Indonesian Journal of Electrical Engineering and Computer Science
Vol. 24, No. 1, October 2021, pp. 554~563
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v24.i1.pp554-563

Identification of user's credibility on twitter social networks

Faraz Ahmad, S. A. M. Rizvi
Department of Computer Science, Jamia Millia Islamia University, New Delhi, India

Article Info

Article history:

Received Nov 2, 2020
Revised Aug 23, 2021
Accepted Aug 30, 2021

Keywords:

Credibility
Emotions
Machine learning
Sentiment
Twitter

ABSTRACT

Twitter is one of the most influential social media platforms, facilitates the spreading of information in the form of text, images, and videos. However, the credibility of posted content is still trailed by an interrogation mark. Introduction: In this paper, a model has been developed for finding the user's credibility based on the tweets which they had posted on Twitter social networks. The model consists of machine learning algorithms that assist not only in categorizing the tweets into credibility classes but also helps in finding user's credibility ratings on the social media platform. Methods and results: The dataset and associated features of 100,000 tweets were extracted and pre-processed. Furthermore, the credibility class labelling of tweets was performed using four different human annotators. The meaning cloud and natural language understanding platforms were used for calculating the polarity, sentiment, and emotions score. The K-means algorithm was applied for finding the clusters of tweets based on features set, whereas, random forest, support vector machine, naive Bayes, K-nearest-neighbours (KNN), J48 decision tree, and multilayer perceptron were used for classifying the tweets into credibility classes. A significant level of accuracy, precision, and recall was provided by all the classifiers for all the given credibility classes.

Corresponding Author:

Faraz Ahmad
Department of Computer Science
Jamia Millia Islamia University
Maulana Muhammad Ali Jauhar Marg, New Delhi, India
Email: faraz159020@st.jmi.ac.in

1. INTRODUCTION

Twitter is one of the most influential social networking platforms that gather millions of users across the world. It provides various features and functionality to their user which helps them in posting content in the form of text, an image, a meme, or a video related to any event/occasion. A tweet could be a user's own opinion or information that was shared from other sources, it could be a question, seeking knowledge for some live event or it is just comment without any relevance. Twitter considered being very useful in high-impact occasions [1], the occasion when everyone is fractious and desired to get correct information related to the ongoing events. Sometimes the uncredible or fake content also disseminates with the authentic ones and ends up creating a disastrous situation. So, it is extremely necessary to build a model that helps social networking users in differentiating between real or fake content.

In this study, machine learning (ML) models are developed based on features set for categorizing the real or fake content by incorporating the classification and clustering algorithms. The study consists of six major steps that helped in developing the model and to produce results that describe how efficiently the clustering and classification algorithms worked while categorizing the tweets. Twitter provides two different application programming interfaces (APIs) for crawling data, Twitter streaming API [2]-[6] and rest API [7]-[14], apart from these two APIs, [15], [16] collected tweets using "Wefollow" directory. In this paper, a data crawler has

been constructed using Twitter rest API for extracting the datasets based on trending hashtags associated with high-impact events/occasions or using Twitter handles of renowned personalities which were either directly or indirectly associated with the events. Along with the dataset a set of features were also extracted like username, number of followers, number of friends, the user is verified by Twitter, user's description, user's geolocation, and many more. The features crawled from Twitter which helped in categorizing the tweets were majorly classified into three categories;, user-based features [3], [15]-[17], content-based features [2], [8]-[12], [15], [18], [19], and hybrid features [1], [3], [4], [13], [14], [20]-[25].

The pre-processing step facilitates discarding the noisy or missing data. The data pre-processing involves the following steps. The tweet posted by the user who sign up on Twitter within one month time period and the profile's information was partially updated were discarded. The users having few numbers of followers, friends, and followings were also discarded. The tweets having less than ten words or contain only mentions @ or hashtags # were also discarded. Tweets written in some other language except English were also discarded.

In the third step, the sentiment, emotions, and polarity features were evaluated using the API provided by the International Business Machines (IBM) Watson Natural Language Understanding and meaning cloud. It helped in further enhancing the feature set associated with the data. The sentiment score was evaluated on a scale of -1 to +1. The emotions feature categorized into five categories like anger, disgust, sadness, fear, and joy. The tweet Polarity was calculated, and it provided results in six different categories. After evaluating these features set the data was undergone into the data annotation step.

A machine learning classification algorithm is supervised in nature, as it can learn by example. In a supervised machine learning approach, several input pairs are used for classifying the number of output pairs based on the mapping of input-output patterns. This explains annotated data are required for the development of the classification model. The data annotation process carried in this paper involved human intervention for labeling the tweets into one of the following given credibility classes which are unacceptable, somewhat unacceptable, neutral, somewhat acceptable, and acceptable.

The fifth step was to develop a machine learning classification and clustering model to categorize the tweets into given credibility classes. The multilayer perceptron (MLP), support vector machine (SVM), naive Bayes (NB), and random forest (RF) classification models were applied for classifying tweets into its respective credibility class [1], [3]-[5], [7]-[14], [21], [23], [26]. Whereas K-Mean algorithm was used for finding the clusters of identical tweets based on the features set. The total variance in the data that is explained by the K-Mean algorithm was 86.6%. However, the machine learning classification algorithms are giving high accuracy with an acceptable level of f1 score. The random forest algorithm gives the best results with 95.79% accuracy with 96.15% area under the curve followed by SVM, naive Bayes, and MLP.

Lastly, the user's credibility was evaluated using the content which he/she had been posted within the last month. The novelty of this work is the evaluation of the credibility of users by focusing on the content which they had posted, and the credibility of those content was calculated using emotions and sentiment-based features only which provide better results than traditional twitter-based features set models that were unable to find the exact credibility of reputed users who fulfil all the features and criteria to become an affluent user and knowingly or unknowingly indulged in posting unethical/fake content.

2. RESEARCH METHOD

2.1. Data preparation for machine learning

The data preparation step consists of data crawling, pre-processing, features generation, and data annotation. In this study, 100,000 tweets were crawled using Twitter rest API. The trending hashtags and the handles of political leaders, social activists, and high-profile celebrities were used for collecting tweets from Twitter social networks. The features set along with the individual user's profile provided by Twitter were also crawled. The features set is divided into two categories, content and user-based features which are shown in Table 1. These features also helped in preprocessing the extracted tweets.

The pre-processing step facilitates discarding the noisy or missing data. A set of rules was prepared by taking help from the previous state of artworks that have been done in this area. The tweets having length fewer than 10 words or tweets posted by those users who sign up on Twitter for less than 1 month were discarded. Tweets whose user's information like profile photo, bio, user description was missing from the profile, were also discarded. The tweets containing only hashtags mentioned or emoticons were also discarded. Tweets written in some other language except English were also removed. Furthermore, tweets were pre-processed based on extracted features set (user's followers count, tweet favorite count, user-created at, tweets created at, user's description is present or not and tweet is possibly sensitive) provided by Twitter. After preprocessing the sentiment, emotions, and polarity features were evaluated using the natural language understanding and meaning cloud API's which are provided by IBM. It helped in further enhancing the feature set associated with

the data. The sentiment score was evaluated at a scale of -1 to +1, where 0 to -1 informed about the negative sentiment score and 0 to +1 informed about the positive sentiment score. The emotions feature categorized into 5 groups like anger, disgust, sadness, fear, and joy. The emotions score was evaluated on a scale of 0 to +1, closer to the value to +1 greater will be the emotion. The polarity features evaluate the tweet in one of the following categories like P+, P, none, neutral, N, N+. anger, disgust, sadness, fear, and joy.

Table 1. Twitter features set and its description

Content Features	Description	User Features	Description
CreatedAt	time at which tweet was posted	UserVerified	user verified by Twitter
Retweet	retweet count on tweet	UserLocation	location while sending
Source	source used for posting the tweet	UserName	user's name
FavoriteCount	number of time tweet liked by others	UserScreenName	user's screen name
TweetGeoLocation	geographic location is on or not	UserFavouritesCount	total number of tweets liked by user
LanguageOfTweet	language in which tweet is written	UserFollowersCount	number of followers
HashtagEntities(#)	number of hashtags within the tweet	UserFriendsCount	number of users the user follows
MediaEntities	is media entities attached with the tweet	UserCreatedAt	time at which user signup on Twitter
UserMentionEntities(@)	number of mentions @ within the tweet	UserStatusesCount	total number of status posted by the user
PossiblySensitive	if tweet contains sensitive words	UserGeoEnabled	user's home location
Retweeted	if the tweet is a retweet	UserTimeZone	user's home time zone
IsProtectedTweet	tweet will not be visible to anyone	UserDescription	user's own description

Furthermore, the number of negative and positive words associated with the tweets was evaluated using the nitroxide radical coupling (NRC) emotion lexicon. All the evaluated emotions and sentiment features are shown in Table 2. The next step is data annotation which requires human effort, a person who was active on Twitter and has a strong knowledge related to ongoing events and trending hashtags were assigned for annotating the data. This step requires a rigorous and time-consuming endeavor. The data was outsourced to four human annotators for labelling the tweets in one of the five given categories for classifying the tweets into credible classes. All the annotators are pursuing a Ph.D. in different fields of sciences and social sciences. The credibility score was given on a likert scale such as, "Acceptable", "Slightly Acceptable" and "Neutral", "Slightly Uncredible", and "Uncredible". The uncredible are those tweet contents that comprise abusive, unethical words related to any person, religion, caste, culture, or society. The validity of credibility class annotation depends upon the "Wisdom of the Crowd". Only the majority votes were considered for every tweet annotation, the rest of the tweets were discarded.

Table 2. Sentiment and emotions features

Features	Description
Sentiment Score	Sentiment score associated with tweet. Score ranges from -1 to +1
Emotion (Joy)	Emotion Joy score associated with tweet. Score ranges from 0 to 1
Emotion (Anger)	Emotion Anger score associated with tweet. Score ranges from 0 to 1
Emotion (Disgust)	Emotion Disgust score associated with tweet. Score ranges from 0 to 1
Emotion (Fear)	Emotion Fear score associated with tweet. Score ranges from 0 to 1
Emotion (Sadness)	Emotion Sadness score associated with tweet. Score ranges from 0 to 1
Polarity	Polarity scale consists of P+, P, None, Neutral N, N+ emotions
Positive	Number of positive words in a tweet.
Negative	Number of negative words in a tweet.

2.2. Social network analysis of tweets and graph generation

The social network analysis of tweets Wasserman and Faust [27] using the provided dataset, established several measures like clustering, density, centralization, modularity, and proportion of isolation. The density depends on the ratio of the number of available links to the total possible links if all the nodes are highly interconnected then cluster density will be higher. Whereas, if nodes are loosely connected, it means the density of the cluster is relatively low. A cluster defines a community of people who are posting tweets

on similar topics, however, the number of clusters depends upon the total number of topics from which tweets are extracted from Twitter. The social network graph is shown in Figure 1 which was generated by using the keywords from the dataset.

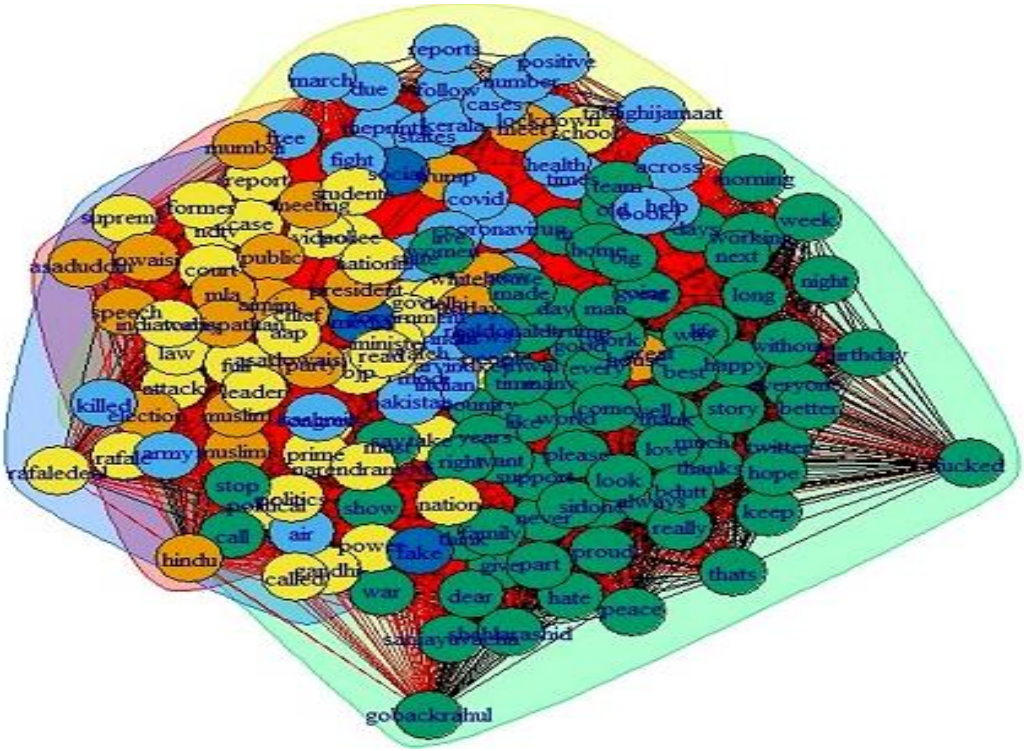


Figure 1. Social network analysis of tweets using label propagation algorithm

This graph has been made using the label propagation algorithm which is semi-supervised and used to find community structure within the provided dataset. This algorithm allocates labels to formally unallocated data points. This algorithm requires no prior information about the parameters beforehand and it runs faster than other community detection algorithms.

In this graph, many overlapping clusters have been found which signifies the different communities used the same words for showing their thoughts and emotions. As most of the tweets were crawled using the handle@ of political leaders, social activists, and some trending hashtags#. The keywords appeared in this graph illustrates that the extracted tweets were taken from some high impacted events/occasion happened in India such as “Rafael Deal”, “Covid19”, “Health Care”, “Nizamuddin Markaz (Muslims)”, “Student Protest March against CAA/NRC”, “Pakistan”, “War”, “Army”, and “Law”. The negative sentiment words such as “Hate”, “Fucked”, and “Fake”. also appeared several times within the dataset.

3. RESULTS AND DISCUSSION

3.1. Machine learning model generation

The K-Mean clustering algorithm was used for finding the clusters of similar objects based on the given feature set. The contingency table produced by the K-Mean algorithm is shown in Table 3. The algorithm provides “within-cluster sum of squares by cluster”: 11794.72, 22243.53, 11490.04, 19315.90, and 22209.82. Whereas the total variance explained by the data is 75.3 % which was calculated using the formula (between_SS / total_SS = 75.3 %). The table clearly explained that the cluster 2, 3 and 4 mostly contained tweets of “Acceptable”, “Slightly Acceptable” and “Neutral” credibility class, whereas cluster 1 and cluster 5 contains tweets of “Slightly Uncredible” and “Uncredible” credibility classes. Moreover, multilayer perceptron (MLP), support vector machine (SVM), J48 decision tree, naive Bayes (NB), K-nearest-neighbours KNN, and random forest (RF) are the six major classification algorithm which was applied for categorizing the data into five different credibility classes. The overall statistics provided by the classifiers are shown in Table 4.

Table 3. Contingency table (K-Mean clustering algorithm)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Acceptable	0	3995	8	5972	2
Neutral	718	2246	14794	0	2
Slightly Acceptable	11	5483	1	11734	1
Slightly Unacceptable	3400	593	987	884	9202
Unacceptable	2298	5	20	9	8050

Table 4. Overall statistics by the classifiers

ML Classifiers/ Statistics	RF	NB	SVM	MLP	J48	KNN
Accuracy	0.9657	0.9539	0.9565	0.9503	0.9654	0.9511
95% CI	(0.9631, 0.9681)	(0.951, 0.9567)	(0.9537, 0.9592)	(0.9479, 0.9529)	(0.9628, 0.9678)	(0.9481, 0.9539)
No Information Rate	0.2514	0.2512	0.2522	0.2498	0.2545	0.2542
P-Value	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16
Kappa	0.9565	0.9415	0.9451	0.9401	0.9561	0.9379

3.2. Performance evaluation

The accuracy, recall, precision, and f1 score were calculated for evaluating the performance of all six classifiers. All these measures were calculated by using the contingency table, which was generated by all the classification models. The contingency tables consist of four different kinds of values, “true positive (TP)”, “true negative (TN)”, “false positive (FP)” and “false negative (FN)”. The TP and FP are the values that were classified correctly, however, the FP and FN are the values that were verified incorrectly [28].

The accuracy of the classifier is the ratio of (TP+FP)/ (TP+FP+TN+FN), whereas, precision is the ratio of TP/(TP+FP) and recall is the ratio of TP/(TP+FN). The F1 score is the weighted average of recall and precision, which is calculated by using the formula 2*(recall * precision)/(recall + precision). The performance evaluation of classifiers based on all credibility classes is shown in Table 5.

Cross-validation is the process of evaluating the efficiency of the machine learning model on different data set while performing the prediction. The machine learning model will be trained on a known data set and tested on an unknown dataset for generating the prediction accuracy. However, for acknowledging the performance of the model on a new dataset and to flag a problem like overfitting the cross-validation should be performed. The K fold cross-validation process involves the partitioning of the data into K subgroups, perform training on subgroups, and testing on another. Besides, for reducing the variability several rounds of training and testing should be performed and lastly, the averaged results are taken for estimating the model’s performance. Table 6 shows the results of 10 fold cross-validation and AUROC.

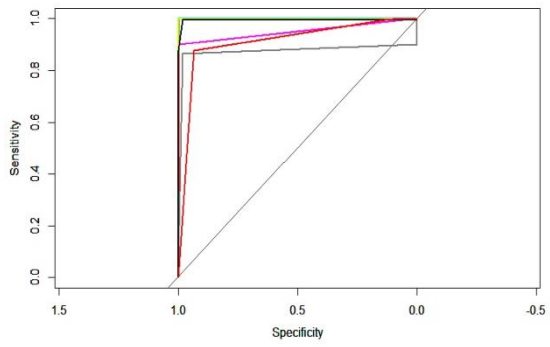
Moreover, for evaluating the performance of the classifier a multiclass receiver operating characteristics (ROC) curve was generated. The ROC shown in Figure 2 is also known as the probability curve which is drawn between ture positive rate v/s false positive rate at different thresholds. The area under the curve (AUC) is the performance measure of ROC. It tells about the degrees of separability between the classes. A higher value of AUC signifies the better prediction performance of the classification model.

Table 5. Evaluating performance of classifiers based on credibility classes

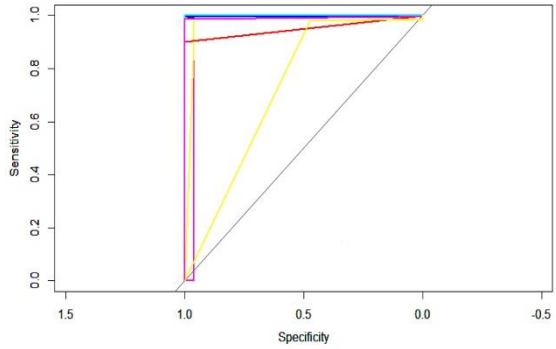
		Acceptable	Slightly Acceptable	Neutral	Slightly Unacceptable	Unacceptable
RF	Precision	0.9962	0.9700	0.9998	0.8895	0.9883
	Recall	0.9990	0.9856	0.9173	0.9572	0.9987
	F1 Score	0.9976	0.9777	0.9568	0.9221	0.9935
NB	Precision	0.9934	0.9900	0.9905	0.8446	0.9520
	Recall	0.9997	0.9467	0.9229	0.9391	0.9993
	F1 Score	0.9965	0.9679	0.9555	0.8893	0.9751
SVM	Precision	0.9973	0.9617	0.9994	0.9071	0.9872
	Recall	0.9993	0.9963	0.9250	0.9460	0.9997
	F1 Score	0.9983	0.9787	0.9608	0.9261	0.9934
MLP	Precision	0.9967	0.9969	0.9597	0.8922	0.8971
	Recall	0.9980	0.9227	0.9930	0.8856	0.9801
	F1 Score	0.9973	0.9584	0.9760	0.8889	0.9368
KNN	Precision	0.9897	0.9699	0.9717	0.8789	0.9505
	Recall	0.9806	0.9759	0.9156	0.9131	0.9980
	F1	0.9852	0.9729	0.9428	0.8956	0.9737
J48	Precision	0.9960	0.9624	0.9151	0.9454	1.0000
	Recall	1.0000	0.9981	1.0000	0.8936	0.9980
	F1	0.9980	0.9799	0.9557	0.9188	0.9990

Table 6. 10-fold cross-validation (accuracy) and AUROC values

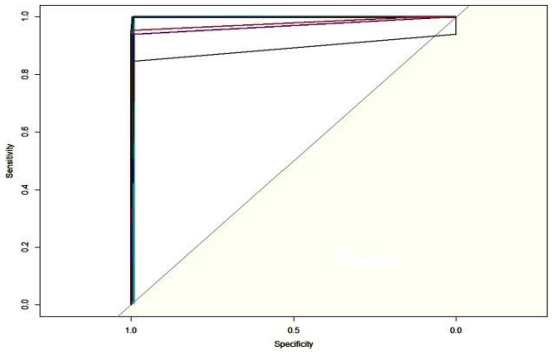
ML Classifiers	Accuracy	AUROC
RF	0.9602789	0.982023
NB	0.9657625	0.980417
SVM	0.9592882	0.972829
MLP	0.9435220	0.972199
J48	0.9669350	0.987609
KNN	0.9672703	0.980897



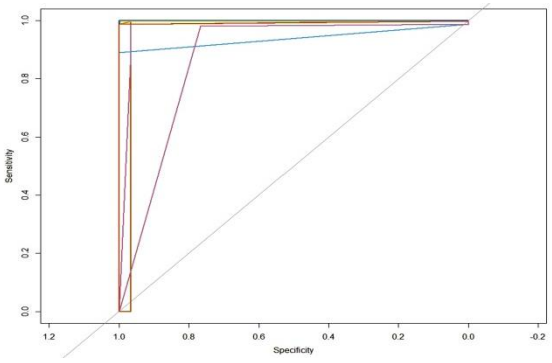
ROC curve for MLP



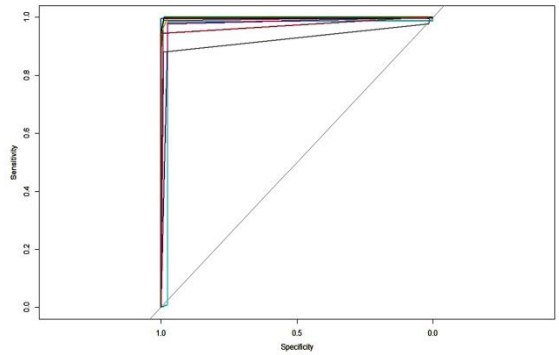
ROC curve for SVM



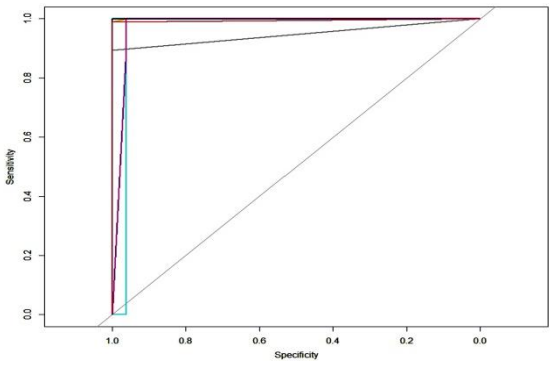
ROC curve for naive Bayes



ROC curve for random forest



ROC curve for KNN



ROC curve for J48

Figure 2. ROC curve

Furthermore, for analysing the features set (“Sentiment”, “Polarity”, “Anger”, “Fear”, “Disgust”, “Sadness”, “Joy”, “Positive” and “Negative”) against the response variable (“Acceptable”, “Slightly Acceptable”, “Neutral”, “Slightly Unacceptable”, “ Unacceptable”) a box plot analysis was performed which is shown in Figure 3. The distribution and variability of the dataset were analysed by making the box plots. It is a graph which is based on five data points, the first one was “Minimum”, then “First Quartile”, “Median”, “Third Quartile” and “Maximum”. It provides information related to the spread out of all the data points. It has been found that most of the negative sentiment tweets were belong to the Unacceptable and Slightly Unacceptable classes, however, the positive sentiment tweets belong to the acceptable, slightly acceptable,

and neutral classes. Unacceptable and slightly unacceptable classes have a higher proportion of negative emotions such as anger, disgust, fear, sadness, whereas the acceptable, slightly acceptable, and neutral classes have higher positive emotion words such as joy. Lastly, the number of positive words and the number of negative word features were plotted, and graphs showed that positive words appeared more in the Acceptable class as compared to the unacceptable class.

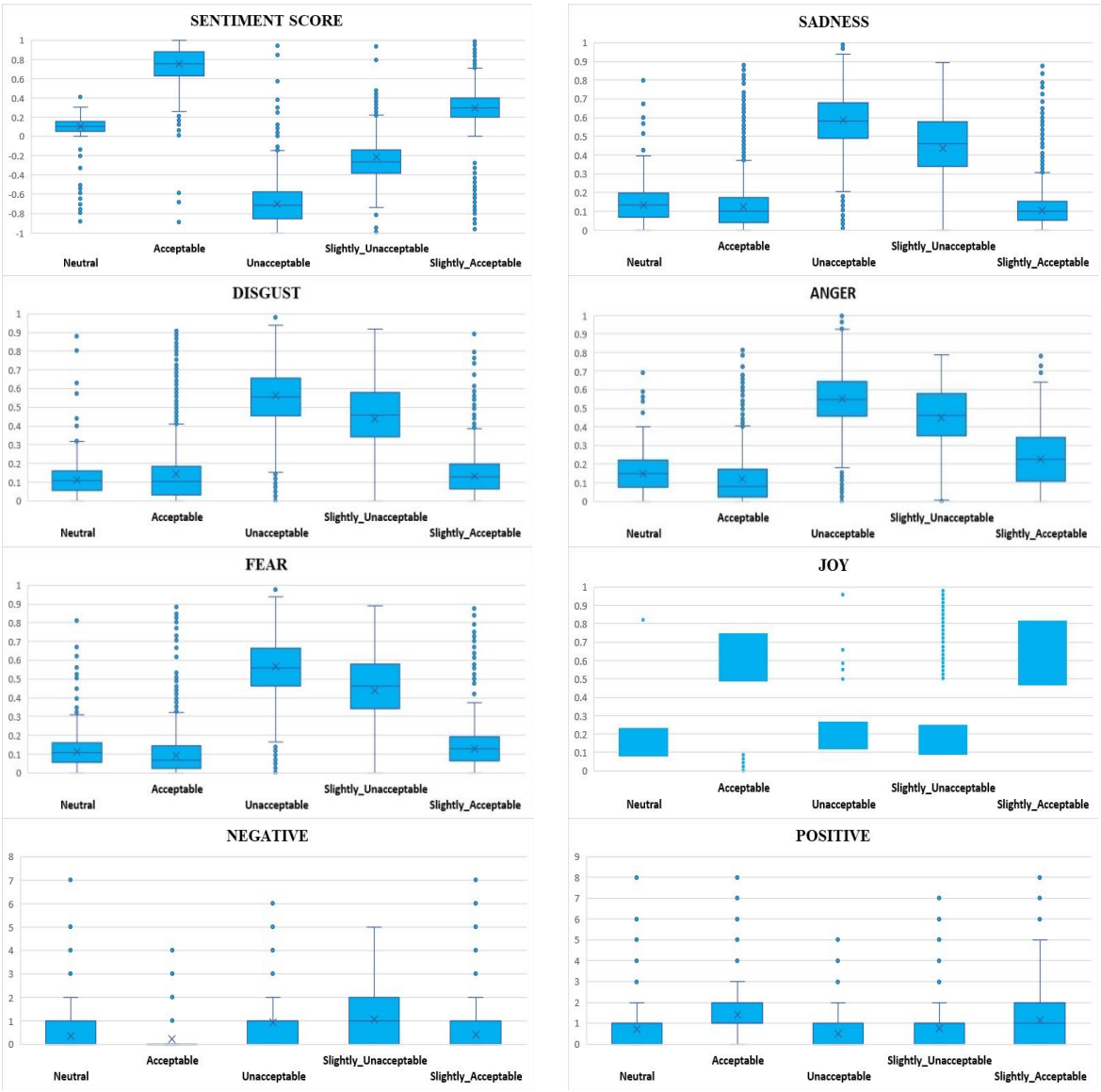


Figure 3. Box plot analysis

3.3. Evaluating user's credibility score

The credibility of the user is defined by the quality of content which he/she is posted, and how it impacted society. In this research, a model has been developed for finding the credibility of posted tweets and based on the above-mentioned features set and machine learning outcome the user's credibility is evaluated. Moreover, the tweet posted by any random user using certain trending hashtags# was also taken into consideration for evaluating the credibility score. The dataset of 10 high profile users and 10 trending hashtags was extracted from Twitter, based on the percentage of uncredible and credible content posted on Twitter the credibility score was calculated. The results shown in Table 7 include the total number of tweets extracted from the user handle or trending hashtags with the percentage of tweets categories into acceptable, slightly acceptable, slightly unacceptable, and unacceptable credibility classes rating which was labelled by

the human annotators. And based on these ratings the user's credibility score was assigned. The tweet contained any hate speeches, abusive, or unethical words are considered unacceptable, and users who indulged in posting such content will be considered Uncredible users. If more than 1% of the total extracted data falls under the category of the unacceptable tweet, then the user who posted those contents will be considered as an uncredible user. However, if more than 10% of the total extracted data falls under the category of the slightly unacceptable tweet, then that user will be considered as a doubtful user. The slightly unacceptable tweets are mostly not harmful to the society as it does not contain any abusive word, however, it cannot be considered as credible as it was user's own opinion for any other person or event/occasion. These kinds of tweets provide no benefit to society and contain words like "Idiot", "Uneducated", and "Stupid". These tweets are often used by people for mocking government policies as per our dataset.

Credible users are those who spread genuine news and share information that is beneficial for society. They never used any negative emotion word for any individual, caste, or religion. They mostly tried to work in favor of society, especially minorities and marginalized people. The result showed that most of the tweets extracted using the user's handle fall into the category of acceptable credibility class and hence their users should be classified as credible users. This is because the data was collected from high-profile Indian leaders, movie actors/actresses, social activists, comedians, and sports players which was already verified by Twitter. However, the tweets extracted using above mentioned hashtags were written by the public for poking other people of the society, making fun of their religion, for playing blame games. These tweets contain abusive and absurd words that are not acceptable and need to be eliminated from any reputed social networking such as Twitter.

Table 7. User's and hashtag's credibility rating

Users	Total extracted tweets	% of Acceptable tweets	% of Un-acceptable tweets	% of Slightly-Acceptable tweets	% of Slightly-Unacceptable tweets	Credibility Rating
User 1	2500	26.44%	0.12%	65.12%	8.32%	Credible
User 2	3000	36.63%	0.03%	53.23%	10.10%	Doubtful
User 3	4000	37.23%	0.70%	45.23%	16.85%	Doubtful
User 4	3000	35.97%	1.43%	52.30%	10.30%	Uncredible
User 5	4500	48.71%	0.24%	41.82%	9.22%	Credible
User 6	4000	41.78%	0.35%	52.53%	5.35%	Credible
User 7	3000	29.63%	0.40%	63.03%	6.93%	Credible
User 8	3000	24.60%	0.47%	54.97%	19.97%	Doubtful
User 9	2500	54.04%	0.08%	42.44%	3.44%	Credible
User 10	2250	38.72%	0.44%	46.88%	13.96%	Credible
Hashtags						
#CAA/#NRC	4600	23.78%	0.96%	51.78%	23.48%	Doubtful
#PKMKB	5000	4.68%	59.72%	9.58%	26.02%	Uncredible
#WhyTheyHateModi	3000	4.03%	73.70%	10.17%	12.10%	Uncredible
#PulwamaAttack	2000	12.30%	6.55%	38.20%	42.95%	Uncredible
#Covid19	6000	29.98%	0.97%	30.88%	38.17%	Doubtful
#TablighiJamaat	5000	2.32%	41.38%	10.50%	45.80%	Uncredible
#GodiMedia	2400	12.46%	54.13%	10.04%	23.38%	Uncredible
#GST	2000	33.85%	0.85%	23.55%	41.75%	Doubtful
#SurgicalStrike	2000	23.85%	10.85%	18.55%	46.75%	Uncredible
#ShameOnYouNewsNation	2500	15.88%	25.04%	19.84%	39.24%	Uncredible

4. CONCLUSION

Twitter is one of the most effective platforms for sharing information such as text, memes, and videos. with millions of users across the world. Such information sharing delivers huge benefits to society by providing the updation about every possible event that happened across the world. However, sometimes the fake/ rumored information was also disseminated with the real one and it becomes quite essential to filter out the uncredible information before it can create chaos within the society. And sometimes the chaos happened is so vast that it can end up in disastrous condition. In this paper, a model has been developed for finding the user's credibility based on the tweets which they had posted on Twitter social networks. In total 100,000 tweets from twenty different users handles and hashtags along with their associated features were crawled. The data was annotated using human annotators into five given credibility classes. The emotions, sentiment, and polarity features that are associated with the tweets were evaluated using natural language understanding (NLU) and meaning cloud API's provided by IBM. All of these features helped in categorizing the tweets into the given credibility classes. The K-Means algorithm was applied to make clusters and to find in which cluster tweet will fall based on the features set, whereas, random forest, SVM, naive Bayes, J48 decision tree, KNN, and multilayer perceptron were applied for classifying the tweets into the labelled credibility classes.

The result shows the significant level of accuracy, precision, and recall was provided by all the classifiers, the best results were given by random forest with 96.59% of accuracy, followed by a J48 decision tree with 96.54%, SVM with 95.65%, naïve Bayes with 95.39%, KNN with 95.11% and MLP with 95.03% accuracy respectively. The last step comprises the evaluation of user's credibility based on the tweets which they had posted on Twitter Social Networks within one month. The novelty of our model is that it works efficiently irrespective of the position or reputation that the user holds within the society. The emotions and sentiment-based features help in filtering the uncredible and rumored content to a great extent and it has been found through experiments that the importance of these novel features gives the leading edge in the research for evaluating the credibility of the posted content as it is not affected by the heuristics of other users. For future references, the work on evaluating the associated emotions and sentiments in posted images will be taken into consideration. Moreover, the authors will try to find the trolling content in OSNs.

REFERENCES

[1] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "TweetCred: Real-Time Credibility Assessment of Content on Twitter," *In: Aiello L.M., International conference on social informatics. Springer, Cham*, vol. 8851, 2014, pp. 228-243, doi: 10.1007/978-3-319-13734-6_16.

[2] T. Takahashi and N. Igata, "Rumor detection on twitter," *In The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems, IEEE*, 2012, pp. 452-457, doi: 10.1109/SCIS-ISIS.2012.6505254.

[3] M. Alrubaian, M. Al-Qurishi, M. M. Hassan, and A. Alamri, "A credibility analysis system for assessing information on twitter," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 661-674, 2016, doi: 10.1109/TDSC.2016.2602338.

[4] M. Alrubaian, M. Al-Qurishi, M. A. Rakhami, M. M. Hassan, and A. Alamri, "Reputation based credibility analysis of Twitter social network users," *Concurrency and Computation: Practice and Experience*, vol. 29, no. 7, p. e3873, 2017, doi: 10.1002/cpe.3873.

[5] K. Lorek, J. Suehiro-Wiciński, M. Jankowski-Lorek, and A. Gupta, "Automated Credibility Assessment on Twitter," *Computer Science*, 16, no. 2, pp. 157-168, 2015, doi: 10.7494/csci.2015.16.2.157.

[6] T. Mitra and E. Gilbert, "Credbank: A Large-Scale Social Media Corpus with Associated Credibility Annotations," *In ICWSM*, 2015, pp. 258-267.

[7] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," *In 13th International Conference on Data Mining*, IEEE, 2013, pp. 1103-1108, doi: 10.1109/ICDM.2013.61.

[8] G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose, "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus," *In Proceedings of the 21st ACM international conference on Information and knowledge management*, 2012, pp. 1980-1984, doi: 10.1145/2396761.2398556.

[9] S. Verma, S. Vieweg, W. Corvey, L. Palen, J. Martin, M. Palmer, and K. Anderson, "Natural language processing to the rescue? extracting" situational awareness" tweets during mass emergency," *In Fifth international AAAI conference on weblogs and social media*, vol. 5, no. 1, July 2011.

[10] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting Personality from Twitter," 2011 *IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 2011, pp. 149-156, doi: 10.1109/PASSAT/SocialCom.2011.33.

[11] H. K. Thakur, A. Gupta, A. Bhardwaj, and D.Verma, "Rumor Detection on Twitter Using a Supervised Machine Learning Framework," *International Journal of Information Retrieval Research (IJIRR)*, vol. 8, no. 3, pp. 1-13, 2018, doi: 10.4018/IJIRR.2018070101.

[12] R. Sicilia, S. L. Giudice, Y. Pei, M. Pechenizkiy, and P. Soda, "Twitter rumour detection in the health domain," *Expert Systems with Applications*, vol. 110, pp. 33-40, 2018, doi: 10.1016/j.eswa.2018.05.019.

[13] J. O. Donovan, B. Kang, G. Meyer, T. Höllerer, and S. Adalii, "Credibility in Context: An Analysis of Feature Distributions in Twitter," 2012 *International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 2012, pp. 293-301, doi: 10.1109/SocialCom-PASSAT.2012.128.

[14] F. Yang, Y. Liu, X. Yu, and M. Yang, "Automatic detection of rumor on Sina Weibo," *In Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, ACM, 2012, p. 13, doi: 10.1145/2350190.2350203.

[15] K. R. Canini, B. Suh, and P. L. Pirolli, "Finding Credible Information Sources in Social Networks Based on Content and Social Structure," 2011 *IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 2011, pp. 1-8, doi: 10.1109/PASSAT/SocialCom.2011.91.

[16] Q. V. Liao, C. Wagner, P. Pirolli, and W. T. Fu, "Understanding experts' and novices' expertise judgment of twitter users," *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, May 2012, pp. 2461-2464, doi: 10.1145/2207676.2208410.

[17] D. Westerman, P. R. Spence, and B. V. Der Heide, "A social network as information: The effect of system generated reports of connectedness on credibility on Twitter," *Computers in Human Behavior*, vol. 28, no. 1, pp. 199-206, 2012, doi: 10.1016/j.chb.2011.09.001.

[18] Y. Bao, C. Yi, Y. Xue, and Y. Dong, "A new rumor propagation model and control strategy on social networks," 2013 *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, 2013, pp. 1472-1473, doi: 10.1109/ASONAM.2013.6785909.

[19] J. Schwarz and M. Morris, "Augmenting web pages and search results to support credibility assessment," *In Proceedings of the SIGCHI conference on human factors in computing systems*, May 2011, pp. 1245-1254, doi: 10.1145/1978942.1979127.
[20] X. Xia, X. Yang, C. Wu, S. Li, and L. Bao, "Information Credibility on Twitter in Emergency Situation," *Lecture Notes in Computer Science*, vol. 7299, Springer, Berlin, Heidelberg, pp. 45-59, 2012, doi: 10.1007/978-3-642-30428-6_4.
[21] Q. Zhang, S. Zhang, J. Dong, J. Xiong, and X. Cheng, "Automatic Detection of Rumor on Social Network," *In Natural Language Processing and Chinese Computing. Lecture Notes in Computer Science*, Springer, Cham, vol. 9362, pp. 113-122, 2015, doi: 10.1007/978-3-319-25207-0_10.
[22] X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah, "Real-time rumor debunking on twitter," *In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, ACM, 2015, pp. 1867-1870, doi: 10.1145/2806416.2806651.
[23] C. Castillo, M. Mendoza, and B. Poblete, "Information Credibility on Twitter," *In Proceedings of the 20th International Conference on World Wide Web*, ACM 2011, pp. 675-684, doi: 10.1145/1963405.1963500.
[24] M. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, "Tweeting is Believing?: Understanding Microblog Credibility Perceptions," *In Proceedings of the ACM Conference on Computer Supported Cooperative Work*, ACM, 2012, pp. 441-450, doi: 10.1145/2145204.2145274.
[25] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Measuring information credibility in social media using combination of user profile and message content dimensions," *International Journal of Electrical & Computer Engineering*, vol. 10, no. 4, pp. 2088-8708, 2020, doi: 10.11591/ijece.v10i4.pp3537-3549.
[26] F. Ahmad and S A M. Rizvi, "Emotion Based Content Credibility Prediction Model For Twitter Social Networks," *International Journal of Scientific and Technology Research*, vol. 9, no. 3, pp. 1253-1259, 2020.
[27] S. Wasserman, and K. Faust, "Social network analysis: Methods and applications," *Cambridge university press*, vol. 8, 1994, doi: 10.1017/CBO9780511815478.
[28] H. R. Esmaeel, "Analysis of classification learning algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 17, no. 2, pp. 1029-1039, 2020, doi: 10.11591/ijeecs.v17.i2.pp1029-1039.

BIOGRAPHIES OF AUTHORS

Dr. Faraz Ahmad has completed Ph.D. in Computer Science from the Department of Computer Science, Jamia Millia Islamia University, New Delhi, India. He has worked as an external faculty in different departments of JMI and Delhi University for the last seven years. His area of interest is Data Mining, Online Social Networks, Machine Learning, Statistical Analysis, and Artificial Intelligence. Currently, he is working as a developer in National Informatics Center, Delhi Secretariat Office.

Dr. S. A. M. Rizvi is working as a Professor in the Department of Computer Science, Jamia Millia Islamia, having more than 30 years of experience in teaching and research. He did his Ph.D. from Dr. R. M. L. Avadh University, India, in 1996. He designed various programs/courses as a Chairman/Member of BOS, Academic Council, and other academic bodies at universities/Higher Educational Institutions (HEI). He Taught in the USA (Credit Hour System), Australia, UAE, and in Indian Educational Systems. He has more than 150 Publications, covering a vast array of topics in Computer Science and Applications. More than 18 Ph.D. awardees till date, with 8 currently registered scholars. He is also a Senior Member of the Computer Society of India (CSI), Old Member of IEEE, ISCA, and IEA.