# Twitter Trend Prediction

(Individual Topic Report)

Amrita Kasaundhan

SJSU ID: 013854204
San Jose State University
Computer Engineering
amrita.kasaundhan@sjsu.edu

*Abstract*— **Social media has become integral part of daily lives in form of connecting to people also, to get information and news from all over the world. While the increase in social media activities, number of such platforms has also increased numeriously. One of such platforms, Twitter, which has emerged as prominent in means of connecting with people, friends, following celebrities and public figures, events, groups, along with involving in a conversion with people across the world. One of the many its features, twitter allows users to include popular topics or hashtags in the post to make the post accessible to maximum number of people. Hence people who wish to increase their following or wish to get post accessible by maximum number of people includes those topics or hashtags in their posts. This paper research topics has proposed to discover hashtags/post which are expected to soon be the trend in future.**

***Keywords—Nearest neighbor, Social media trend prediction, clustering***

## I. Introduction

Social networking has also become popular activity all over the world with the increase of internet access. One of many such networking sites, Twitter has also become popular for the reason of connecting with celebrities, politicians, other public figures along with friends and family members. Users of all ages and backgrounds can sign up on Twitter, can speak their mind, share their thoughts, news. People can stay up to date in terms of global information along with share the information.

Twitter enabled rapid topic diffusion, and the common practice of retweeting further draws attention to select ideas. Thus, Twitter has been uniquely able to centralize, motivate and influence massive discussions which is called 'trends' through its constantly evolving trending list of words, phrases, and hashtags (words with the # sign affixed to them). "Trends" keeps updated based on constant real-time tweet activity, users can identify relevant and prominent themes at any given time and follow conversations accordingly. Following trends in turn allows users to both learn about new topics that would be absent from a 'local feed'-based social media application (like Facebook) and also to discover new perspectives from individuals who are from completely different social circles. Hence this way, following trending lists on Twitter enables users to stay informed, relevant, and connected with the broader online community

Twitter serves as key in terms of personal and financial development these days. Eg: like celebrities and corporations alike command massive followings on social media application, and users who wish to break into this world often aim to do the same. Twitter also has become a casual and convenient way of connecting with an incredibly diverse audience.

## II. Ease of Use

In recent years, there has been an explosion in the availability of data — data that demands to be analysed and gives valuable insights. Of many, Twitter is one of the most popular such platforms for social networking which has shown unprecedented and unbeatable growth in the past few years. The twitter data is so enormous that researchers have to delve deeper through the data and extracted extremely crucial and practically relevant data which represents an information. Such large quantities of data bound to have both opportunities and challenges. Enough of data that can reveal the hidden underlying structure in a process of interest by making computations over data at such huge scale is a challenge. Although, advances in distributed computing have made it easier to exploit the structure in large amounts of data to do inference information. Twitter identifies the potential of the buzzing topics and developed a proprietary algorithm that enables them to predict topics that are likely to trend.

The goal of this research paper aims for two purposes : first, it hopes to predict topics and hashtags that will likely become trend before they actually trend, and second, it recommends pre-trending topics to users based on their individual tweet history and preferences. To accomplish this , several machine learning models have been applied to predict emerging trends from a corpus of over a million real-time tweets. After collecting the list of predicted pre-trending topics,it recommends a selection of them to users based on information derived from previous Twitter activity.

## III. Literature Review

One of the most important parts of this project was research papers of highly qualified professors who had previously done extensive research in the field of social network analysis. Many researchers currently are working and have worked previously on a technique of extraction and analysis of huge amount of twitter data trend for trend prediction by doing trend analysis.

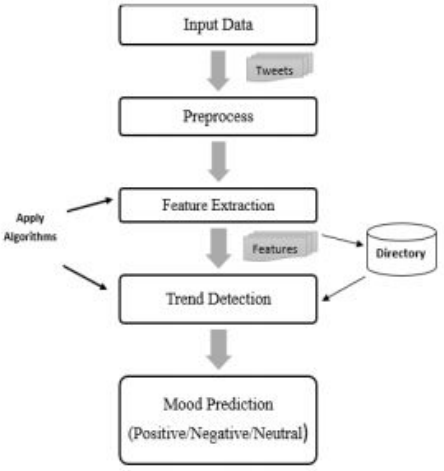
In [2] author applied six trend detection method on the data and compared the results which discovered that standard natural language processing technique perform well for social streams on particular topic.In [5], the author applied various machine learning algorithms like Naïve Bayes, Decision Trees and Support Vector Machine for sentiment analysis on the data which was obtained from twitter. This research has followed tweets classification in which one task was to perform preprocessing of the data, and the other task was to calculate Term Frequency-Inverse Document Frequency (TF-IDF) and stemming of the words. Researchers used the same dataset for three algorithms and performance has been evaluated on the basis of all the different information retrieval metrics precision.. In [6] Towards More Systematic Twitter Analysis: Metrics for tweeting activities- The major focus of this paper was on the power of hashtags. The research describes how hashtags been used as a means for prediction of trends in twitter. In [7] author propose a model which predicts public opinion on political event by applying different classifier which predict whether mood is positive or negative. In [8], the researchers propose a way to get the pre labeled data from twitter which can be used to train the SVM classifier model. The twitter hashtags were used to judge the polarity of tweet. A study on the these classifier was also conducted which showed the result with an accuracy of more than 85%. The authors [9] propose a new technique to classify the sentiment of tweets as positive or negative. The results of machine learning algorithms for twitter sentiment analysis by using distant supervision were represented by the authors. The authors used tweets with emotions which were used as noisy labels. Researchers claims that the machine learning algorithms such as Naive Bayes, Maximum Entropy and SVM, when get trained with emotion tweets, can have better accuracy which is more than 80%. The study also highlights preprocessing process of classification for higher accuracy. In [10] sentiment analysis is performed using SVM in which two pre-classified datasets of tweets are used, further, researchers to do comparative analysis on data set, they use measures Precision, Recall and F-Measure.

## IV. Data Collection and Analysis

Twitter is about what's happening in the world, and about context that people around the world talking about right now.Twitter can access via world wide web also via twitter mobile app. Twitter provides data to the public in form of twitter API's to the developers for business and research work. To get Twitter API keys and Access token, I created Twitter Student Developer Account (https://developer.twitter.com/). Python provides a library named Tweepy which was used to extract the data from Twitter API and results were stored in dictionary objects where the name of the tweet used as the keys and tweet count used as the values. These results could also be stored in .txt file if the data huge and if several algorithms is supposed to be applied in that data, I stored the results in a dictionary for the reason of simplicity as this work was limited to individual task. For this work, I collected tweets name along with the tweet volume count in two intervals.

Graphical representation of the model that we proposed for Twitter trend prediction. The model will be having following step:

- Data Collection from tweets
- Pre-Process the collected tweets
- Feature Extraction
- Trend Detection
- Trend Prediction




*A.* Dataset: Tweet data was collected by using Twitter developer API Tweepy. It's function is to downloads tweets in JSON format. Keyword, hashtag, username etc can be used to download tweets related to them .

*B.* Pre-processing: Extracted tweets were pre-processed in several stages. After downloading tweets , extracted text data formed such that it discards video, audio, image etc stores only text which is retrieve form tweets. The next task is to removes stemming words, stop words, word tokens etc.

*C.* Feature Extraction: After preprocessing next module is extract features which is done in two way through Term frequency calculation with texts.

*D.* Trend Detection: We can interfere trend by count the frequency of words from tweets.

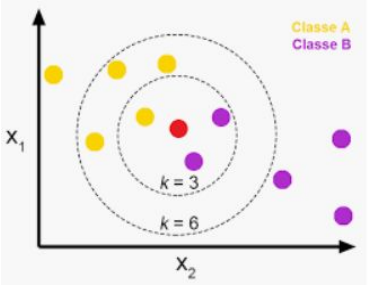
## V. Models

The aim of this research paper is to develop an algorithm that could predict soon to be trending topics or hashtags. The research used the Twitter Api to collect corpus of tweets and finally predicts top 10 tweets which are likely to trend soon. Then it to rank upon compared with actual tweets.

A.HashTag Ranking using KNN:

The model ranked the tweets which are most commonly used within corpus amount of tweets. Trending topics in the form of hashtags took consideration while counting its frequency. Tweets hashtag needed to be collected along with frequencies, we needed to interface with the Twitter Streaming API Tweepy, an easy-to-use wrapper, to do this. Researchers went ahead in this research covering tweets of all languages. Since research aims to collect as many tweets as possible, and set about constructing a query that would capture almost all languages tweets. By collecting counts from the most commonly used words on Twitter (including stop words like 'a', 'the', and 'like'). After collecting the tweets from twitter api, we built a dictionary object for frequency count to track the number of occurrences of each hashtag along with hashtags.Our initial implementations hit a variety of memory errors because of huge dataset which had to be re-write with optimization.in this research corpus tweets was extracted for training the model.

K-nearest neighbors algorithm also called KNN is a non-parametric method used for classification and regression. In our research we used it for classification purpose.It takes input which consists of the k(for this research it was k = 10) closest training examples in the feature space. The training samples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm is to store the feature vectors and class labels of the samples. In the classification phase, k(k= 10) is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label(frequency count for this research) which is most frequent among the k training samples nearest to the query point. Euclidean distance measure is used to calculate the distance between two variables. To Run the machine learning algorithm KNN on collected hashtags yielded the following results:




## VI. Conclusion

This method was quite successful at discovering trends which are soon to be trending. This is probably because a relatively large proportion of tweets actually uses hashtags along with its frequency at two given intervals. Probably its one of the reasons was higher bar hashtags to make it onto the list of top trending. Also research includes collecting tweets along with frequency at different times and days to consider if the time plays a role while predicting the soon to be popular hashtags. Surprisingly, results shows that tweets started at morning during weekdays is likely to trend at short interval of time within the same time zone is started, and the tweets started at evening during weekends has higher chance of trend.

.

```
sorting by frequency for iteration 0
Value at index 0  Name Emmy, freq: [196611, 218867], PredictedFreq: 0
Value at index 1  Name #خسوف_النسر, freq: [173039, 223507], PredictedFreq: 0
Value at index 2  Name Eric Garner, freq: [139387], PredictedFreq: 0
Value at index 3  Name #العدوة_السنوية8, freq: [138617], PredictedFreq: 0
Value at index 4  Name Thor 4, freq: [136722, 168567], PredictedFreq: 0
Value at index 5  Name #SeninSanemin, freq: [104187, 111305], PredictedFreq: 0
Value at index 6  Name Kellyanne, freq: [101037, 110564], PredictedFreq: 0
Value at index 7  Name Toffoli, freq: [85021, 99377], PredictedFreq: 0
Value at index 8  Name Taika Waititi, freq: [67701], PredictedFreq: 0
Value at index 9  Name #vonderLeyen, freq: [53689, 56933], PredictedFreq: 0
sorting by frequency for iteration 1
Value at index 0  Name #خسوف_النسر, freq: [173039, 223507], PredictedFreq: 0
Value at index 1  Name Emmy, freq: [196611, 218867], PredictedFreq: 0
Value at index 2  Name Thor 4, freq: [136722, 168567], PredictedFreq: 0
Value at index 3  Name #SeninSanemin, freq: [104187, 111305], PredictedFreq: 0
Value at index 4  Name Kellyanne, freq: [101037, 110564], PredictedFreq: 0
Value at index 5  Name Toffoli, freq: [85021, 99377], PredictedFreq: 0
Value at index 6  Name #vonderLeyen, freq: [53689, 56933], PredictedFreq: 0
Value at index 7  Name Johnny Clegg, freq: [43280, 47757], PredictedFreq: 0
Value at index 8  Name COAF, freq: [34711, 41083], PredictedFreq: 0
Value at index 9  Name Seri, freq: [39055, 41007], PredictedFreq: 0
```

.

## VII. References


[1] http://docs.tweepy.org/en/v3.5.0/

[2] http://www-personal.umich.edu/~qmei/pub/sigir2014-kong.pdf

[3] Inc. Twitter. Company Facts. 2019. url: https://about.twitter.com/company.

[4] Inc. Twitter. Using Hashtags on Twitter. 2019. url: https : / / support . twitter.com/articles/49309

[5] http://www.bharathsrivatsan.com/files/tweets.pdf

[6] https://www.digitalocean.com/community/tutorials/howto-authenticate-a-python-application-with-twitter-usingtweepy-on-ubuntu-14-04

[7] http://news.mit.edu/2012/predicting-twitter-trendingtopics-110

[8] https://link.springer.com/chapter/10.1007/978-3-319-20294-5_49