Twitter Likes Prediction Using Content and Link based Features

TEHMINA AMJAD*, HAFSA ZAHRA

Department of Computer Science and Software Engineering, International Islamic University, Islamabad, Pakistan

**Corresponding author's e-mail: tehminaamjad@iiu.edu.pk,*

Abstract

Twitter, a microblogging network, allow its users to post content in real-time according to their interest and share ideas, thoughts and information with each other. Contents can be an image, a movie, a link to a news article or a short message known as "Tweet". Although Twitter provides a list of most popular topics, called *Trending Topics*, but users are usually concerned about a small quantity of tweets from their own topic of interest. It is rather challenging to predict which kind of information is expected to attract interest of more users in such a large collection of tweets and can become more popular within short time interval. In this study, we use the "likes" of tweet as a measurement for the popularity among the Twitter users and study the interesting problem of Tweet Likes Count Prediction (TLCP) to explore the characteristics for popularity of tweets for top *Trending Topics* in the near future. Valuation of possible popularity is of great importance and is quite challenging. For a particular Tweet, we measure the impact of three main attributes (Tweet Content, Number of followers and Geographical Location) for TLCP by using prediction models and evaluate their performance using F-measure. A real world dataset from Twitter was extracted covering tweets from August 4, 2016 till August 21, 2016. Experimental results show that Bayesian Network outperform 70% performance with combined features (Tweet, Followers, Location) on likes as a best predictive model than others on the basis of Accuracy, Precision, Recall and F-measure.

Keywords: Online Social Network, Twitter, Trending Topics, Tweet Likes Prediction, Classification, Prediction.

INTRODUCTION

Among today's online social networks, one of the most popular microblogging sites is Twitter (https://twitter.com). It provides a fast mode of communication and information spread. Retweet is considered to bekey instrument for information propagation and researchers have paid great attention on analyzing and predicting retweet behavior. The user-generated *"Tweet content"* in Twitter is composed of short messages known as tweets, containing up to 140 characters, which can also contain images or links to news articles or videos [1].We only used the "*Tweet Contents"* that are in the form of English text. For

example: "ustad rahat fateh khan rocks independence concert". This tweet content is concerned with a Trending topic (Rahet Fateh Ali Khan). Tweet content is basically considered as a tweet and is used among the eight features in the features definition. Users can like a particular tweet using the "Like" button which is a heart shaped symbol. One of the most remarkable thing about Twitter is its Real-time nature. Hence, on a large number of topics, tweets are valuable features that may reflect real-time news relevant to happening events that occur in any place of the world. The time of popularity of a tweet can vary depending upon its topic like, sports, economic crisis, elections, celebrities, death, singer's latest album, airline crash news, mother's day and so on. A content of a tweet which is discussed and liked by more users in Twitter, within a few days to a few months or even a year, makes a tweet successful or popular. This observation is basically taken from Twitter, as when a topic is discussed more by users, Twitter automatically generate that topic by using the hash tag at the start of Trending Topic. e.g. *#World Bank*. Thus, Trending Topic in the Twitter can be in the form of keywords, phrases and hash tags. In this study we used two methods to find the top *Trending Topics*. First, we used Latent Dirichlet Allocation (LDA) [2] for finding the *Trending Topic* for calculation of topic rank by assigning the ranking positions calculated by LDA on the basis of average likes. Second, we predict the *Trending Topic* by using three main features and then combined the features to check the impact of likes through four Classification models.

Comprehensive studies have been conducted in literature to predict the retweet behavior, but analysis of likes and its impact on information spread has not yet gained attention of researchers. In this study, we attempt to predict tweet likes using different features on five separate trending topics by applying prediction models on a dataset with five different trending topics. The interesting problem of TLCP faces several challenges: The first challenge for TLCP is to find out the most effective features that are significant for future likes prediction from several aspects such as tweet content, number of followers of a user and impact of its geographical location. This study introduces a set of features that can be used for prediction of likes of tweets. Second challenge for TLCP is to conjoin all appropriate features to distinguish the possibly interesting tweets. Thus for TLCP, we first defined eight features on the basis of tweet likes and secondly we applied different prediction models using three important features to predict tweet likes. We rank and predict the impact of features on the basis of likes so "Likes" signalize an important attitude towards the Trending Topics.

LITERATURE REVIEW

Background

Twitter provides a lot of features to its users, some of which are discussed here in this section. A *Tweet* is basically a short message of up to 140 characters. It can contain images, links to news articles**,** video or text messages that contains (keywords, URL, hash tags, RT symbols) and so on. Each user has a *Twitter user id.* For getting that id each user has to make a twitter account for which user provides personal information about him in a profile page.

User personal information can be: User id, User URL, User Screen Name, Location, Date of Birth, etc.

Twitter follows a specific format by using topics, hash tags that starts with "#" symbol. Likewise to reply or comment a specific post "@" symbol is used with the start of those username. It provides a feature namely *Hash tag* which contains keyword or phrase. Users use hash tag (#) symbol to categorize those tweets with keywords.

Twitter further gives a concept of *followers*. It allows users to follow others by making a social network graph. For example, in Twitter there are two users named as "a" and "b". If user "a" finds post of user "b" interesting and wants to follow user "b" then "a" can do so by clicking on "Follow" icon and can access tweets posted by user "b".

Likes are a way for users to show their liking for a particular content. Liking means to appreciate some content and it is represented by a heart symbol in twitter. Twitter users can forward other's tweets to their profile so that those user's followers can get notification about shared tweet, forwarded by a specific user. This process is known as *Retweet*. The topics that are discussed by more Twitter users are called *Trending Topics*.

Related Work

From review of the related literature we found work done regarding retweet prediction, activity prediction, trending topics prediction either (Trending or non-Trending).The identity of source of the tweet and the retweeter are most significant features for the prediction of retweet behavior[3]. A Bayesian approach was applied for predicting the popularity of a tweet using the time series path of its retweets[4]. Yang *et al*. studied the retweet behavior of the tweeter users and found that 25.5% of the tweets are actually retweeted from the friend's blogs[5]. They proposed a factor graph model to predict users' retweeting behaviors. Similarly, Kupavskii *et al*. studied the retweet behavior and they trained a model to forecast that how many retweets a given tweet can attain over a fixed time period [6]. Matchbox Model [7] was used to predict the future retweets using data of what was retweeted within an assured time window (1 hour). Maximum retweets within one hour was represented as a positive feedback and lack of retweets was represented as negative feedback. Kathy Lee *et al*. [8] emphasized on classification of the trending topics provided by twitter into general categories with high accuracy for better information retrieval. Real-time classification of twitter trends explores the types of triggers that spark trends on Twitter, by studying the earliest tweets that produce a trend [9], [10]. Weerkamp *et al*.[11] predict twitter activities of users that enable them to share their posts with everyone. The main focus is in future activities of users and plans of twitter users rather than their current or past activities. For future certain timeframe (tonight, next week, tomorrow) and microblog messages, attempt to predict the smart activities for future timeframe. Das *et al*. [12] proposed a machine learning based approach for prediction of twitter trends. Rosa *et al*. [13] proposed Twitter Topic Detection to detect most popular topics from a large collection of trending topics in Twitter. Twitter Topic Fuzzy Fingerprints method is used first and then compare this method with two text based classifiers namely, K-Nearest Neighbors (KNN) and Support Vector Machines (SVM). Yan *et al*.[14] proposed to predict the future citations

to examine the popularity among scientists. From a large collection of research articles, it is challenging to predict which type of research articles are more cited by researchers. Dataset is taken from computer science domain. Several features (Topic Rank, Diversity, Author Rank, Recency, Authority, Sociality, H-Index, Venue Rank, Venue Centrality, and Productivity) defined to predict the popularity rate for each article in near future.

We compared our results by applying the different Classification Models like Bagging and Random Subspace for Prediction rather than SVM, AdaBoostM1, LogitBoost, Matchbox Model etc. that are used in existing studies. Moreover, in the first part we took the idea of our eight different features from base paper (Citation count prediction: Learning to estimate future citations for literature) [14] for ranking of these Twitter features from different aspects. In this study our main focus is on Twitter likes count prediction for *Trending Topics* with the use of several features as input to study the correlative characteristics for popularity. The details of the data and method are discussed in section 3.

TWITTER LIKES COUNT PREDICTION

Features Definition

For TLCP, we have calculated some features that can play significant role for prediction of likes. These features are defined as follows:

Topic Rank. This feature calculates the probability distribution over topics allocated to each Tweet "t". We apply unsupervised LDA [2] to find the topic probability of each topic. That is, for each 5 topics, the topic model calculates $p(topic_i|t)$, the inferred probability of topic "i" in tweet "t". The topic distribution T (t) is:

$$T(t) = \{p(topic_1|t), p(topic_2|t) \ldots p(topic_5|t)\} \quad (1)$$

Then to estimate the total likes of a specific topic from tweet "t", signified by LT $(topic_i|t)$, allocate the likes of the tweets LT (t) permitting the topic distribution $T(t)$, i.e.,

$$LT(topic_i|t) = LT(t) \times p(topic_i|t) \quad (2)$$

Likes of all 5 topics are obtained by using this formula:

$$LT(topic_i) = \sum_{t \in T} LT(topic_i|t) \quad (3)$$

Formula (3) is the sum of total likes of all the tweets for five *Trending topics* that are obtained by formula (2).

Where T is the topic collection. Topic popularity is considered as top ranked Topic on the basis of Average Likes. Table 1 shows the topic ranks for selected topics.

Table 1: Topic Rank (Popularity)

Sr. No.	TOPIC RANK (Popularity)	Average Like counts
1.	World Bank	11393812.61
2.	RIO Olympics	10695198.14
3.	Rahat Fateh Ali Khan	10657601.55
4.	Loreal	10638107.69
5.	Nawaz Shareef	10360027.22

Diversity. We obtain the concept of diversity of a tweet from its topic distribution. When a tweet has a wide range of viewers, it is likely to be liked by tweeter users who are interested in various trending topics. To calculate the like's range for a Topic tweet, we calculate the diversity of the collection of Tweets (T) topic distribution according to this formula:

$$\text{Diversity (t)} = \sum_{t=1}^{|T|=5} -\text{p (topic}_i|\text{t)} \cdot \log \text{p (topic}_i|\text{t)} \quad (4)$$

Tweet Index (T-index). To measure the productivity and impact of the tweets of a particular topic, T- index is a useful indicator. We calculate T-Index by taking average of each user's tweets.

User	Tweet Likes	
Hafsa	I Love Pakistan	344
	Excellent effort for Metro Project	400
	Loreal is best fashion brand	290
	Sum= 344+400+290=1034	
	Average (T-index)= 1034/3=344.6	

Therefore, to predict average likes for each user in Twitter we consider T-index as a candidate variable.

Retweet (RT). To determine the spread of a tweet we determine that a tweet was retweeted or not. In our dataset we identify it with yes or no for each tweet. RT(yes) means that the tweet was retweeted and similarly

RT(no) represents that tweet was not retweeted.

Twitter User Rank. We calculate the Twitter user rank according to user average tweet likes count (T-index). Each user has his/her own probability of tweet likes. We analyze all users against their average tweet likes and allocate each of them a rank position number. We do

this task for all 5 trending topics and represent top 10 users for each trending topic dataset individually.

Authority. We associate the authority of a tweet with number of times it has been retweeted. We analyze all tweets against retweet counts and then allocate them a rank position number. We calculate the authority of tweets for selected 5 categories and find top 10 leading tweets from overall records on the basis of highest retweet counts.

Geographical Location Rank. We calculate user location rank on the basis of their T-index. Geographical location of a user can also attract more users. The rank of a geographical location is calculated by finding the average number of likes for that location. For 5 selected categories we calculated top 10 locations on the basis of T-index.

Sociality. Twitters users can have different number of followers. Due to that 'follower' relationship a network of twitter users emerges in which users are nodes that are connected with each other via edges represented by follow relationship. It can be assumed that a tweet from a widely connected user has a higher chance to be liked by a wide range of followers. For 5 selected topics, we calculated sociality by calculating the Correlation formula in MS Excel to calculate the values of sociality for each type of correlation. Correlation between the followers and retweet indicates that a user who is having a maximum number of followers in his/her profile is having a larger chances of user's tweet to be retweeted by his/her followers but in that case if the tweet content is more informative or useful according to user's interest.

1. Correlation between followers and likes count
2. Correlation between followers and retweet count
3. Correlation between Average follower and Average Retweet count
4. Correlation between Average follower and T-index

Prediction Models

K-Nearest Neighbor (KNN). KNN algorithm is a method for classifying objects based on closest training examples in the feature space by a majority common vote amongst its k nearest neighbors [14]. It is widely used for statistical estimation, pattern recognition and regression. We have used it by simply assigning the property value (in our case, likes) for the object (i.e., tweet t) to be the average of the values of its k nearest neighbors to predict the value based on a similarity measure. The neighbors are taken from a set of objects for which real likes counts are known.

Bayesian Network (BN). BN is a directed acyclic graph whose nodes represent events in a domain [15]. These events are connected with directed links, which represent an association or a causal relationship between them. BN can be considered as a network of events connected by the probabilistic dependencies between them [16]. In our case, we calculate the conditional probability according to Bayes theorem as P(Y | X), which is the probability of the event Y (likes) conditional on a given outcome of event X (the selected feature)

Bagging. Bagging is described by a set of classifiers, called base learners, (C1, C2… Ck) that are obtained from a set of bootstrap samples (D1, D2 …Dk) to form an ensemble method

for prediction [17]. The objective is to predict new samples by a set of classifiers and prediction is finalized by taking a majority vote. To achieve very high accuracy, Boosting, Random forest classifier, decision trees (e.g. CART) were combined for individual as well as combined features.

Random Subspace. Random subspace method, also called attribute bagging adds an additional layer of randomness to bagging. Unlike the standard trees where each node is split using the best split among all variables, a random forest splits each node by randomly choosing from the best of predictors at that node[18].

EXPERIMENTS AND EVALUATION

Dataset

Micro Strategy is a Business Intelligence application software that enables users to crawl dataset from social network sites. A 64-bit architecture namely Micro Strategy platform, maintains OLAP reports that are stored in memory as datasets[19].To import the dataset from social network like Twitter, Facebook, Dropbox, Google Drive and so on. User has to configure online to connect with internet and to login with his/her account e.g. (Twitter, Facebook) to enter a specific query to import the required dataset. Attributes are created and used for further operations. Dataset can be exported in the form of excel and Pdf. We extract the Twitter Trending Topics dataset from this software and then export dataset into an excel file. We enter queries for *Trending Topic*s to extract the dataset from August 4, 2016 to August 21, 2016.The tweets related to five Trending Topics of August 2016 namely, Politics (Nawaz Shareef), Economy (World Bank), Sports (Rio Olympics 2016), Fashion Brand (Loreal), Music (Rahet Fateh Ali Khan) were extracted. After preprocessing, the statistics of the dataset are given in Table 2.

Table 2: Statistics of Twitter dataset

Sr. No	Trending Topics	Records	Tweets
1.	ECONOMY: (World Bank)	1592	1853
2.	FASHION: (Loreal)	1183	1662
3.	SPORTS: (RIO Olympics 2016)	1611	1996
4.	MUSIC: (Rahat Fateh Ali Khan)	610	777
5.	POLITICS: (Nawaz Shareef)	642	1070
Total		**5638**	**7358**

In Table 2 each *Trending Topic* consists of different number of records. Records means the total number of rows that are having a complete information about Tweet, Location, Date, Followers, and Retweet against Twitter user. Likewise, there is also different number of tweets against each user for each *Trending Topic*. The amount of tweets is more

than the number of records because some users are having more than one tweet for a specific *Trending Topic*. So this is the basic difference between Record and Tweet. For example: A user having 3 tweets against a Trending Topic(Nawaz Shareef) so we extract the total number of tweets against each individual user from the total number of record(# of rows) from the dataset for five different *Trending Topics*.

Initially the overall extracted dataset size is huge but after preprocessing, a lot of missing data in records are excluded from the dataset so that's why the dataset size remains small.

Performance Evaluation and Feature Analysis

For empirical evaluation WEKA[20], [21] was used to predict top tweet likes for different trending topics in Twitter by applying four prediction models (KNN, BN, Bagging, Random Subspace) using 10 fold cross validation. For performance evaluation we calculated accuracy, precision, recall and F-measures. The results were then compared to find which model provides better accuracy as compared to others. WEKA is a set of machine learning and data mining algorithms. It supports methods for classification, regression, data pre-processing, clustering, association rules and visualization.

In TLCP, we define three main features that are significant for future likes from several characteristics which are as follows: (Tweet content, Followers, and Geographical location). We first checked the impact of these three important features individually for the prediction of likes and then checked combined impact of all applying predictive models.

Accuracy is defined as the percentage of unseen (Testing data) that are correctly classified, predicted by a model on the given dataset. We have computed the accuracy on ranked number of likes for five different *Trending Topics* to evaluate their results on the basis of that how much estimated value is much closer to the predicted value. Accuracy gives precise results in a dataset for both true positive and true negative. We calculate Accuracy by using this formula:

$$Accuracy = tp+tn \,/\, tp+fp+fn+tn$$

Table 3:Accuracy (A) and F-Measure (F) for all prediction models with different and combined features

	KNN		BN		Bagging		Random Subspace	
	A	F	A	F	A	F	A	F
Tweet content	66.97%	0.666	68.12%	0.685	65.59%	0.663	66.59%	0.662
Followers	25.85%	0.259	33.55%	0.295	32.15%	0.305	29.97%	0.272
Location	35.89%	0.356	38.01%	0.360	36.61%	0.351	36.76%	0.354
Combined	60.42%	0.603	70.00%	0.693	65.79%	0.657	67.35%	0.67

Table 4: Precision (P) and Recall (R) for all prediction models with different and combined features

	KNN		BN		Bagging		Random Subspace	
	P	R	P	R	P	R	P	R
Tweet content	0.671	0.67	0.747	0.681	0.843	0.656	0.692	0.666
Followers	0.259	0.258	0.29	0.335	0.298	0.321	0.282	0.312
Location	0.355	0.359	0.413	0.38	0.381	0.366	0.374	0.368
Combined	0.603	0.604	0.717	0.691	0.779	0.658	0.714	0.673

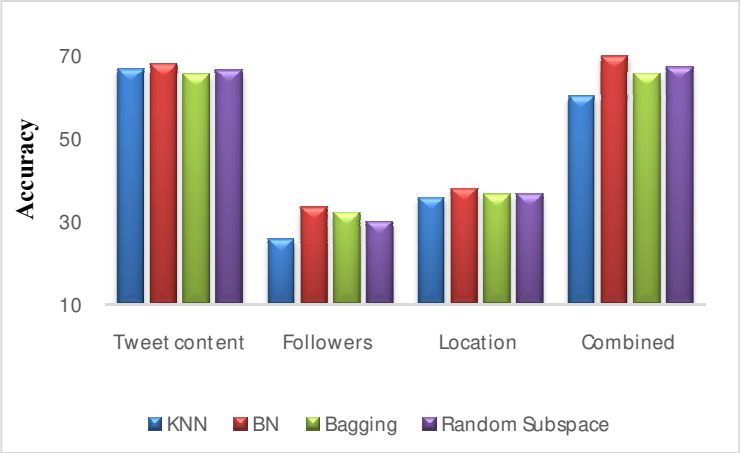

Figure 1: The Accuracy of prediction models

From Table 3, Table 4 and Figure 1, we notice that BN performs better than all other models. Using individual features, Tweet content was the most powerful feature for prediction of likes with an accuracy of 68.12 % using Bayesian Network. It is clear that Tweet contents are likely to be the more important, informative and interested to be liked by many users for leading Trending Topics. Surprisingly, Followers (33.55 %) and Geographical Location (38.01 %) is shown to have the minimum impact as compared to Tweet content. BN performs better with all features when combined for prediction, with an accuracy of 70 %. It appears that Tweet content is more important as compared to the Followers. The reason can be the fact that sometimes followers can be inactive on Twitter when the content was posted by user. That's why followers were unable to read or like the content. Likewise, Location also relies on user's interest. If users from different locations are interested in a specific global topic (e.g World Bank), those users will be able to like and share the contents which are relevant to World Bank and if some users are not interested in that topic, no chances for that topic to be liked from different locations.

KNN performs better using Tweet content and combined features for prediction of likes with an accuracy of 66.97% and 60.4197 % respectively. However, KNN shows least performance when the features used are Followers (25.8459 %) and Location (35.89 %). The reason behind the performance is the methodology of KNN that tries to find the most related neighbors and take the most correlated neighbor's like as the estimated likes. However, it consumes less data from the huge amount of dataset by applying Ten-fold cross validation. Bagging and Random Subspace classifiers perform better than KNN classifier. BN outperforms than these classifiers.

Bagging gives an accuracy of 65.59% for Tweet content, 32.15% for followers, 36.61% for Location and 65.79% for the combined features. Likewise, Random Subspace shows accuracy of 66.59% for Tweet contents, 36.76% for Location, 29.97% for Followers, and 67.35% for combined features. According to results, the impact of Location and Followers is much smaller as compared to the tweet content for the prediction of likes. The more comprehensive will be the tweet content, more is the chance to get more likes irrespective of the number of followers you may have or whatever is your geographical location. However all prediction models perform better when the features were combined for the predictions. For all combined features BN outperforms all other models with an accuracy of 70%.

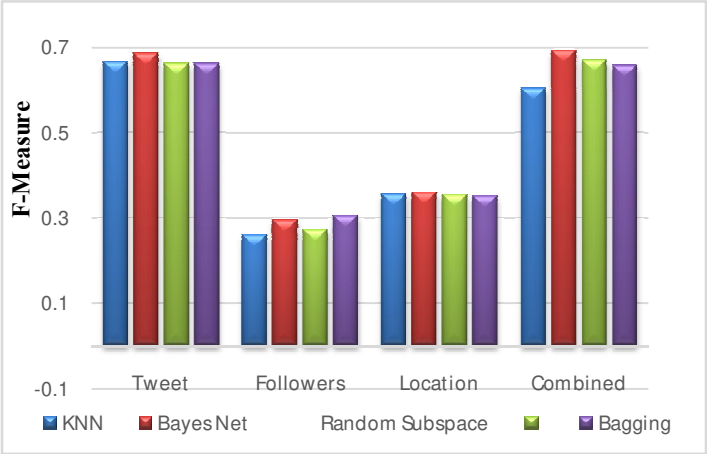

Figure 2: The F-measure of prediction models

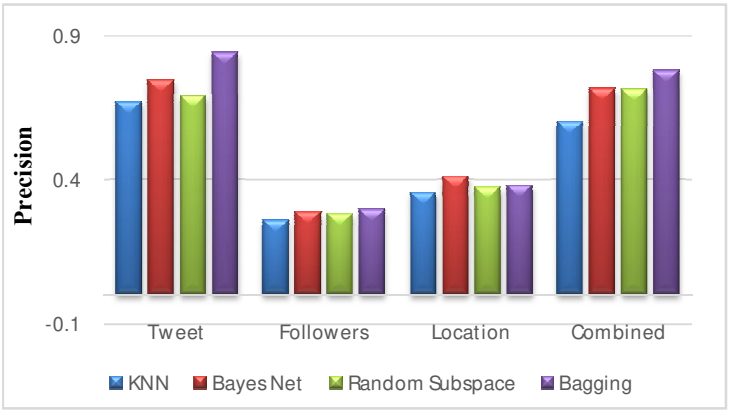

Figure 3: The Precision of prediction models

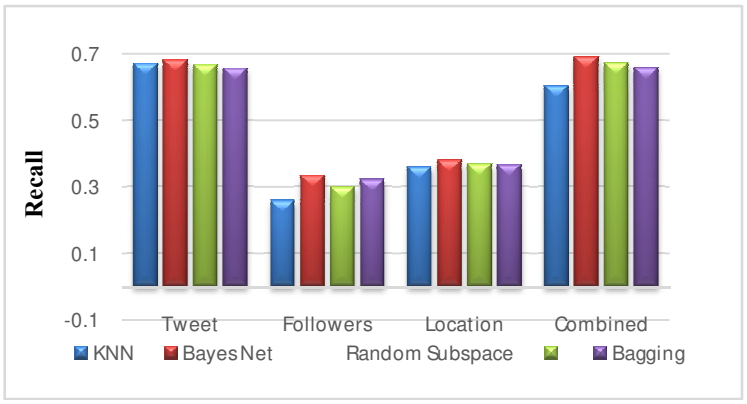

Figure 4: The Recall of prediction models

In Figure 2, F-measure is computed by four prediction models and we can say that values of F-measure in Bayes Net for Tweet content and combined features are greater than other features. In Figure3 by applying the Bagging model, the values of Precision for Tweet content and combined are greater than other features in other models. In Figure 4 the value of recall for Tweet content and for combined features is greater than other features in other models. So we can say that overall performance of Tweet content and combined features are well predicted by Bayes Net prediction model rather than others and these both features have more impact on number of likes in Twitter for *Trending Topics*.

In this paper, we predict the top most Trending topics whose tweets can be further discussed and liked by users in future and to check the impact of features (Tweets, Location, and # of Followers) individually and combined on the basis of likes by using four predictive models in WEKA.

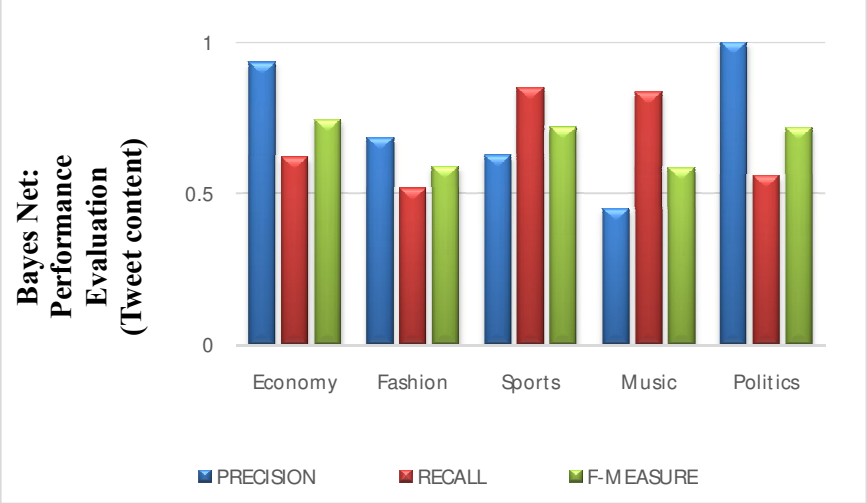

Figure 5: Bayes Net (Tweet, Likes)

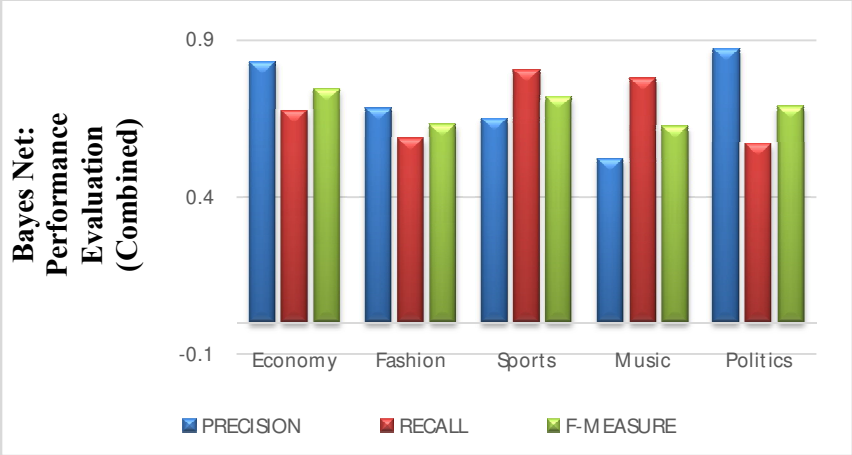

Figure 6: Bayes Net (Combined, Likes)

According to Figure 5and Figure 6.The overall performance of Bayesian Network, the impact of (Tweet vs Combined) on the basis of *Likes* shows that Economy has largest F-measure performance than other *Trending Topics*. It means according to Trending Topic: Economy (World Bank) with Tweet and Combined variables has the largest impact on the number of likes in future.

CONCLUSION AND FUTURE WORK

In this paper we studied TLCP for different *Trending Topics* in Twitter which predicts the future likes for tweets for five trending topics. We calculated eight different features and assigned them a ranking position on the basis of tweet likes for all the different Trending Topics separately. We applied prediction models including KNN, BN, Bagging and Random Subspace on likes using three main features (Tweet content, Followers, Geographical Location) individually as well as combined features. BN performs better as compared to other classifiers on the basis of Accuracy, Precision, Recall and F-measure. Using individual features, Tweet content was the most powerful feature for prediction of likes with an accuracy of 68.12 % using BN. It is clear that Tweet contents are likely to be the more important, informative and interested to be liked by many users for leading Trending Topics. Surprisingly, Followers (33.55 %) and Geographical Location (38.01%) is shown to have the minimum impact as compared to Tweet content. BN performs better with combined features for prediction with an accuracy of 70%. On individual basis, KNN performs better on Tweet content and for combined variables to check the impact on the basis of likes by correctly classifying with accuracy of 67% for Tweet content and 60.4197% for combined. KNN has least performance for Location (35.89 %) and Followers (25.8459 %) impact on number of likes. The reason behind this performance is that KNN tries to find the most related neighbors and take the most correlated neighbor's likes as the estimated likes though consumes less data from the huge amount of dataset by applying Ten-fold cross validation. Bagging and Random Subspace classifiers performs better as compared to KNN classifier. The main advantage of BN is that it computes all the instances, even missing data can be handled by it effectively. Actual data can be (incrementally) combined with predicted data to better estimate the accurate knowledge by making probabilistic predictions. In our proposed work, since content based Tweet dataset is relatively imbalance, this work can be extended by handling imbalance datasets in order to further improve and validate the performance results presented in our existing work. In future we will experiment the prediction with all features proposed in this study instead of using three features only. This will give more comprehensive results for prediction of tweet likes.

REFERENCES

[1] Z. Xu and Q. Yang, "Analyzing user retweet behavior on twitter," in *Proc. Int. Conf. Adv. Soc.Netw.Anal.Mining* (ASONAM), 2012, pp. 46–50 [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.208.5687&rep=rep1&type=pdf

[2] D. M. Blei *et al.*, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003 [Online]. Available: http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[3] T. R. Zaman *et al.*, "Predicting information spreading in twitter," in *Workshop on Comput.Soc.Sci.Wisdom of Crowds*, 2010, vol. 104, no. 45, pp. 17599–17601.

[4] T. Zaman *et al.*, "A Bayesian approach for predicting the popularity of tweets," *Ann. Appl. Stat.*, vol. 8, no. 3, pp. 1583–1611, 2014.

[5] Z. Yang *et al.*, "Understanding retweeting behaviors in social networks," in *Proc. 19th ACM Int.Conf. Inf.Knowl.Manag*. 2010, pp. 1633–1636 [Online]. Available: http://www.cs.cmu.edu/~./ziy/pubs/CIKM10-Yang-et-al-Understanding-Retweeting.pdf

[6] A. Kupavskii *et al.*, "Prediction of retweet cascade size over time," in *Proc. 21st ACM Int. Conf. Inf.Knowl.Manag*., 2012, pp. 2335–2338 [Online]. Available: https://www.researchgate.net/profile/Andrey_Kupavskii/publication/262368746_Prediction_of_retweet_cascade_size_over_time/links/5585706108ae71f6ba8e8fee/Prediction-of-retweet-cascade-size-over-time.pdf

[7] D. H. Stern *et al.*, "Matchbox: large scale online Bayesian recommendations," in *Proc. 18th Int.Conf. World Wide Web*, 2009, pp. 111–120 [Online]. Available: http://www2009.eprints.org/12/1/p111.pdf

[8] K. Lee*et al.*, "Twitter trending topic classification," *IEEE 11th Int. Conf. Data Min. Workshops (ICDMW)*, 2011, pp. 251–258.

[9] A. Zubiaga*et al.*, "Real-time classification of twitter trends*," J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 3, pp. 462–473, 2015 [Online]. Available: https://arxiv.org/pdf/1403.1451.pdf

[10] A. Zubiaga *et al.*, "Classifying trending topics: A typology of conversation triggers on twitter," in *Proc. 20th ACM Int.Conf. Inf.Know.Manag*., 2011, pp. 2461–2464 [Online]. Available: http://pensivepuffin.com/dwmcphd/syllabi/infx598_wi12/papers/twitter/zubiaga. TwitterTrending.CIKM11.pdf

[11] W. Weerkamp *et al.*, "Activity prediction: A twitter-based exploration," in *SIGIR Workshop Time-aware Inf. Access*, 2012 [Online]. Available: https://pure.uva.nl/ws/files/2475778/163223_taia2012_activities.pdf

[12] A. Das *et al.*, "Predicting Trends in the Twitter Social Network: A Machine Learning Approach," in Panigrahi *et al*., (eds) *Swarm, Evol.MemeticComput.SEMCCO 2014. Lecture Notes in Computer Science*,vol 8947.pp. 570–581 [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-20294-5_49

[13] H. Rosa *et al.*, "Detecting a tweet's topic within a large number of Portuguese Twitter trends," in *OASIcs-OpenAccess Ser. Inform.*, 2014, vol. 38 [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2014/4569/pdf/17.pdf

[14] R. Yan *et al.*, "Citation count prediction: learning to estimate future citations for literature," in *Proc. 20th ACM Int.Conf. Inform.Knowl.Manag.*, 2011, pp. 1247–1252 [Online].Available: https://pdfs.semanticscholar.org/6637/c218fdf448543e837e315e2109 ae455e7977.pdf

[15] T. D. Nielsen and F. V. Jensen, *Bayesian Networks and Decision Graphs*. Springer Science & Business Media, 2009.

[16] C. Van Koten and A. R. Gray, "An application of Bayesian network for predicting object-oriented software maintainability" *Inf. Softw. Technol.*, vol. 48, no. 1, pp. 59–67, 2006 [Online]. Available: https://ourarchive.otago.ac.nz/bitstream/handle/10523/919/dp2005-02.pdf

[17] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996 [Online]. Available: https://link.springer.com/content/pdf/10.1007/BF00058655.pdf

[18] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001 [Online]. Available: https://doi.org/10.1023/A:1010933404324

[19] M. Rouse, "MicroStrategy," Tech Target, 2017.[Online]. Available: http://searchbusinessanalytics.techtarget.com/definition/MicroStrategy.

[20] G. Holmes *et al.*, "Weka: A machine learning workbench," in *Proc. 1994 2nd Australian New Zealand Conf.Systems, Intelligent Inform.*, 1994, pp. 357–361 [Online].Available: https://researchcommons.waikato.ac.nz/bitstream/handle/10289/1138/uow-cs-wp-1994-09.pdf

[21] R. Arora, "Comparative analysis of classification algorithms on different datasets using WEKA," *Int. J. Comput.Appl.*, vol. 54, no. 13, 2012 [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.258.9202&rep=rep1&type=pdf