International Journal of System Dynamics Applications
Volume 11 • Issue 2 • July-December 2022

User Activity Classification and Domain-Wise Ranking Through Social Interactions

Ravindra Kumar Singh, National Institute of Technology, Jalandhar, India
https://orcid.org/0000-0003-1142-1954

Harsh Kumar Verma, National Institute of Technology, Jalandhar, India

ABSTRACT

Twitter has gained a significant prevalence among users across numerous domains in the majority of the countries and among different age groups. It servers a real-time micro-blogging service for communication and opinion sharing. Twitter is sharing its data for research and study purposes by exposing open APIs that make it the most suitable source of data for social media analytics. Applying data mining and machine learning techniques on tweets is gaining more interest. The most prominent enigma in social media analytics is to automatically identify and rank influencers. This research is aimed to detect the user's topics of interest in social media and rank them based on specific topics, domains, etc. A few hybrid parameters are also distinguished in this research based on the post's content, post's metadata, user's profile, and user's network feature to capture different aspects of being influential and used in the ranking algorithm. Results concluded that the proposed approach is effective in both the classification and ranking of individuals in a cluster.

KEYWORDS

Apache Kafka, Apache Spark, Elasticsearch, Real-Time Analytics, Social Media Analytics, User Classifications, User Ranking

1. INTRODUCTION

Social media has evolved very dynamically and has become a primary source of feedback, trends, debate, and sentiments across various domains (Singh et al., 2020; Grover et al., 2019). The various social media platforms have become an important channel in collaboration and sharing opinions, thoughts, and experiences (Jansen et al., 2009) due to its publicly broadcasting and uncontrolled interactions on the posts. Messages posted on twitter could be viewed by everyone. Most of the users are interested in multiple domains and sharing their opinion on them. On the contradictory side, few users are very much inclined towards a specific domain and share their viewpoints related to that domain, they generally have a huge follower base and their posts are serving as food for thought or point of discussions among followers and the whole network. Such users are known as influencers. This social impact could be observed in all the business segments, online advertising and promotions have been considered as an important aspect to maintain a good brand reputation on social media platforms (Vernier et al., 2018). Business leaders are looking for different ways to effectively promote their offerings and be competent in social media networks. In such cases, Influencers play a crucial

DOI: 10.4018/IJSDA.20220701.oa5

role to help the companies to reach the correct audience/customers. In this scenario, one of the biggest challenges is to find a way to determine the most influential user on the social media network.

Indeed, it's not possible to have a context-independent universal influencer, so automatically discovering the influencer in a specific domain, location or any other stringent criteria is an add on challenge (Montangero et al., 2015). Most of the approaches to solve this problem are simply magnitude-based and taking the absolute count of the concerned feature to rank, that is not an effective approach because these are not highly reliable. A high number of followers or friends don't guarantee the higher engagement on the posts, similarly, user-provided information too could not be reliable, sometimes users fake it. As an example, the user's short bio section is not reflecting the core interest of the user even sometimes it is misguiding the facts.

So this research is mainly focused to classify the user's social media activities and rank them on various domains, particularly politics, sports, cinema, business, technology, and others. Additionally, this research has distinguished a few hybrid parameters based on the post's content, post's metadata, user's profile, and user's network feature to capture different aspects of the influencer's ranking.

The rest of the article is arranged as follows. Section 2 presents a brief overview of related work regarding user activity classifications in social media and ranking influencers of specific topics. Section 3 presents an overview of the research design and methodology. Section 4 describes the experimental setup parts of this research by describing the datasets, methods of preprocessing, selections of machine learning models, and their evaluation parameters. Section 5 focuses on the experimentation part along with the analysis and discussion of the results. Finally, section 6 concludes the research and also presents recommendations for further work.

2. BACKGROUND

Twitter is not only utilized for communication and opinion sharing but also considered as a source of recommendation systems and promotion activities. In such cases, finding the users interested in the concerned field or domain is very evident, and boosting the effectiveness of the recommendations and promotions. User's categorization based on their interests will help this cause and limit the targeted users from the huge user base of Twitter. Predicting the behavior and interests of web users is an evolving area of research, and it is a very challenging task to reliably classify users among various categories (Rahman et al., 2019). User behavior classification, profiling, modeling, and prediction for various use cases in different domains, such as commerce, banking, trend analysis, education, medicine, etc., are the hottest area of research in the data analytics field (Sawita Yousukkee, 2016).

The research highlighted the user's interest from Twitter trends using a modeling approach to tweets (Shahzad et al., 2017). They collected the tweets of particular trends and label it with an appropriate category and utilized it for model training after preprocessing of the tweets and later on these models are predicting the categories of the given tweets. This approach is using the Support Vector Machine (SVM) for the predictions and classifies the topics of interest of a user on Twitter. This approach could be utilized in information filtering and prediction systems, especially in personalized recommendation systems, etc. Another research has designed a real-time system for Twitter user profiling based on a supervised machine learning approach to categorize Twitter users into various interest categories like Politics, Entertainment, Entrepreneurship, Journalism, Science & Technology, and Healthcare (Raghuram et al., 2016) based on Tweet-based, User-based and Time-series based features. They utilized numerous classifiers like Support Vector Machines, Naive-Bayes, k-Nearest Neighbours, Decision Tree, and Logistic Regression, and obtained up to 89.82% accuracy in classification. A characteristics analysis of online user's behavior concludes that feature extraction techniques, such as principal component analysis (PCA), independent component analysis (ICA), and self-organizing maps (SOM), helps in detecting anomalies in user behavior (Deshpande et al., 2017). A research investigated the user behavior in an e-commerce site for predicting the buying intention of users with the help of deep belief networks and stacked denoising auto-encoders and concluded that feature extraction from high-dimensional data achieves better predictions (Vieira, A., 2016).

Additionally, the research highlighted the importance of graph theory in anomalies detection in user behavior (Bindu et al., 2016). This research also concluded that data mining, pattern recognition, graph theory, and machine learning techniques are used to extract features from social media networks. The understanding of user behaviors in social media networks is a challenging task for social network analysis and anomalies detection. A study of social media user behavior characteristics based on their engagement at different time-span of the day has revealed positive outcomes in detecting spammers (Chen et al., 2015). A linear regression and association rule mining techniques are employed to classify user's behavior based on their past web usage activities to strengthen network security and privacy policies, and results concluded the positive impact of the implementations (Rahman et al., 2017). An insightful research in e-commerce for the classification and prediction of user's shopping trends of specific products are also studied and identified better prediction accuracy (Raju et al., 2018). A recurrent neural-network-based human behavior prediction system is proposed to predict the next activity of the user using a long-short term memory network model by analyzing activities (Almeida, A., & Azkune, G., 2018). It is proposed to provide better governance for mental impairment patients. A study attempted to find Twitter influencers and they compared three measures of influence namely the number of followers, the number of retweets, and page-rank of the tweet. They concluded that the number of followers and the page-rank measures are effective (Kwak et al., 2010). A subsequent study claimed that only the absolute number of followers is not a good measure to rank the influencers on Twitter (Weng et al., 2010). They proposed an algorithm that uses tweet related features and network-related features to distinguish the influencers in a particular domain. A probabilistic clustering approach is proposed to identify influencers on Twitter in specific domains by using 15 different features to rank users, instead of using network analysis techniques to speed-up the computation (Pal et. al., 2011). Additionally, a research aimed at identifying the most influential users on Twitter based on current friends added on the user's network (Gupta et al., 2013).

Although numerous researches have been proposed for user classification and ranking in the literature, their usages are too specific and couldn't be used to solve the generic use cases. In this research, bag-of-bi-gram based models are used to classify users across various identified categories. Furthermore, this research is ranking the users of a category based on specific parameters by using machine learning models.

3. PROPOSED METHOD

This research is aimed to classify the user's activities in specific categories and then rank the users of specific categories. In this regard, we have proposed two methods, one for classification of user's activities in various topics and another is to rank the users of specific topics.

3.1 User's Activity Classifications

Bag-of-words are a very popular method of feature extraction, it is very simple and flexible and can be used in numerous ways for extracting features from documents. A bag-of-words is the text representation of words that defines the occurrence of words within a document. Bag-of-words could be utilized with machine learning algorithms to build a bag-of-words model. Similarly, if we build vocabulary using bi-grams of text, it would be known as the Bag-of-bi-grams model. In this research we are proposing to use the Bag-of-bi-grams model to classify the user's activity in various topics, it is useful to reduce the dimensionality of document vectors and optimize the performance of the classifier. The implementation of the Bag-of-bi-grams model could be broken down into the following steps:

- Data collection

- Data cleaning using preprocessing methods
- Design the vocabulary of bi-grams
- Determine the weightage of vocabulary in the documents
- Create document vectors
- Predict the class of the documents using supervised machine learning models

In this research, we are calculating the weightage of vocabulary words by using TF-IDF and utilizing binary classifiers for each category/topic in which we want to classify the twitter activity. It would leverage the complexity of posts where a single post is falling in multiple categories. Furthermore, a binary classifier's accuracy is always higher than any multi-class classifier. Below are a few data cleaning steps we follow in this research:

- Ignore case
- Remove punctuation
- Remove stop-words
- Handle misspelled words
- Apply Word Stemming

Post feature extraction we would utilize the various binary classifiers to predict whether given post belongs to this category or now.

3.2 User's Ranking in Specific Category

In this research, we are ranking the users of Twitter using their profile features, network features along with four hyper parameters to gain better accuracy in ranking. These hyper parameters are derived from tweet's features and user's network features. These are explained in section 3.2.2 followed by useful terms to explain these parameters.

3.2.1 Hyper Parameter's Useful Terms

3.2.1.1 Context Weightage

Context weightage is the weightage of text message based on its associated metadata like hashtag, mentions, link address, embedded images. Its is given as follows in Table 1.

3.2.1.2 Activity Weightage

A activity refers to any of below task of the user:

- Post
- Reply
- Re-tweet
- Quoted-tweet
- Favorite

While activity weightage are the weightage that should be given on any above mention activities by the user. It have a very close co-relation with Context Weightage of the text, if the activity utilize any text message. It is given as follows:

Post: 0.8 + Context Weightage of Post * 0.2
Quoted Tweet: 0.6 + Context Weightage of Quoted Tweet * 0.25
Reply: 0.45 + Context Weightage of Reply * 0.3

Table 1. Context weightage of the text

Hashtag	Mention	Link	Image	Context weightage
0	0	0	0	0
0	0	0	1	0.4
0	0	1	0	0.5
0	0	1	1	0.8
0	1	0	0	0.6
0	1	0	1	1.0
0	1	1	0	1.0
0	1	1	1	1.0
1	0	0	0	0.5
1	0	0	1	0.7
1	0	1	0	0.8
1	0	1	1	1.0
1	1	0	0	1.0
1	1	0	1	1.0
1	1	1	0	1.0
1	1	1	1	1.0

Re-tweet: 0.4
Favorite:0.25

- **Total Activity Weightage:** It the the sum of Activity Weightage of all the activities performed by the user. It could be computed using equation (1):

$$\alpha TAW = \sum_{ForallActivities} ActicityWeightage \tag{1}$$

- **Relevant Activity Weightage:** It the the sum of Activity Weightage of all the activities of specific topic performed by the user. It could be computed using equation (2):

$$\alpha RAW = \sum_{ForallRelevantActivities} ActicityWeightage \tag{2}$$

3.2.2 Hyper Parameters

Below are a few parameters that are being considered as the most relevant coefficients to measure the effectiveness of influencers.

3.2.2.1 Domain Activity Coefficient (αDA)

It would measure the individual's interests in a specific domain. It would be helpful to identify whether an individual belongs to a concerned domain or just sharing random posts and being eligible for the influencer in that domain. Below are a few parameters to determine the value of the coefficients:

- Activity weightage on total activities (αTAW)
- Activity weightage on to activities of specific topic (αRAW)
- Percentage of unique links shared across all shared links (LU)

Domain Activity Coefficient (αDA) could be computed using equation (3):

$$\alpha DA = \left(\frac{\alpha TAW}{\alpha RAW}\right) * LU \tag{3}$$

3.2.2.2 Percentile Coverage Coefficient (αPC)

It would measure the individual's contribution to the relevant domain. It would be helpful to identify the percentage share of individuals in relevant domains and to find its activeness on social media platforms. Below are a few parameters to determine the value of the coefficients:

- Activity weightage on to activities of specific topic (αRAW)
- Percentile Coverage Coefficient (αPC) could be computed using equation (4):

$$\alpha PC = \frac{\sum_{ForallUsershavinglower\alpha TAW} Count1}{TotalNumberofUsers} \tag{4}$$

3.2.2.3 Activity Recency Coefficient (αAR)

It would be helpful to identify the rising stars in the concerned domain. Below are a few parameters to determine the value of the coefficients:

- Activity weightage on to activities of specific topic (αRAW) per month.
- Activity Recency Coefficient (αAR) could be computed using equation (5):

$$\alpha AR = \left(\frac{log\left(\alpha RAWoflastmonth\right)}{log\left(Avg\left(\alpha RAW\right)ofallmonths\right)}\right) * \left(\frac{log\left(\alpha RAWoflastmonth\right)}{log\left(Max\left(\alpha RAW\right)ofallmonths\right)}\right) + 0.1 * No.ofMonths \tag{5}$$

3.2.2.4 Human Person Handle Coefficient (αHPH)

It would distinguish whether the contributor is a human or bot at the same time distinguish whether the handle belongs to any company, organization, group, or a single contributor. The idea behind utilizing this coefficient is to eliminate the paid or self-advertising on social media posts. Below are a few parameters to determine the value of the coefficients.

Human Person handle Coefficient (αHPH) could be computed using equation (6):

$$\alpha HPH = \left(1 - \alpha Bot\right) \times \left(1 - \alpha Org\right) \tag{6}$$

So finally, we are using machine learning models to predict the rank of the user using above mentioned parameter along with user's profile parameters and user's network related parameters.

4. EXPERIMENTAL SETTING

This section will provide detailed information about the tweet collection strategies, data preprocessing methods, selection of machine learning algorithms, and their evaluation parameters along with the system configuration to setup this framework.

4.1 System Configuration

All the bench-marking and result calculation of this research was done on a 64-bit operating system laptop having the configuration of 16 GB memory, 4 core i7 processor, 1 TB hard disk with 2 GB graphics card. It is operating on Ubuntu 18.04 Linux operating system, furthermore, the following packages are installed on it:

- Python 3.6
- Elasticsearch 5.6.4
- MongoDB 3.6.3
- Spark 2.3.4
- Apache Flume 1.6.0
- Apache Kafka 2.2.0

4.2 Data Set

This research has considered six topics to classify user's topics of interest, namely politics, sports, cinema, business, technology, and others for the analysis purpose. This dataset is further limited for Indian tweets of the English language only, so the keywords were considered accordingly and the tweet collected against them through Flume using the Tweepy library of python using Twitter free API and data was sunk directly in the Spark cluster. Collected tweets stats are given in Table 2.

4.3 Preprocessing Methods

Preprocessing is a technique of data cleaning to maximize the effectiveness of the analytics result. In this study, we are applying the following preprocessing techniques (Singh et al., 2020):

- Handling Special Characters
- Handling Punctuation
- Non-English Word Elimination
- Slang Word Removal
- Emoji to Words
- Hashtags to Words
- Handling Para Languages
- Stop Words Removals
- Remove words of 1 character length
- Apply Word Stemming
- Tense Identification

Table 2. Statistics of experimental data

Topic Name	Number of Posts	Number of Users
Politics	10987	6219
Sports	10219	5932
Cinema	10070	7523
Business	9316	8271
Technology	9157	7216
Others	12063	7894

- POS Tagging
- Handling Para Languages

4.4 Data Processing Methods and Technologies

In this research, we are using Apache Flume to aggregate the tweets using Apache Kafka as a channel and buffer to store tweets until they are picked for processing. For processing, we are using Apache Spark to utilize distributed computing benefits. All preprocessing, processing, and feature forming steps are performed by Apache Spark using Python and after the processings data is stored in MongoDB for model's use. In this research, we have trained six different binary classifier to predict the tweets in each category. The analytics results are also stored in MongoDB.

4.5 Selection of Machine Learning Algorithms

This research is aimed to classify the users in various domains and to rank them in corresponding domains, it could be performed by supervised machine learning algorithms. In this research, we are utilizing the following 3 machine learning algorithms based on their working principle and effectiveness (Singh et al., 2020; Vieira, A., 2016; Shahzad et al., 2017; Raghuram et al., 2016; Rahman et al., 2017) on this problem space.

4.5.1 Logistic Regression (LR)

Logistic regression (LR) is a generalized case of linear regression models. It is one of the popular methods to predict the probability of the outcomes (categorical response). It utilizes a logistic function to measure the association between the input feature vector and the instance class. It could be utilized to predict binary or multi-class outcomes based on the use cases. For binary outcome, it uses binomial logistic regression, and for the multi class outcome, it uses multinomial logistic regression (Elzayady et al., 2018).

4.5.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a very effective supervised machine learning model. It seeks a hyper-plane representation in data set to classify data points in the mentioned class with the margins between them as far apart as possible (Singh et al., 2020; Shahzad et al., 2017). More generally, SVM secures higher accuracy due to its multidimensional support.

4.5.3 Long Short Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a deep learning model, it's a type of recurrent neural network. It is capable of learning order dependence effectively in sequence prediction problems and designed to remember the history of positive values for a short period. It's internal mechanisms called gates, there are three gates in its architecture, input gate to read in the input, output gate to write the output to the next layers, and forget gate to decide which data to be remembered and which data to forget (Singh et al., 2020). LSTM is highly relevant in complex problem domains like speech recognition, machine translation, and more.

4.6 Evaluation Parameters

The outcome of classification tasks could be categorized in the following four classes:

- True Positive (TP) represents the correctly predicted positive class.
- True Negative (TN) represents the correctly predicted negative class.
- False Positive (FP) represents the incorrectly predicted positive class.
- False Negative (FN) represents the incorrectly predicted negative class.

There are various evaluation metrics for classification algorithms, among them, the following four evaluation parameters are utilized to evaluate the performance for the analytics (Singh et al., 2020) in this research.

4.6.1 Accuracy

It is the ratio of correctly classified samples versus the total no. of samples. It lies between the range of 0 to 1 that respectively represents the worst to best value. A high accuracy represents that model classified substantially most of the results correctly. It could be computed using equation (7):

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

4.6.2 Precision

It is the ratio of the no. of predicted true positives by model versus the total no. of predicted positives by model. It lies between the range of 0 to 1 that respectively represents the worst to best value. A high precision represents that model classified substantially more relevant results than irrelevant ones. It could be computed using equation (8):

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

4.6.3 Recall

It is the ratio of the no. of predicted true positives by model versus the total no. of actual positive. It lies between the range of 0 to 1 that respectively represents the worst to best value. A high recall represents that model correctly classified most of the relevant results. It could be computed using equation (9):

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

Table 3. Topic wise post's classification distribution

Topic Name	Number of Tweets	Distribution Percentage
Politics	14816	46.21
Sports	4159	12.97
Cinema	5140	16.03
Business	3280	10.23
Technology	1382	4.31
Others	3286	10.25

4.6.4 F1 Score

It is the harmonic-mean of precision and recall. It lies between 0 to 1, where 0 and 1 respectively represent the worst and the best value. A high score represents a perfect precision and recall of the prediction. It could be computed using equation (10):

$$F1Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

5. RESULTS AND DISCUSSIONS

This section provides the statistical output of this research, it includes the accuracy measurement of the implemented machine learning models using the proposed methods of this research.

5.1 Topics Wise User's Classification

In this research, we are using the bag-of-bigrams to classify any tweets in various topics, this activity is backed by a various supervised machine learning algorithm to predict the classes post feature forming. In this section, we are evaluating the average posts distributions of various categories and measuring the effectiveness of these machine learning models in the classification tasks.

5.1.1 Topics Wise Post's Classification for 1000 User's Posts

In this analysis, we took a random corpus of 1000 users and classified all their tweets in various categories using our bag-of-bi-gram method. The topic's distribution is given in Table 3 and its pie-chart illustration is given in Figure 1 to better represent the distribution.

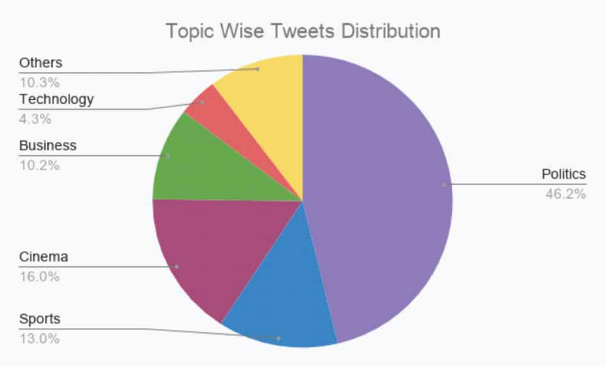
This distribution represents that most users in our corpus are interested in politics.

5.1.2 Topics Wise Post's Classification Effectiveness Measurements

We have prepared a tweets training dataset of 1k for each topic with 50% positive (related to the topic) and 50% negative (tweets not related to the topics) tweets. In this analysis, we are using a 10 fold cross-validation method, in which each model was trained on 900 tweets and validated against 100 tweets. The result of the effectiveness measurement of the proposed method is given in Table 4 along with its graphical representation in Figure 2.

In this bench-marking, LSTM performed well on Accuracy, Precision, and F1 score scale on the other side SVM secured well in Recall base analysis. So LSTM could be considered the best among them, it secured 0.875 accuracy, 0.908 precision, 0.852 recall, and 0.879 F1 score on the classification of topic "Cinema". SVM secured second place by achieving 0.850 accuracy, 0.842 precision, 0.856

Figure 1. Topic wise tweets distribution of 1000 user's tweets



recall, and 0.859 F1 score. LR was on the last position in this bench-marking and achieved 0.825 accuracy, 0.828 precision, 0.823 recall, and 0.826 F1 score.

So this method achieved the highest 87.5% accuracy using LSTM based classifier on "Cinema" topics and proofs the effectiveness of the method.

5.2 User Ranking

In this research, we are ranking the users based on a few hybrid parameters along with the user's network and user's profile based features using three machine learning models. In this section, we are bench-marking the effectiveness of these models for various topic-wise clusters of users. For conducting this bench-marking we have prepared the ranking of 500 users per topic cluster as the training dataset and we utilized a 10 fold cross-validation method, in which each model was trained on 450 records and validated against 50 records. The result of the effectiveness measurement of the proposed method is given in Table 5 along with its graphical representation in Figure 3.

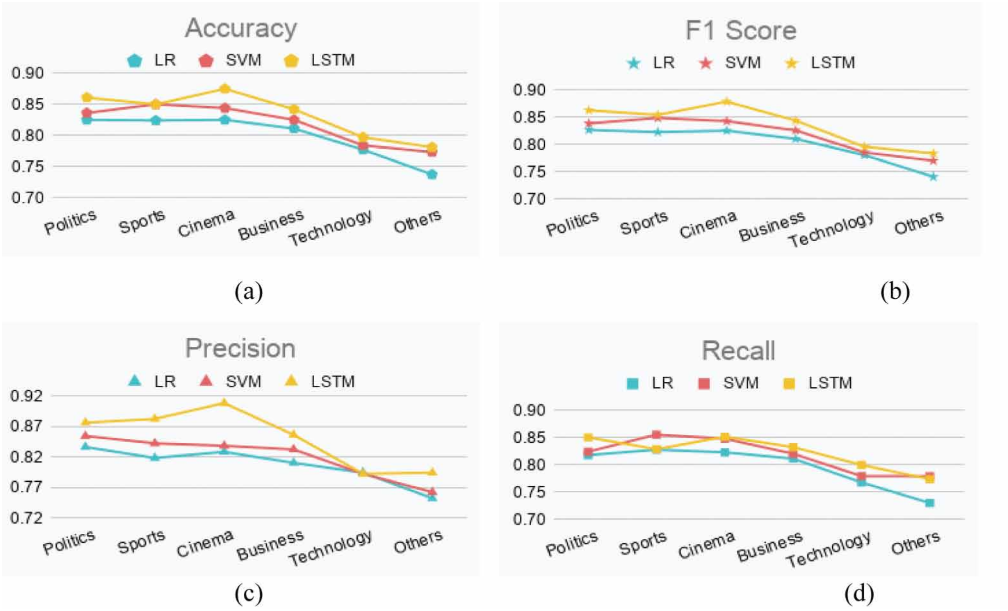
By analyzing the ranking predictions, it's observed that LSTM is leading with its performance over SVM and LR, it secured best on the ranking of users in Sports topic and achieved 0.924 accuracy, 0.952 precision, 0.902 recall, and 0.926 F1 score, similarly, SVM achieving 0.911 accuracy, 0.942 precision, 0.887 recall, and 0.914 F1 score and secured second place. LR was in the last position in this bench-marking and achieved 0.893 accuracy, 0.924 precision, 0.870 recall, and 0.896 F1 score.

So this method achieved the highest 92.4% accuracy using LSTM based model prediction. It is significant evidence for the effectiveness of the method.

Table 4. Effectiveness measurements of machine learning algorithms for classification of tweets in various topics

Topic Name	Logistic Regression (LR)				Support Vector Machine (SVM)				Long Short Term Memory (LSTM)			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Politics	0.825	0.836	0.818	0.827	0.836	0.854	0.824	0.839	0.861	0.876	0.850	0.863
Sports	0.824	0.818	0.828	0.823	0.850	0.842	**0.856**	0.849	0.850	0.882	0.829	0.855
Cinema	0.825	0.828	0.823	0.826	0.844	0.838	0.848	0.843	**0.875**	**0.908**	0.852	**0.879**
Business	0.811	0.810	0.812	0.811	0.825	0.832	0.821	0.826	0.842	0.856	0.833	0.844
Technology	0.777	0.794	0.768	0.781	0.784	0.792	0.780	0.786	0.797	0.792	0.800	0.796
Others	0.737	0.752	0.730	0.741	0.773	0.762	0.779	0.770	0.781	0.794	0.774	0.784

Figure 2. (a) to (d) are displaying the bench-marking of LR, SVM, and LSTM on the classification of tweets on various topics



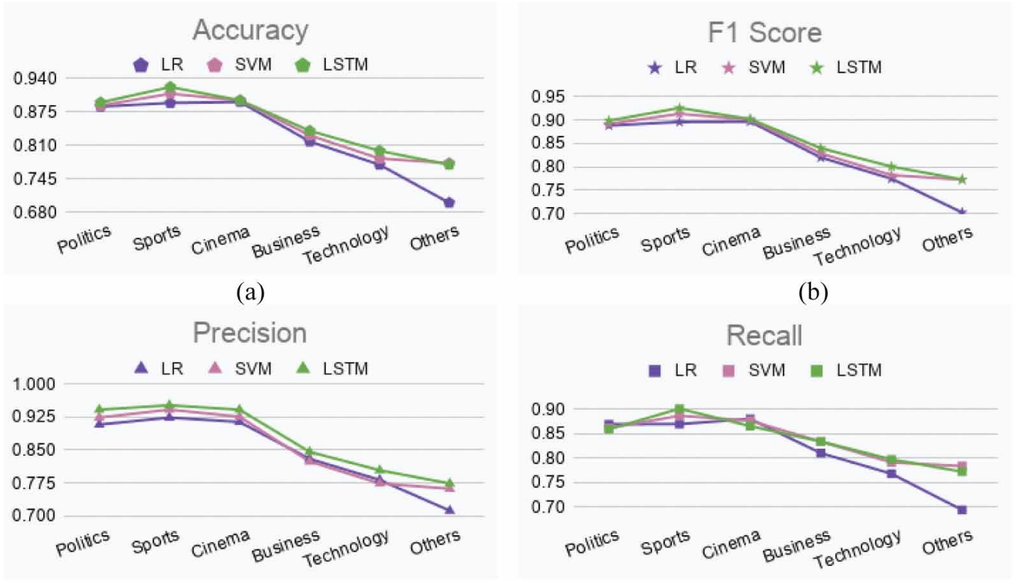
(a) (b)

(c) (d)

Table 5. Effectiveness measurements of machine learning algorithms on user ranking in various topics

Topic Name	Logistic Regression (LR)				Support Vector Machine (SVM)				Long Short Term Memory (LSTM)			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Politics	0.886	0.908	0.870	0.888	0.888	0.924	0.862	0.892	0.894	0.942	0.859	0.899
Sports	0.893	0.924	0.870	0.896	0.911	0.942	0.887	0.914	**0.924**	**0.952**	**0.902**	**0.926**
Cinema	0.895	0.914	0.881	0.897	0.898	0.926	0.877	0.901	0.898	0.942	0.866	0.902
Business	0.818	0.830	0.811	0.820	0.830	0.824	0.834	0.829	0.839	0.846	0.834	0.840
Technology	0.773	0.782	0.768	0.775	0.785	0.774	0.791	0.783	0.800	0.804	0.798	0.801
Others	0.699	0.712	0.694	0.703	0.776	0.762	0.784	0.773	0.773	0.774	0.772	0.773

Figure 3. (a) to (d) are displaying the bench-marking of LR, SVM, and LSTM on the ranking of users in various topics



(a) (b)

(c) (d)

6. CONCLUSION AND FUTURE WORK

Twitter is indeed the first choice for businesses to promote their offerings through influencers. In this cause of action, it's a very challenging task to identify the most prominent influencer for specific businesses. The open boundaries of cross-domain conversations, support of multiple languages, and unrestricted information sharing are making this selection even more complex. This research is aimed to classify the social interactions of users in various categories and to recognize the most suitable influencer(s) for any specific conditions based on tweets, user-profile, and user-network related features. In this research, we have used 3 machine learning models for classification and ranking purposes, these are namely LSTM, SVM, and Logistic Regression. User's activity (tweets) is classified in various categories using bag-of-bigrams models and achieved 87.5% accuracy, 90.8% precision, 85.2% recall, and 87.9% F1 score by using the LSTM model. Additionally, this research has achieved 92.4% accuracy, 95.2% precision, 90.2% recall, and 92.6% F1 score in the user's ranking by using the LSTM model. In both experiments, LSTM outperformed the SVM and Logistic Regression algorithms. These results are significant evidence of the effectiveness of the proposed methods.

This research could be further extended to regressively utilize the user's network, tweeting activity (Pal et al., 2011), and tweeting patterns features in the user's ranking could be optimized the ranking time by using word hashing and other techniques.

REFERENCES

Almeida, A., & Azkune, G. (2018). Predicting human behaviour with recurrent neural networks. Socio-Cognitive and Affective Computing. *Applied Sciences (Basel, Switzerland)*, *8*(305), 1–13.

Bindu, P., & Thilagam, P. S. (2016). Mining social networks for anomalies: Methods and challenges. *Journal of Network and Computer Applications*, *68*, 213–229. doi:10.1016/j.jnca.2016.02.021

Chen, G., Wang, N., Zhang, F., & Jiang, H. (2015). Understanding the time characteristic of user behavior on online forums. 2015 IEEE international conference on big data, 2300–2306. doi:10.1109/BigData.2015.7364019

Deshpande, D., & Deshpande, S. (2017). Analysis of various characteristics of online user behavior models. *International Journal of Computers and Applications*, *161*(11), 5–10. doi:10.5120/ijca2017913127

Elzayady, H., Badran, K. M., & Salama, G. I. (2018). Sentiment Analysis on Twitter Data using Apache Spark Framework. *13th International Conference on Computer Engineering and Systems*.

Grover, P., Kar, A. K., & Ilavarasan, P. V. (2019). Impact of corporate social responsibility on reputation—Insights from tweets on sustainable development goals by CEOs. *International Journal of Information Management*, *48*, 39–52. doi:10.1016/j.ijinfomgt.2019.01.009

Gupta, P., Goel, A., Lin, J., Sharma, A., Wang, D., & Zadeh, R. (2013). Wtf: The who to follow service at twitter. *22nd International Conference on World Wide Web,* 505–514.

Jain, H., & Fatema, N. (2019). Layer recurrent neural network based intelligent user activity classification model using smartphone. *Journal of Intelligent & Fuzzy Systems*, *35*(5), 5085–5097. doi:10.3233/JIFS-169793

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter Power: Tweets as Electronic Word of Mouth. *Journal of the American Society for Information Science and Technology*, *60*(11), 2169–2188. doi:10.1002/asi.21149

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is twitter, a social network or a news media? *19th International Conference on World Wide Web*, 591-600. doi:10.1145/1772690.1772751

Montangero, M., & Marco Furini, M. (2015). TRank: Ranking Twitter users according to specific topics. In *12th Annual IEEE Consumer Communications and Networking Conference*. IEEE. doi:10.1109/CCNC.2015.7158074

Pal, A., & Counts, S. (2011). Identifying topical authorities in microblogs. *WSDM '11: Proceedings of the fourth ACM international conference on Web search and data mining*, 45–54. doi:10.1145/1935826.1935843

Raghuram, M. A., Akshay, K., & Chandrasekaran, K. (2016). *Efficient User Profiling in Twitter Social Network Using Traditional Classifiers. In Intelligent Systems Technologies and Applications. Advances in Intelligent Systems and Computing* (Vol. 385). Springer.

Rahman A., Dash, S., Luhach, A. K., Chilamkurti, N., Baek, S., & Nam, Y. (2019). A Neuro-fuzzy approach for user behaviour classification and prediction. *Journal of Cloud Computing, 8*(1).

Rahman, A., Saleh, A. A., & Abraham, A. (2017). User behavior classification and prediction using fuzzy rule based system and linear regression. *Journal of Information Assurance & Security*, *12*(3), 86–93.

Raju, S. S., & Dhandayudam, P. (2018). Prediction of customer behaviour analysis using classification algorithms. *AIP Conference Proceedings*, *1952*(1), 020098. doi:10.1063/1.5032060

Shahzad, B., Lali, M. I., Nawaz, M. S., Aslam, W., Mustafa, R., & Mashkoor, A. (2017). Discovery and classification of user interests on social media. *Information Discovery and Delivery*, *45*(3), 130–138. doi:10.1108/IDD-03-2017-0023

Silva, L. D., & Riloff, E. (2014). User Type Classification of Tweets with Implications for Event Recognition. *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, 98-108. doi:10.3115/v1/W14-2714

Singh, P., Dwivedi, Y. K., Kahlon, K. S., Pathania, A., & Sawhney, R. S. (2020). Can twitter analytics predict election outcome? An insight from 2017 Punjab assembly elections. *Government Information Quarterly*, *37*(2), 101444. doi:10.1016/j.giq.2019.101444

Singh, R. K., & Verma, H. K. (2020). Effective Parallel Processing Social Media Analytics Framework. *Journal of King Saud University - Computer and Information Sciences*.

Vernier, M., Farinosi, M., & Foresti, G. L. (2018). *Twitter Data Mining for Situational Awareness. Encyclopedia of Information Science and Technology* (4th ed.). IGI Global.

Vieira, A. (2016). *Predicting online user behavior using deep learning algorithms*. arXiv:1511.06247v1 [cs.LG]

Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010). Twitterrank: Finding topic-sensitive influential twitterers. *3rd ACM International Conference on Web Search and Data Mining*, 261–270. doi:10.1145/1718487.1718520

Yousukkee. (2016). Survey of analysis of user behavior in online social network. In *Management and Innovation Technology International Conference (MITicon)*. IEEE.

Ravindra Kumar Singh is a research scholar pursuing his Phd in Computer Science Department from NIT Jalandhar. He achieved his M.Tech in Computer Science from NIT Jalandhar in 2011 and his research profile includes 15+ international journals and international conferences. Apart from his research activities he is very stable in his professional career and contributed well in various companies, most recently he possessed the role of Technical Architect in Germany based leading health care service provider company powered by AI, ML, Big Data and Block-chain Technologies.

Harsh Kumar Verma is Professor and Head of Computer Science and Engineering Department in NIT Jalandhar. His research profile includes 30+ international journals, 40+ conference publication, book chapters and research projects. He supervised 15+ research scholars for PhD in various domain in computer science. Apart from his research contribution he has been very active in other administrative activities and organized various technical events.