Trend Analysis on Twitter for Predicting Public Opinion on Ongoing Events

Tejal Rathod
Research Scholar,
Kadi Sarva Vishvavidhyalaya
Computer Eng. Department
LDRP-ITR, Gandhinagar

Mehul Barot
Assistant Prof.
Computer Engineering Department
LDRP-ITR, Gandhinagar

ABSTRACT
Twitter is most popular social media that allows its user to spread and share information. It Monitors their user postings and detect most discussed topic of the movement. They publish these topics on the list called "Trending Topics". It show what is happening in the world and what people's opinions are about it. For that it uses top 10 trending topic list. Some topic will trend at some point in the future and others will not. We wish to predict which topics will trend. And apply algorithm to find out what public opinion about that topic which use to predict mood. In this paper, we propose model which use machine learning algorithm and classify sentiment of twitter message. For that we collect tweet, preprocess that tweet, find trending topic and apply multi classifier algorithm which predict public mood. We are going to use different measure such as precision, recall, F-measure. We will going to achieve better accuracy.

General Terms
Machine learning algorithm, information retrieval, classification.

Keywords
Social media, Twitter, Twitter Trending Topic, Topic Detection, Text mining, Polarity detection.

1. INTRODUCTION
Social media is a rich resource of information about actual world action of all type twitter is one of them. It is most popular micro blogging site which allow their user to share information and short message which is called tweet. Where millions of people tweet every day. Twitter exchange wide variety of local and real world event. Twitter having two features [2]:

- The shortness of tweets, which cannot go beyond 140 characters, it facilitates Creation and sharing of messages in a few seconds
- Easiness of spreading message to a large number of user with in little time.

Twitter has standard syntax which listed follow [3]:

- User Mentions: when a user mentions another user in their tweet, Place @-sign before the corresponding username. Like @ Username
- Retweets: Re-share of a tweet which is posted by another user called retweet. By coping original tweet user consider that message of interest to other.
- Replies: when a user wants to reply an earlier tweet, they place the @username mention at the beginning of the tweet, e.g., @username I have question on what you say.
- Hashtags: Hashtags included in a tweet tend to group tweets in conversations or represent the main terms of the tweet, it usually referred to topics or common interests of a community. It is differentiated from the rest of the terms in the tweet in that it has a leading hash, e.g., #hashtag.

Twitter give list of most discussed topic at the movement which is called "Trending topic". It shows what people discussing what is going on their mind.

Following image shows how trend shows on twitter:-

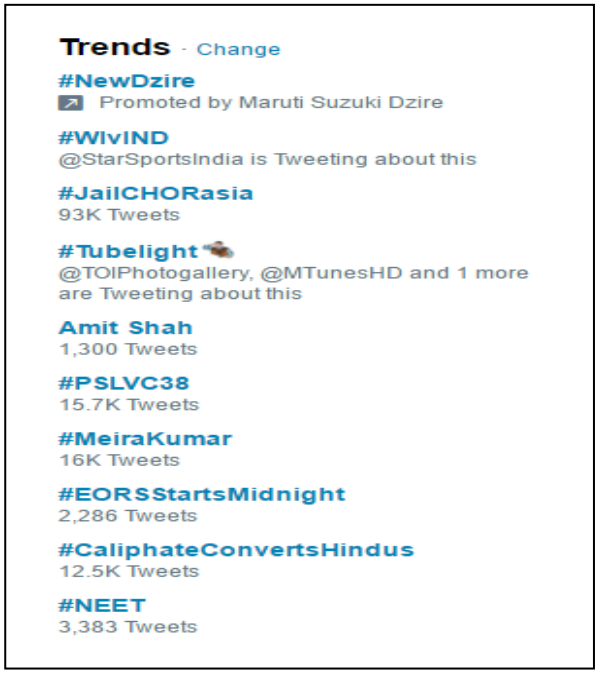


Fig 1: Twitter Top ten Trend list

In this paper we propose model which is use to predict public opinion what they talking about. We can predict polarity about different events, sports, Economy, politics etc. We collect tweets about particular event and predict public opinion about that event for that first we have to do pre-processing of tweets then apply feature extraction and find out

polarity by applying machine learning algorithm. For polarity detection we can use two type of classification. Binary classification and multiclass classification.

In binary classification we have to predict public opinion in two category like positive or negative. Where is multiclass classification we can use more than two category like positive, negative, neutral. Or strong positive, positive, medium, negative and strong negative and ranking by numbers 1, 2, 3, 4, and 5 (1-2 negative and 4-5 positive).

2. LITERATURE SURVEY

Trend analysis and based on that predicting public opinions. It plays important role, many researcher working on automatic technique of extraction and analysis of huge amount of twitter data. In [1] author compare six trend detection method and find that standard natural language processing technique perform well for social streams on particular topic. They conclude that n-gram give best performance other than state-of-art techniques. In [4], the authors have used three different machine learning algorithms Naïve Bayes, Decision Trees and Support Vector Machine for sentiment classification of Arabic dataset which was obtained from twitter. This research has followed a framework for Arabic tweets classification in which two special sub-tasks were performed in pre-processing, Term Frequency-Inverse Document Frequency (TF-IDF) and Arabic stemming. They have used one dataset with three algorithms and performance has been evaluated on the basis three different information retrieval metrics precision, recall, and f-measure. In [6] author proposed supervised learning techniques to classify twitter trending topic for that they use text based and network based classifier and conclude C5.0 gave best performance. In [19] author propose model which predict public opinion on political event by Appling different classifier which predict that whether mood is positive or negative. In [26], the authors proposed a way to get the pre labeled data from twitter which can be used to train SVM classifier. They used the twitter hash tags to judge the polarity of tweet. To analyze the accuracy of proposed technique, a test study on the classifier was conducted which showed the result with the accuracy of 85%.

The authors in [27] introduced a new technique to classify the sentiment of tweets as positive or negative. They presented and discussed the results of machine learning algorithms for twitter sentiment analysis by using distant supervision. Training data, the authors used consisted of tweets with emotions which were used as noisy labels. According to authors, the machine learning algorithms such as Naive Bayes, Maximum Entropy and SVM when trained with emotion tweets can have accuracy more than 80%. The study also highlighted the steps used in preprocessing stage of classification for high accuracy. In [28] sentiment analysis perform using SVM in that two pre classified datasets of tweets are used then do comparative analysis, they use measures Precision, Recall and F-Measure.

3. TOOLS AND TECHNOLOGY

In proposed model coding is on python for we have to install python, anaconda. Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Python supports modules and packages, which encourages program modularity and code reuse. Anaconda is a freemium open source distribution of the Python and R programming languages for large-scale data processing, predictive analytics, and scientific computing, that aims to simplify package management and deployment. In anaconda we use jupyter notebook. The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more.

4. PROPOSED MODEL

Description about model which we are proposed as given below.

The model having following steps:

- Data collection of tweets
- Pre-process tweet
- Feature Extraction
- Trend Detection

Calculate mood Tendency (Positive, Negative, and Neutral).Following figure shows proposed model:-

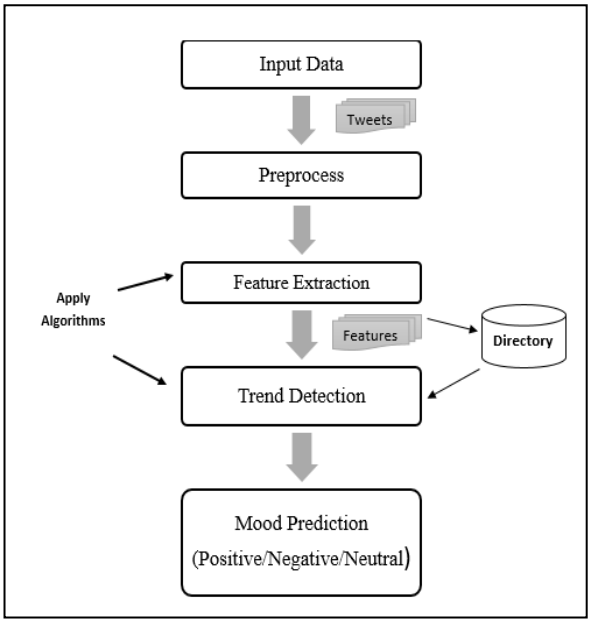


Fig 2: Proposed model for Trend detection and polarity detection

1) Dataset:

Collect tweet data through twitter streaming API. Which download tweets in JSON format. We can apply keyword, hashtag, username to download tweets related to them.

2) Pre-processing:

Tweet pre-processing module having several stages. After downloading tweets we have to extract text data form that and discard video, audio, image etc .store English text which is retrieve form tweet. Then remove @, #, url and other punctuation form tweets and apply stop word remove, word tokenize.

3) Feature Extraction:

After pre-processing stage next module is Feature extraction which is done in two way through Term frequency calculation and pos tagging

4) Trend Detection and Mood Prediction

We can determine trend by using TF-IDF calculation. And predict positive, negative, neutral mood tendency by applying machine learning algorithms. Apply sentiment classification.

5. CLASSIFICATION TECHNIQUES

There are different types of classifiers that are generally used for text classification which can be also used for twitter sentiment classification.

A. SVM Classifier [24]
The main goal of Support Vector Machine is maximize margin. SVM separates the tweets using a hyper plane. SVM uses the a discriminative function defined as

$$g(X) = w^T \emptyset(X) + b \quad (1)$$

'X' is the feature vector, 'w' is the weights vector and 'b' is the bias vector. 'w' and 'b' are learned automatically on the training set.

SVM having hard margin and Soft margin. There are linearly separable method and Non-linear separable method. For linearly separable method we have following equation [22]:

$$f(x) = \sum \alpha_i y_i X_i^T X \quad (2)$$

Where α_i is Lagrange multiplier, y_i is class and $x_i$is input. This is Equation for Hard margin and for soft margin we use slack variable.

For non- linearly separable method we use different kernel tricks like linear, polynomial, radial basis function etc.

B. Nave Bayes Classifier [24]

Nave Bayes is probabilistic model [7]. This Classifier makes use of all the features in the feature vector and analyzes them individually as they are equally independent of each other. The conditional probability for Naive Bayes can be defined as

$$P\left(\frac{X}{y_i}\right) = \prod_{i=1}^{m} P\left(\frac{x_i}{y_j}\right) \quad (3)$$

' X' is the feature vector defined as $X = \{x_1, x_2 \ldots x_m\}$and yj is the class label. Here, in our work there are different independent features like emoticons, emotional Keyword, count of positive and negative keywords, and count of positive and negative hash tags which are effectively utilized by Naïve Bayes classifier for classification. Nave Bayes does not consider the relationships between features. So it cannot utilize the relationships between part of speech tag, emotional keyword and negation.

C. Logistic Classifier

Logistic regression [25] is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, etc.).

The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest:

$$logit(p) = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k \quad (4)$$

Where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

odds= p/(1-p) = (Probability of presence of characteristic)/(Probability of absence of characteristic)

And

$$logit(p) = ln\left(\frac{p}{1-p}\right) \quad (5)$$

D. Decision Tree

Decision tree [24] builds classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy). Leaf node (e.g., Play) represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. C4.5 is an algorithm used to generate a decision tree.

E) KNN classifier

K nearest neighbours [24] is a simple algorithm that stores all available cases and predict the numerical target based on a similarity measure (e.g., distance functions).

The algorithm assumes that it is possible to classify documents in the Euclidean space as points. Euclidean distance is the distance between two points in Euclidean space. The distance between two points in the plane with coordinates p=(x, y) and q=(a, b) can be calculated

$$d(p,a) = \sqrt{(x-a)^2 + (y-b)^2} \quad (6)$$

6. IMPLEMENTATION AND RESULTS

Dataset having 1000 tweets after pre-processing we have 988 tweets. Which is used in weka tool. Then apply different classifier which generate results. Results having information retrieval measure like Precision, Recall, F-measure, accuracy, Root mean squared error etc.

Results are shown as below:

Correctly classified vs. incorrectly classified instances: Out of 988 instances different classier classify instance in different parts.

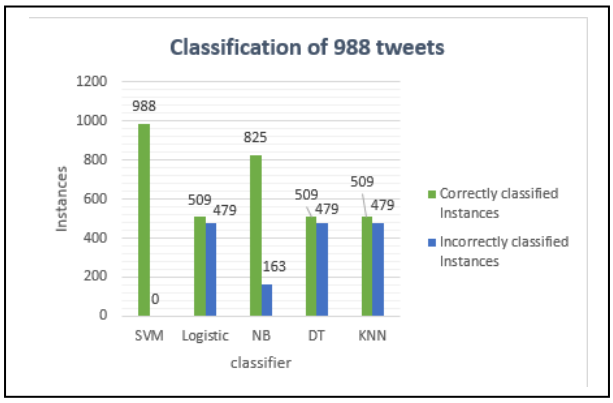
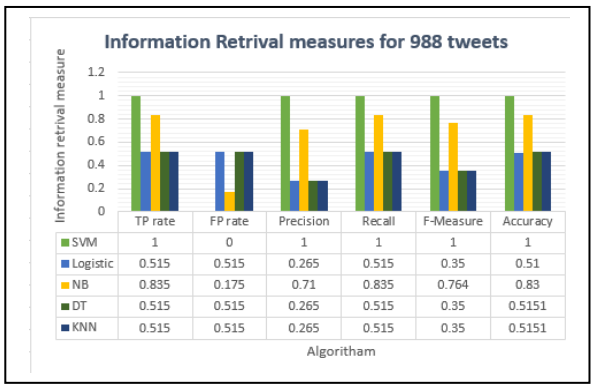


Fig 4: correctly classified vs. incorrectly classified instances line chart

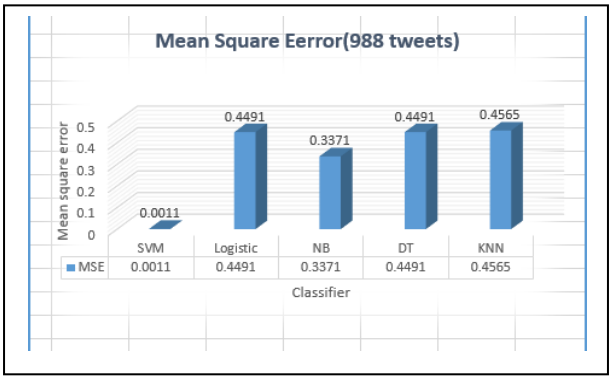
Information retrieval measure: This field having different measures like precision, recall, F-measure, accuracy we compare them and analysis their results based on the graph which are shown as below:



	TP rate	FP rate	Precision	Recall	F-Measure	Accuracy
SVM	1	0	1	1	1	1
Logistic	0.515	0.515	0.265	0.515	0.35	0.51
NB	0.835	0.175	0.71	0.835	0.764	0.83
DT	0.515	0.515	0.265	0.515	0.35	0.5151
KNN	0.515	0.515	0.265	0.515	0.35	0.5151

Fig 5: Information retrieval measures

Mean square error: By observing results of different classifier we can say that SVM having less mean square error compare to other classifier.



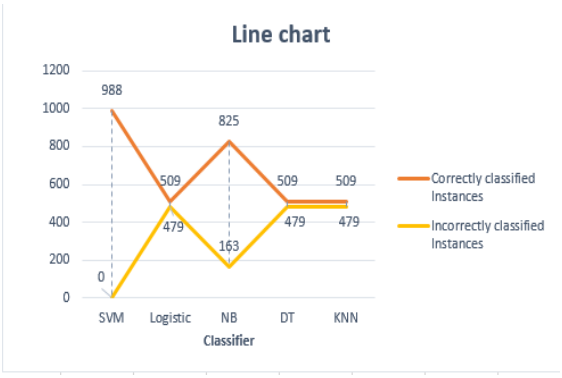


Fig 6: Mean square error

7. CONCLUSION

Tweet having short message we use that for predicting public opinions on sports, Economy, ongoing events etc. We are finding keyword in tweet and predict whether it is having weightage positive or negative by applying machine leaning algorithms. We can apply multi classification algorithms like SVM, Naïve Bayes, Logistic classification, KNN and Decision tree. We observe that Information retrieval measures like precision, recall and F-measure. We get results so by observing the results we can say SVM having less mean square error so it is good classifier for this type of dataset. In future we can test this with python coding and find best classifier.

8. REFERENCES

[1] Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Göker, Ioannis Kompatsiaris, Senior Member "Sensing Trending Topics in Twitter" IEEE, and Alejandro Jaimes IEEE Transactions On Multimedia, Vol. 15, No. 6, October 2013.

[2] Soyeon Caren Han, Hyunsuk Chung, Do Hyeong Kim, Sungyoung Lee, and Byeong Ho Kang "Twitter Trending Topics Meaning Disambiguation" Springer International Publishing Switzerland 2014.

[3] Arkaitz Zubiaga, Damiano Spina, Raquel Mart´ınez, V´ıctor Fresno "Real-Time Classification of Twitter Trends" Journal of the American Society for Information Science and Technology copyright @ 2013.

[4] Altawaier, M. M., & Tiun, S. (2016) "Comparison of Machine Learning Approaches on Arabic Twitter Sentiment Analysis" International Journal on Advanced Science, Engineering and Information Technology, 6(6), 1067-1073.

[5] Rong Lu and Qing Yang, "Trend Analysis of News Topics on Twitter",International Journal of Machine Learning and Computing Vol. 2, No. 3, June 2012

[6] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, Alok Choudhary, "Twitter Trending Topic Classification" 2011 11th IEEE International Conference on Data Mining.

[7] Erwin B. Setiawan, Dwi H. Widyantoro, Kridanto Surendro, "Feature Expansion using Word Embedding for Tweet Topic Classification" IEEE, 2016.

[8]http://www.socialmediatoday.com/social-networks/heres-why-twitter-so-important-everyone

[9]http://www.newsmedialive.com/wpcontent/uploads/2015/10/TWITTER.jpg

[10] http://www.twitter.com

[11] Yubao Zhang, Student Member, IEEE, Xin Ruan, Student Member, IEEE, Haining Wang, Senior Member, IEEE, Hui Wang, and Su He "Twitter Trends Manipulation: A First Look Inside the Security of Twitter Trending" IEEE transactions on information forensics and security, vol. 12, no. 1, january 2017.

[12] Amina Madani, Omar Boussaid, Djamel Eddine Zegour "Real-time trending topics detection and description from Twitter content" Springer-2015.

[13] Arkaitz Zubiaga, Damiano Spina, Raquel Martinez, Victor Fresno, "Real-Time Classification of Twitter Trends" American Society for Information Science and Technology 2013.

[14] Arkaitz Zubiaga, Damiano Spina, Víctor Fresno, Raquel Martínez "Classifying Trending Topics: A Typology of Conversation Triggers on Twitter" ACM 2011.

[15] María del Pilar Salas-Zárate, José Medina-Moreira, Paul Javier Álvarez-Sagubay "Sentiment Analysis and Trend Detection in Twitter" Springer 2011.

[16]https://statinfer.com/204-6-8-svm-advantages-disadvantages-applications/?c=361cde8465e4

[17]https://www.slideshare.net/ashrafmath/naive-bayes-15644818

[18]http://www2.cs.man.ac.uk/~raym8/comp37212/main/node264.html

[19] A. Hernandez-Suarez, G. Sanchez-Perez, V. Martinez-Hernandez, H. Perez-Meana,K. Toscano-Medina, M. Nakano and V. Sanchez "Predicting Political Mood Tendencies based on Twitter Data"

[20] http://www.kdnuggets.com/2017/06/which-machine-learning-algorithm.html

[21] Amina Madani,Omar Boussaid, Djamel Eddine Zegour "Real-time trending topics detection and description from Twitter Content" Springer 2015.

[22]https://web.stanford.edu/class/cs276/handouts/lecture14-SVMs.ppt

[23] https://www.dtreg.com/solution/view/29

[24]https://github.com/ctufts/Cheat_Sheets/wiki/Classification-Model-Pros-and-Cons

[25]https://www.quora.com/What-are-applications-of-linear-and-logistic-regression

[26] Zgheib, W. A., & Barbar, A. M. A Study using Support Vector Machines to Classify the Sentiments of Tweets.

[27] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(2009), 12.

[28] Munir Ahmad, Shabib Aftab, Iftikhar Ali "Sentiment Analysis of Tweets using SVM" International Journal of Computer Applications November 2017.