet:* This feature is used to determine the number of URLs per tweet as Spammers will in general have more URLs per tweet to do phishing and malware downloading using these links provided [2, 8]. This value is greater for spam tweet than a normal tweet.
- *Text:* This is a significant component of breaking down whether a tweet is spam or not. As spammers will in general post tweets that may contain unlawful content, pornographic content, duplicate words, violation of privacy in the tweet, etc. [7, 9].
- *Length of the tweet:* Spammers tend to post shorter messages as compared to legitimate users [8]. Since it is a dynamic feature and depends on the tweet user posting the tweet. Since spammers might have become smarter and they might have changed the trend so to use this feature we will better analyze the trend used by recent spammers in tweets.
- *Number of digit usage per tweet:* Legitimate users tend to have a smaller number of digits count per tweet. So, digit count for spam tweet has value greater as compared to non-spam tweets [9].
- *The number of capital words per tweet:* Spammers will in general have enormous use of capital words in their tweets. This isbecausethey need to cause the appearance of tweet not the same as would be expected tweet so it to can move toward becoming everybody focal point to peruse it. This feature is the newly introduced feature in this work based on the manual analysis of the tweet.
- *The number of spam words per tweet:* Spammers will in general have high concentrations of spam words in their tweets. So, for spam tweets spam count is greater than normal tweets [5].
- *The number of repeated words per tweet:* Spam tweets will in general have high concentrations of repeated words. So, to repeat words per tweet count is greater for spam tweets and less for normal tweets [2].
- *Favorited:* Spam tweets are not usually favorited by the users. So, spam tweets have less probability of being favorited by the users in comparison to legitimate tweets. This is the new feature added in this work. By manual analysis of tweet, we conclude spam tweets are less favorited by users than legitimate tweets.
- *Tweet Id:* This is the one-of-a-kind identifier used to recognize the tweets posted by the client.

B. ***User /Account Based Features***

- *Statuses count:* This feature is very helpful to distinguish between a spam tweet user and a normal tweet user. Since spam records are normally obstructed by Twitter rapidly in the

wake of being recognized so their age is less when contrasted with authentic clients. So, spam records will in general have a less status count as compared to the legitimate users [8].

- Listed count: This element will assist us with segregating among spam and real tweet clients. As spam accounts as subscribed to more user groups, hence listed count is greater for spam users than for legitimate ones [8]. To analyze the trend followed for current spam tweets related to listed count we incorporated this feature.
- *Follower count:* This component will assist us with distinguishing between spam tweet user and normal tweet user. Since spammers will in general have more followers so this count is more for spam tweets than for non-spam [5, 8].
- *Friends count:* This element will assist us with analyzing to recognize spam tweet user and normal tweet user. Since spammers tend to have less friends count as compared to legitimate tweet users [8].
- *Following count:* Spammers tend to have low following users in comparison to legitimate users [5, 9].
- *Favourites count:* This component is useful to examine whether a tweet has been liked by how many users. Legitimate tweets are favorited more often as compared to spam tweets so favorites count for spam tweets is less as compared to non-spam tweets [9].
- *Verified:* This element encourages us to recognize spam and non-spam tweet account according to Twitter standards accounts that are checked are treated as real and those that come up short are treated as spam profiles. Verified accounts have a blue tick in their account description. Verified profiles usually don't post spam tweets. As spam tweets tend to have verified feature set to false [9].
- *Follower/Friends Ratio:* Ratio between number of friends and the number of followers is used to calculate this feature. If the result of the ratio is too small, then the probability of being a spam account will increase.
- *Username:* Name of the account user. This feature is used to uniquely identify a particular user.

### 5.3 Algorithm

The pseudo code for the self-training based semi-supervised technique used for labeling the spam and non-spam tweets is explained below. The algorithm relies on the assumption that one's own high confidence predictions are correct and threshold for selection is taken as 0.90 using in majority of recent works. The conceptual level block diagram of the algorithm is shown below Fig. 2.

**Step 1:** Let L be the set of labelled data and U be set of un-labelled data, 'h' underlying classifier and T is the threshold for selection

**Step 2:** Repeat (U! = empty).

***Step 3:*** *Train classifier h with training data L.*

***Step 4:*** *Classify data in U with h.*

***Step 5:*** *For each Xi ∈ U*

I. *Assign label based on classification confidence.*
II. *Sort newly labeled examples based on confidence*
III. *Find a subset U' of U with the most confident scores*

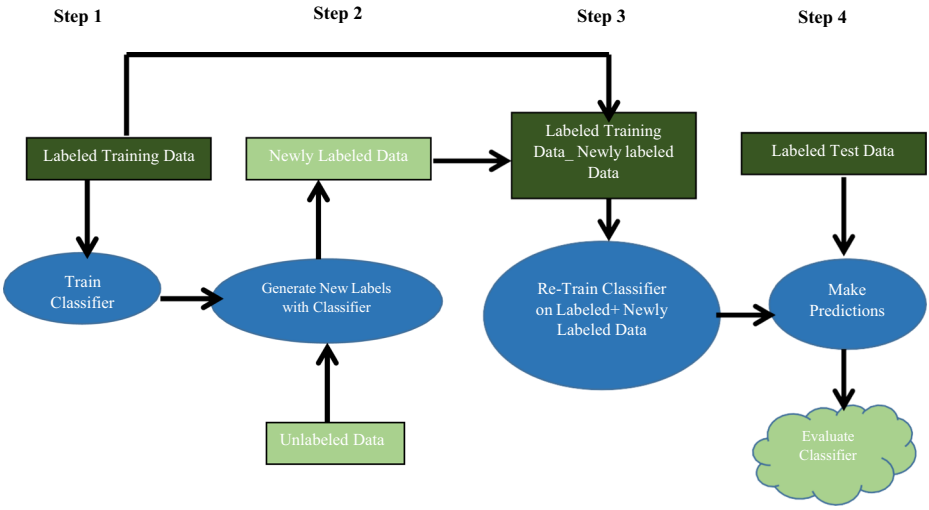


**Fig. 2** Block diagram of self-training algorithm

IV. $L + U' = L$
V. $U - U' = U$

***Step 6:*** *Retrain classifier h with new training set.*

### 5.4 Label the tweets using semi-supervised labeling

In this phase, we have designed a semi-supervised approach based on the self-training technique to label the Twitter datasets. The dataset is assigned a label initially based on tweet users. For example, if a tweet user is legitimate, it is expected and assumed that every one of its tweets is non-spam, and in the event that it is spammer every one of its tweets is spam. But this assumption cannot be always true and we cannot treat this label as the true label of the dataset. To assign the actual label to the dataset we have used semi-supervised labelling. Two algorithms were used as base classifiers for this labelling technique. Firstly, we apply the semi-supervised support vector classifier as a base classifier for the self-training technique. In this technique, we kept the threshold value equal to 0.90 for selecting the high confidence predictions for each iteration. We have applied the SVC algorithm using k fold validation with a number of splits equal to 5 to predict the labels based on statistical properties of the tweet and tweet user. The predictions of each fold for the test split are saved. At the end of 5 folds, all the test predictions are combined to form the predicted dataset. Similarly, we have applied the self-training-based decision tree classifier with k fold validation with a number of splits equal to 5 on the same dataset with the same set of features and the same threshold for selection of high confidence predictions at each selection step. After that, the third algorithm applied to the dataset was the K means clustering algorithm based on tweet content. After the clusters are obtained, we assigned the label to tweets based on the label assigned by K means clustering algorithm. Then we applied a voting classifier to predict the final label to the dataset using the predictions of the

three classifiers as illustrated in Fig. 1. The label assigned by the voting classifier is treated as a true label of the dataset that can be used for further analysis.

## 6 Experimental results and discussion

The proposed work is a very new approach in this field which planned for structuring a semi-supervised labeling technique for labeling a huge dataset where manual labelling is merely impossible and furthermore expensive. All experiments were carried out on PC with 32 GB RAM. The code snippets for all models were written in Python language and run-on Anaconda Jupyter 64-bit software. In this section, we have analyzed the dataset and assessed the performance of the proposed technique. The computational complexity of the proposed technique is O ($n^2$). We have designed the three models for labeling the tweets. The first one is K means clustering algorithm which performs tweet labeling based on tweet content. The second one is a support vector machine classifier and the third one is a decision tree classifier and both work on the tweet and account- based features. In both of these algorithms, we assign labels to the tweet only if its accuracy is more than 90%. Further it has been found that the semi-supervised labeling based on self-training approach using SVC and DT classifier is really best than unsupervised labeling using K means clustering. This approach can also be viewed as an ensemble of semi-supervised and unsupervised approach where the labels predicted by all the three algorithms were combined using the majority voting scheme. The label predicted by the voting classifier is the actual label we assign to the dataset. The cluster validation indices for the K means clustering algorithm shows a significant and effective value. As the Jaccard Similarity value of 0.859 approximately indicates that tweets in the two clusters are completely different (Since value equal to 1 indicates two clusters are well separated). Similarly, the mean square error or mean squared deviation between the two clusters is minimum and value equal to 0.149.

Out of a total of 11674 tweets, the count of actual non-spam tweets is 5272 and the spam tweets count is 6462. Applying the semi-supervised approach of Support Vector Classification (SVC) by accepting the 20% of data is labeled for each training dataset in each fold and then increasing the labeled dataset using a self-training approach for each fold of cross-validation. After training the classifier using 5-fold cross-validation and combining the results of all folds. Then after analyzing the predicted results, the classifier predicts a total of 6323 tweets as spam tweets and 5351 as non-spam tweets. The predicted accuracy of SVC classifier obtained is 97%. The confusion matrix for the same is shown in Table 2. Applying the semi-supervised Decision Tree classifier by following the same approach, the classifier predicts the number of spam tweets equal to 6298 and non-spam tweets equal to 5376. The accuracy of the classifier obtained is 96%. The confusion matrix for semi-supervised

**Table 2** Confusion Matrix for Semi-supervised Support Vector Classifier

| **Actual Class** | | **Predicted Class** | | |
|---|---|---|---|---|
| | **Positive** | **Positive** | **Negative** | |
| | | TP=6194 (Correctly identified spam tweets) | FP=208 (Predicted spam but these are actually non-spam tweets) | Actual Spam= 6402 |
| | **Negative** | FN=129 (Predicted non-spam tweets but they were actually spam) | TN=5143 (Correctly identified non-spam tweets) | Actual non- spam= 5272 |
| | Predicted Spam= 6323 | Predicted non-spam=5351 | **Total=11,674** | |

**Table 3** Confusion Matrix for Semi-supervised Decision Tree Classifier

| Actual Class | | Predicted Class | | |
|---|---|---|---|---|
| | **Positive** | **Positive** | **Negative** | |
| | | TP=6142 (Correctly identified spam tweets) | FP=260 (Predicted spam but these are actually non- spam tweets) | Actual Spam=6402 |
| | **Negative** | FN=156 (Predicted non-spam tweets but they were actually spam) | TN=5116 (Correctly identified non-spam tweets) | Actual non-spam=5272 |
| | | Predicted Spam=6298 | Predicted non-spam=5376 | **Total=11,674** |

**Table 4** Confusion Matrix for K-Means Clustering

| Actual Class | | Predicted Class | | |
|---|---|---|---|---|
| | **Positive** | **Positive** | **Negative** | |
| | | TP=6036 (Correctly identified spam tweets) | FP=366 (Predicted spam but these are actually non- spam tweets) | Actual Spam=6402 |
| | **Negative** | FN=1382 (Predicted non-spam tweets but they were actually spam) | TN=3890 (Correctly identified non-spam tweets) | Actual non- spam=5272 |
| | | Predicted Spam=7418 | Predicted non-spam=4256 | **Total=11,674** |

SVC is shown in Table 3. Further, we applied the K Means clustering algorithm on the dataset. After clustering the class label associated with the dataset is compared with the label produced by the clustering algorithm. If both the labels are same then data is assumed to be correctly classified. Hence K Means Clustering predicted a total of 7418 spam tweets and 4256 non-spam tweets. The accuracy of the classifier obtained is 85% and details regarding predicted and actual true positive, false positive, true negative and false negative is shown in Table 4. The predictions of the three algorithms implemented are combined using a voting classifier so that we can predict the final label to the dataset. The classifier showed an accuracy of 99% while predicting 6462 tweets as spam and 5212 tweets as non-spam as shown in Fig. 3. The labels predicted by the voting classifier are assigned to the

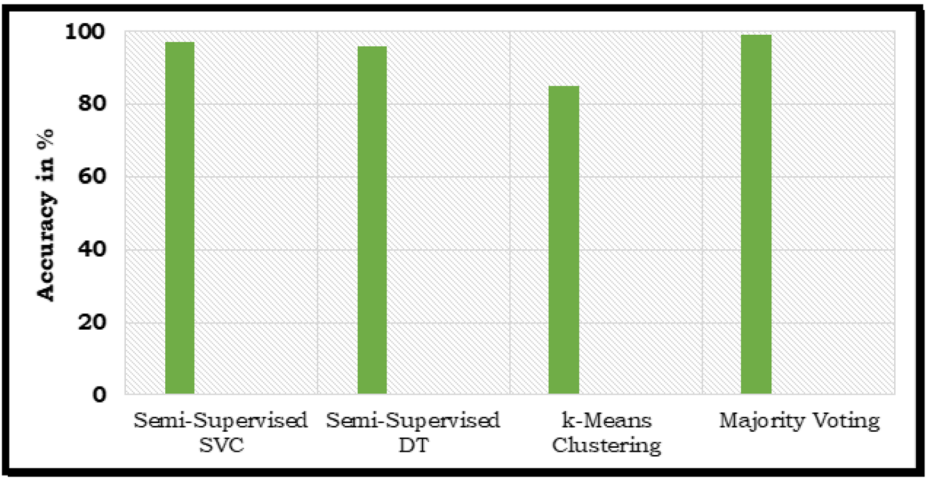


**Fig. 3** Comparison of accuracy of different classifiers

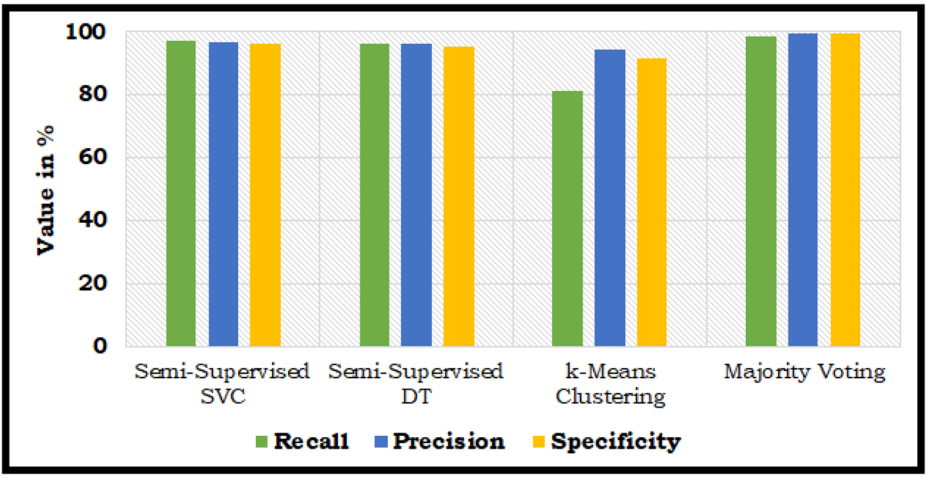


**Fig. 4** Comparison of performance measures

actual dataset as a true label and further processing is done on this labeled dataset. As appeared in Figs. 3 and 4 and Table 2, we inferred that the label assigned to tweets initially based on tweet user is about 99% true. This implies that spam users mostly post spam tweets and legitimate users mostly post legitimate tweets. Also, the rest of the spam tweets are those posted by legitimate users and non-spam tweets by spam users. The label predicted by the voting classifier is treated as an actual label for the tweets and further analysis is done on it.

In Table 5, the confusion matrix parameters corresponding to the voting classifier are enlisted which clearly reflect that the classifier has a low false-positive rate and low false-negative rate. The classifier has a low false-positive rate of 0.63% than the false-negative rate of 1.44%, respectively.

The performance measures appeared in Table 6 is reflective of the fact that the classifiers are accurate in its prediction.

The performance measures of the novel principled labeling technique is also evaluated using specificity, precision and recall. The high value of recall and precision as listed in Table 3 indicates that the Voting classifier predicts low false-negative and false-positive

**Table 5** Confusion Matrix of Voting classifier

| **Actual Class** | | **Predicted Class** | | |
|---|---|---|---|---|
| | **Positive** | **Positive** | **Negative** | |
| | | TP=6369 (Correctly identified spam tweets) | FP=33 (Predicted spam but these are actually non-spam tweets) | Actual Spam=6402 |
| | **Negative** | FN=93 (Predicted non- spam tweets but they were actually spam) | TN=5179 (Correctly identified non-spam tweets) | Actual non- spam=5272 |
| | | Predicted Spam=6462 | Predicted non-spam=5212 | **Total=11,674** |

**Table 6** Performance measures of the Semi-supervised Labelling

| Classifier | F1 Score (%) | Precision (%) | NPV (%) | Recall (%) | Specificity (%) | FPR (%) | FNR (%) |
|---|---|---|---|---|---|---|---|
| Semi - Supervised SVC | 97.34 | 96.75 | 97.55 | 97.95 | 96.11 | 3.88 | 2.04 |
| Semi - Supervised DT | 96.71 | 95.93 | 97.04 | 97.52 | 95.16 | 4.83 | 2.47 |
| K-Means | 87.34 | 94.28 | 73.78 | 81.36 | 91.40 | 8.59 | 18.63 |
| Majority Voting Scheme | 99.01 | 99.48 | 98.23 | 98.56 | 99.36 | 0.63 | 1.44 |

rates. The recall, precision, and specificity of the voting classifier showed an increase in values than the other three classifiers which clearly indicates the classifier is quite accurate in prediction.

## 7 Conclusion and future scope

The work in this paper is based on the labeling of Twitter datasets using semi-supervised labeling approach. We have effectively prepared a recent dataset of Twitter using streaming APIs on which the labeling was performed. The accuracy of all the classifiers used for semi-supervised labelling has a value more than 90%, which is itself reflecting the fact that models are quite exact in their prediction. This approach can be used for labelling datasets of different platforms. We can further use this recent dataset for different purposes (like finding the spam patterns in different tweet features for legitimate and spam users, perform sentiment analysis of tweets to predict the behavior of a tweet user, and so on) as we have extracted all the maximum possible features from the downloaded twitter data. In the future, this research work can be implemented using CUDA based programming to speed up the process of labelling in case of large datasets.

### Declarations

**Conflict of interest** Authors declared that they have no conflict of interest in this work.

## References

1. Abuliash M, Fazil M (2018) A hybrid approach for detecting automated spammers in twitter. IEEE Trans Inform Forensics Security 13(11):2707–2719
2. Al-Zoubi AM, Alqatawna J and Faris H (2017) Spam profile detection in social networks based on public features. 8th Int Conf Inform Comm Syst (ICICS), 130-135
3. Bazzaz Abkenar, S., Mahdipour, E., Jameii, S., & Haghi Kashani, M. (2021). A hybrid classification method for twitter spam detection based on differential evolution and random forest. Concurrency And Computation: Practice And Experience. https://doi.org/10.1002/cpe.6381.
4. Benevenuto F, Magno G, Rodrigues T and Almeida V (2010) Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference* (CEAS), 6:12–22.
5. Chakraborty A, Sundi J and Satapathy S (2012) SPAM: a framework for social profile abuse monitoring. *CSE508 report, Stony Brook University, stony brook, NY.*
6. Eshraqi N, Jalali M and Moattar MH (2015) Detecting Spam tweets in twitter using a data stream clustering algorithm. International Congress on Technology, *Communication and Knowledge* (ICTCK), 347–351

7. Gautam G and Yadav D (2014) Sentiment analysis of twitter data using machine learning approaches and semantic analysis. Seventh Int Conf Contemp Comp (IC3), 1-6.
8. Herzallah W, Faris H, Adwan O (2018) Feature engineering for detecting spammers on twitter: modelling and analysis. J Inf Sci 44(2):230–247
9. Inuwa-Dutse I, Liptrott M, Korkontzelos I (2018) Detection of spam-posting accounts on twitter. Neurocomputing 315:496–511
10. Lin PC and Huang PM (2013) A study of effective features for detecting long-surviving twitter spam accounts. 15th Int Conf Adv Comm Technol (ICACT), 841-846.
11. Liu C and Wang G (2016) Analysis and detection of Spam accounts in social networks. 2nd IEEE Int Conf Comp Comm, 2526-2530
12. Peikari M, Salsms S, Nofech-Mozes S, Martel A (2018) A cluster-then-label semi-supervised learning approach for pathology image classification. Sci Rep 8(1):1–13
13. Sedhai S, Sun A (2018) Semi-supervised spam detection in twitter stream. IEEE Trans Comp Soc Syst 5(1): 169–175
14. Stringhini G, Kruegel C and Vigna G (2010) Detecting spammers on social networks. Proceed 26th Ann Comp Sec Appl Conf (ACSAC), 1-9
15. Sun, N., Lin, G., Qiu, J., & Rimba, P. (2020). Near real-time twitter spam detection with machine learning techniques. Int J Comp Appl. 1-11. https://doi.org/10.1080/1206212x.2020.1751387



## Affiliations

**Tabassum Gull Jan[1] · Surinder Singh Khurana[1] · Munish Kumar[2]**

Tabassum Gull Jan
tabassumgull2012@gmail.com

Surinder Singh Khurana
surinder.seeker@gmail.com

[1] Department of Computer Science & Technology, Central University of Punjab, Bathinda, India

[2] Department of Computational Sciences, Maharaja Ranjit Singh Punjab Technical University, Bathinda, India