A Minor Project Report on

SentiMeter: An Android Application for Sentiment Analysis of Twitter Data Using KNN and NBayes Classifiers

Submitted in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Engineering in Software Engineering under

Pokhara University

Submitted by:

Kanchan Singh, 161716

Poshan Pandey, 161724

Priska Budhathoki, 161726

Under the supervision of

Asst.Prof. **Himal Acharya**

Date: 5th Jan 2020

Department of Software Engineering

NEPAL COLLEGE OF INFORMATION TECHNOLOGY

Balkumari, Lalitpur, Nepal.

Abstract

Social media websites have emerged as one of the platforms to raise users' opinions and influence the way any business is commercialized. The opinion of people matters a lot to analyze how the propagation of information impacts the lives in a large-scale network like Twitter. Sentiment analysis of the tweets determines the polarity and inclination of the vast population towards a specific topic, item or entity. These days, the applications of such analysis can be easily observed during public elections, movie promotions, brand endorsements, and many other fields. In this project, we are going to extract live tweets via creating an app using Twitter developer key and we are going to exploit the fast and in-memory computation of Twitter data using classifiers KNN (K-Nearest Neighbors) and NBayes (Naive Bayes) in Java to perform sentiment analysis. The primary aim is to provide a method for analyzing sentiment score in noisy twitter streams from android application. This paper reports on the design of sentiment analysis and extracting a vast number of tweets. Results classify user's perception via tweets into positive and negative. Secondly, we discuss various techniques to carry out a sentiment analysis on twitter data in detail.

Keywords: Java, KNN, NBayes, Twitter, Sentiment Analysis

Contents

List of Tables

List of Figures

1. Introduction

1.1 Domain Introduction

Sentiment analysis is also known as “opinion mining” or “emotion Artificial Intelligence” and alludes to the utilization of natural language processing (NLP), text mining, computational linguistics, and bio measurements to methodically recognize, extricate, evaluate, and examine emotional states and subjective information. Sentiment analysis is generally concerned with the voice in client materials; for example, surveys and reviews on the Web and web-based social networks.

As the internet is growing bigger, its horizons are becoming wider. Social Media and Microblogging platforms like Facebook, Twitter, Tumblr dominate in spreading encapsulated news and trending topics across the globe at a rapid pace. A topic becomes trending if more and more users are contributing their opinion and judgments, thereby making it a valuable source of online perception. These projects generally intended to spread awareness.

Large organizations and firms take advantage of people's feedback to improve their products and services which further help in enhancing marketing strategies. One such example can be leaking the pictures of the upcoming iPhone to create a hype to extract people's emotions and market the product before its release. Thus, there is a huge potential of discovering and analyzing interesting patterns from the infinite social media data for business-driven applications.

1.2 Motivation

We have chosen to work with twitter since we feel it is a better approximation of public sentiment as opposed to conventional internet articles and web blogs. The reason is that the amount of relevant data is much larger for twitter, as compared to traditional blogging sites.

Moreover, the response on twitter is prompter and also more general (since the number of users who tweet is substantially more than those who write web blogs on a daily basis).

Sentiment analysis of public is highly critical in macro-scale socioeconomic phenomena like predicting the stock market rate of a particular firm. This could be done by analyzing overall public sentiment towards that firm with respect to time and using economics tools for finding the correlation between public sentiment and the firm's stock market value. Firms can also estimate how well their product is responding in the market, which areas of the market is it having a favorable response and in which a negative response (since Twitter allows us to download stream of geotagged tweets for particular locations).

If firms can get this information they can analyze the reasons behind the geographically differentiated response, and so they can market their product in a more optimized manner by looking for appropriate solutions like creating suitable market segments. Predicting the results of popular political elections and polls is also an emerging application to sentiment analysis.

1.3 Problem Statement

Twitter has 330 million monthly active users (as of 2019 Q1). Of these, more than 40 percent that is about 134 million people use the service on a daily basis (Twitter, 2019). The average time spent on Twitter clocks in at 3.39 minutes per session (Statista, 2019). That is a whole lot of data. These data can be used to analyze their opinion about any event, movie, product or politics. Such data can be used for better purpose and for better customer experience.

1.4 Project Objectives

The objective of this project is to extract data from twitter and use those data to find the real-time trend and the opinion of the public so that to use them in business objectives, social campaigns, marketing, and other promotional strategies. It can be used during elections, movie premier, promotions, etc. to find the opinions of the audience or public and act accordingly. Our aim is to

provide the people with a means to find the opinion of the public about their product or ideology or principle.

1.5 Project Scope and Limitations

As the available social platforms are shooting up, the information is becoming vast and can be extracted to turn into business objectives, social campaigns, marketing, and other promotional strategies. It can be used during elections, movie premier, promotions, etc. to find the opinions of the audience or public and act accordingly.

The limitations of this application are:

- Cannot identify humor and sarcasm.
- Does not consider the context of tweets.
- The current classifier does not consider neutral sentiments.
- For now, this project is limited to the English language.
- Cannot handle bigrams.

1.6 Significance of the Study

Sentiment Analysis of Twitter Dataset has a number of applications like promotion, politics, election, etc. Twitter Sentiment Analysis can be used to develop their business strategies, to assess customers' feelings towards products or brand, how people respond to their campaigns or product launches and also why consumers are not buying certain products. In politics, Twitter Sentiment Analysis is used to keep track of political views, to detect consistency and inconsistency between statements and actions at the government level. Sentiment Analysis Dataset is also used for analyzing election results. Twitter Sentiment Analysis is also used for monitoring and analyzing social phenomena for predicting potentially dangerous situations and determining the general mood of the blogosphere.

2. Literature Review

This section summarizes some of the scholarly and research works in the field of Machine Learning and data mining to analyze sentiments on the Twitter and preparing prediction model for various applications. As the available social platforms are shooting up, the information is becoming vast and can be extracted to turn into business objectives, social campaigns, marketing and other promotional strategies as explained in [1]. The benefit of social media to know public opinions and extract their emotions are considered by authors in [2] and explained how twitter gives advantage politically during elections. Further, the concept of the hashtag is used for text classification as it conveys emotion in few words. They suggested how previous research work suffered from lack of training set and misses some features of target data. They opted two-stage approach for their framework- first preparing training data from twitter using mining conveying relevant features and then propounding the Supervised Learning Model to predict the results of elections held in the USA in 2016. After collecting and preprocessing the tweets, training data set was created first by manual labeling of hashtags and forming clusters, next by using online Sentimental Analyzer VADER which outputs the polarity in percentage. This approach reduced the number of tweets or training set and further they applied Support Vector Machine and Naive Bayes classification algorithm to determine the polarity of tweets. Multistage classification approach was used where an entity classifier receives a general class of tweets and categorize them with respect to individual candidates for comparison. The metric they used to determine the winner was the “Pvt ratio” which is a Positive number of tweets to the total count of tweets for respective candidate.

Sentiment Analysis by researchers Imran et al. [3] exploited the technology 'Apache Spark' for fast streaming of tweets and presented the approach Stream Sensing to handle real-time data in the unstructured and noisy form. They conducted the approach on twitter data to find some useful and interesting trends which further can be generalized to any real-time text stream. The unsupervised learning approach is used to locate interesting patterns and trends from tweets processed on Apache Spark. Inspired by the

approach described by Zhu et al. [4] and Li et al. [5] for mining data by selecting time window, authors [3] opted for sliding window method for capturing the live streams of tweets. The common approach found in almost all relevant research works constitutes data collection using Twitter API, preprocessing of data, filtering of data then approaches in feature extraction, classification and pattern analysis makes the distinction. Authors used a sliding window of 5 minutes during data collection and further created Term Document Matrix (TDM) for feature extraction. The pattern analysis was carried out by using the score of TF-IDF for finding the most important keywords as explained by Wu et al [5]. The trending topic or hashtag is fed and tweets relevant to it are filtered to form TDM and computing the weights of TF-IDF to find the most important words is the key idea of this sentiment analysis.

3. Methodology

We worked on following methodologies for the application of knowledge, skills, tools, and techniques to a broad range of activities in order to meet the requirements of our project.

3.1 Software Development Life Cycle: Waterfall Model

In "The Waterfall" approach, the whole process of software development is divided into separate phases. In this Waterfall model, typically, the outcome of one phase acts as the input for the next phase sequentially. The above illustration is a representation of the different phases of the Waterfall Model. We will be using the waterfall model approach for the development of our project. It is very simple to understand and use. In a waterfall model, each phase must be completed before the next phase can begin and there is no overlapping in the phases. Waterfall approach was first SDLC Model to be used widely in Software Engineering to ensure the success of the project.

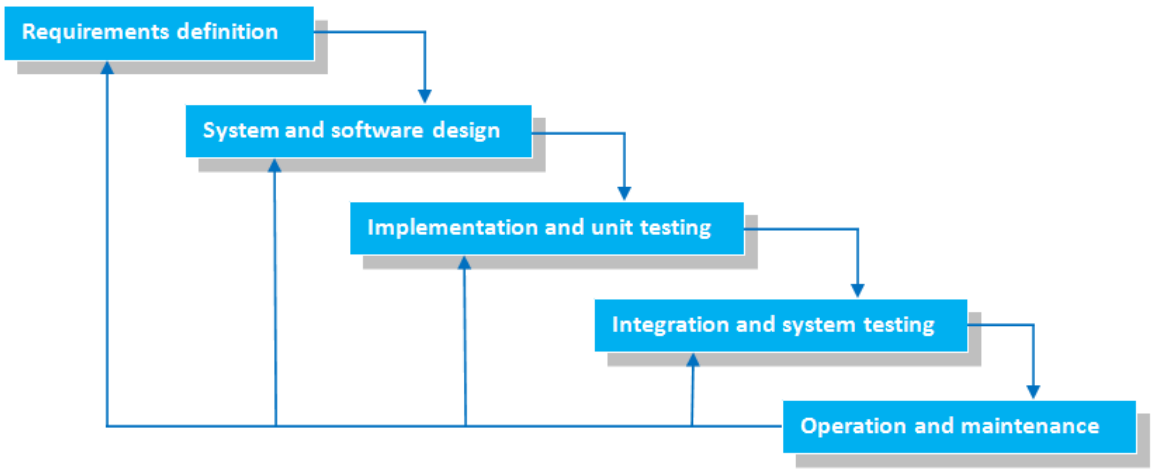

Figure 1: Waterfall Model

The sequential phases in the Waterfall model are:

• **Requirement definition** - All possible requirements of the system to be developed are captured in this phase and documented in a requirement specification document.

• **System and Software Design** - The requirement specifications from the first phase is studied in this phase and the system design is prepared. This system design helps in specifying hardware and system requirements and helps in defining the overall system architecture.

• **Implementation and unit testing** - With inputs from the system design, the system is first developed in small programs called units, which are integrated into the next phase. Each unit is developed and tested for its functionality, which is referred to as Unit Testing.

• **System Testing** - All the units developed in the implementation phase are integrated into a system after testing of each unit. Post integration of the entire system is tested for any faults and failures.

• **Deployment of system** - Once the functional and non-functional testing is done. The product is deployed in the customer environment or released into the market.

•**Operation and Maintenance** - There are some issues which come up in the client environment. To fix those issues, patches are released. Also, to enhance the product some better versions are released. Maintenance is done to deliver these changes in the customer environment.

3.2 Classifiers Used:

3.2.1 Naïve Bayes (NBayes):

This is a classification method that relies on Bayes' Theorem with strong (naive) independence assumptions between the features. A Naive Bayes classifier expects that the closeness of a specific feature (element) in a class is disconnected to the closeness of some other elements. For instance, an organic fruit might be considered to be an apple if its color is red, its shape is round and it measures approximately three inches in breadth. Regardless of whether these features are dependent upon one another or upon the presence of other features, a Naïve Bayes classifier would consider these properties independent due to the likelihood that this natural fruit is an apple. Alongside effortlessness, the Naive Bayes is

known to out-perform even exceedingly modern order strategies. The Bayes hypothesis is a method of computing for distinguishing likelihood P(a|b) from P(a), P(b) and P(b|a) as follows:

$$P(a/b) = [P(b/a) * P(a)]/P(b)$$

Where P(a/b) is the posterior probability of class given as given predictor b and P(b/a) is the likelihood that is the probability of predictor b given the class a. The prior probability of given class a is denoted by p(a) and that of predictor b is P(b). The Naive Bayes is widely used in the task of classifying texts into multiple classes and was recently utilized for sentiment analysis classification.

3.2.2 K-Nearest Neighbors (KNN):

KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. To evaluate any technique, we generally look at 3 important aspects:

• Ease to interpret the output

• Calculation time

• Predictive Power

Let us take a few examples to place KNN in the scale:

KNN algorithm fairs across all parameters of considerations. It is commonly used for its ease of interpretation and low calculation time.

	Logistic Regression	CART	Random Forest	KNN
1. Ease to interpret output	2	3	1	3
2. Calculation time	3	2	1	3
3. Predictive Power	2	2	3	2

Table 1: KNN & other classifiers comparison

Let’s take a simple case to understand this algorithm. Following is a spread of circles (RC) and squares (GS):

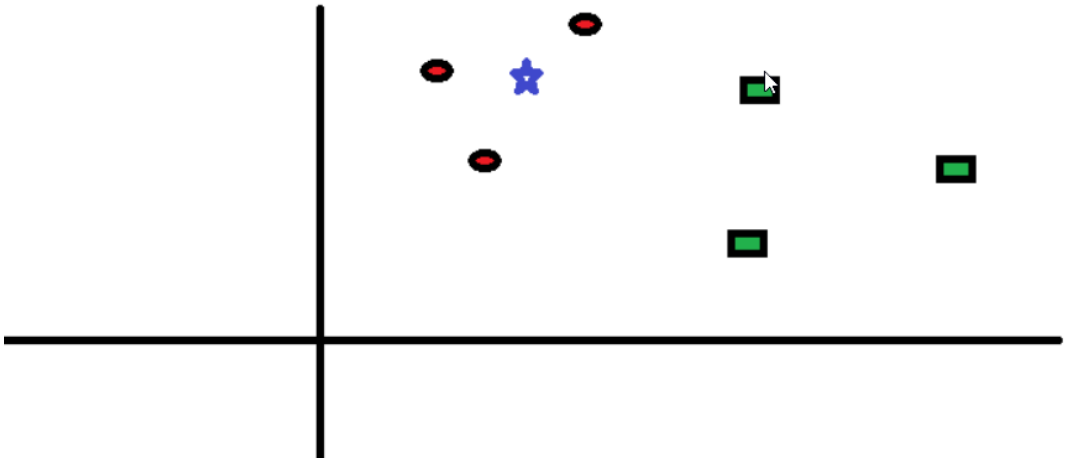

Figure 2.1: How KNN works?

You intend to find out the class of the star (BS). BS can either be RC or GS and nothing else. The “K” is the KNN algorithm is the nearest neighbors we wish to take a vote from. Let’s say K = 3. Hence, we will now make a circle with BS as center just as big as to enclose only three data points on the plane. Refer to the following diagram for more details:

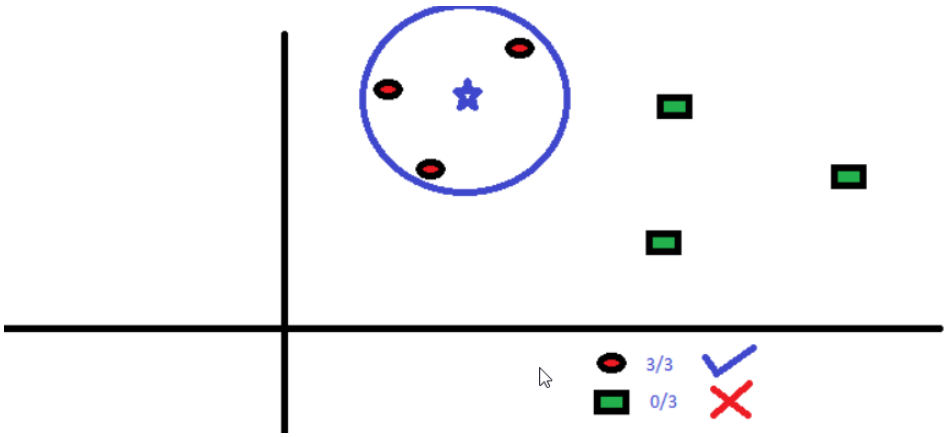

Figure 2.2: How KNN works 2? [6]

The three closest points to BS are all RC. Hence, with a good confidence level, we can say that the BS should belong to the class RC. Here, the choice became very obvious as all

three votes from the closest neighbor went to RC. The choice of the parameter K is very crucial in this algorithm.

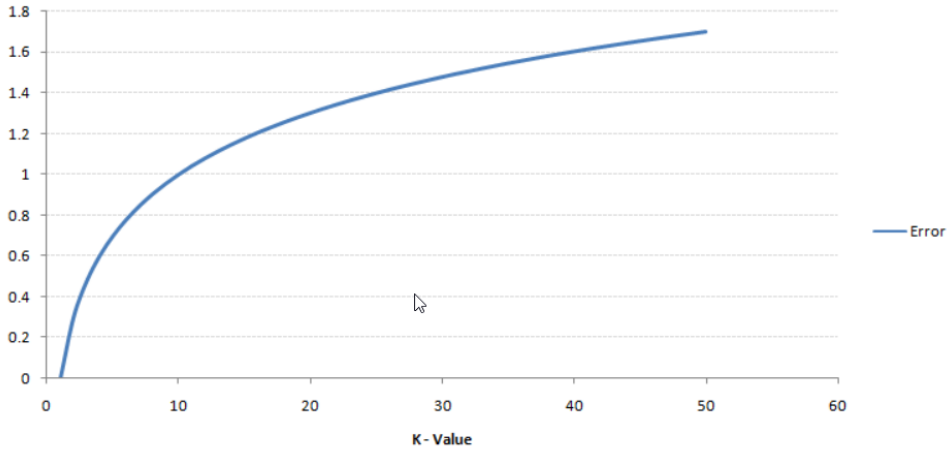

Figure 3: Relation of value of K with error

4. Technical Description

This application is implemented in android using Java programming language. We will use twitter's developer's API to extract data from the twitter and we will use Naïve Bayes and KNN classifiers approach in java for the sentiment analysis.

4.1 Technologies Used

S. N	Technology	Purpose
1	Java	For android development and algorithm implementation.
2	Twitter Developer API	For Twitter data extraction
3	XML	For android UI development.

Table 2: Technologies Used

4.2 Tools Used

S. N	Tools	Purpose
1	Android Emulator	Testing App
2	Android Studio	IDE for Android Development
3	Draw.io	Drawing Charts and tables
4	Figma	Designing UI look
5	GitHub	Managing Team work
6	Google Chrome	For learning and research
7	MS PowerPoint	Making Presentation
8	Microsoft Visio	For Gantt Chart
9	MS Word	For Documentation

Table 3: Tools Used

4.3 Accuracy

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix is used. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. It allows easy identification of confusion between classes e.g. one class is commonly mislabeled as the other. Most performance measures are computed from the confusion matrix.

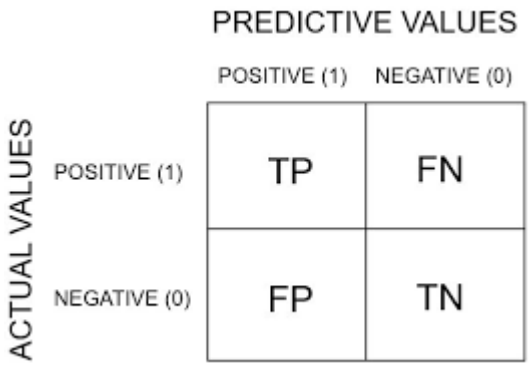

Figure 4: confusion Matrix

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

Figure 5: Confusion Matrix Formula

We used our model to calculate the sentiment of a movies reviews datasets from Kaggle in which there were 50% positive review and 50% negative review [7].

We observed the following result during our analysis:

	Predicted		
Actual		**Positive**	**Negative**
	Positive	**31.667%**	**68.333%**
	Negative	**82%**	**18%**

Table 4: Calculated Result for measuring accuracy via confusion Matrix

By calculating we got 24.834% accuracy.

5. Project Task and Time Schedule

The project is completed a bit late than the proposed time. Since there was Dashain and Exams in between we were unable to complete it by the proposed time. We completed our project by 4th January, 2020. Here is the project tasks and time schedule.

5.1 Project Tasks

S. N	Project Task	Approx. Duration (in Days)
1	Deciding the project	3
2	Analyzing and Research [9] [10]	5
3	Proposal Documentation	7
4	Twitter developer key extraction	4
5	Learn about the classifiers	4
6	Developing basic UI	6
7	Importing the Tweets	3
8	Implementing the tokenization	10
9	Implementing Models	9
10	Testing the application	5
11	Report Documentation	6

Table 5:Project Tasks and approx. duration

5.2 Gantt Chart

ID	Task Name	Start	Finish	Duration
1	Deciding the project	7/25/2019	7/29/2019	3d
2	Analyzing and Research	7/30/2019	8/5/2019	5d
3	Proposal Documentation	8/1/2019	8/9/2019	7d
4	Twitter developer key extraction	8/15/2019	8/20/2019	4d
5	Learn about the classifiers	8/20/2019	8/23/2019	4d
6	Developing basic UI	8/26/2019	9/2/2019	6d
7	Importing the Tweets	9/4/2019	9/6/2019	3d
8	Implementing the tokenization	9/10/2019	9/23/2019	10d
9	Implementing Models	10/29/2019	11/8/2019	9d
10	Testing the application	12/19/2019	12/25/2019	5d
11	Report Documentation	12/26/2019	1/2/2020	6d

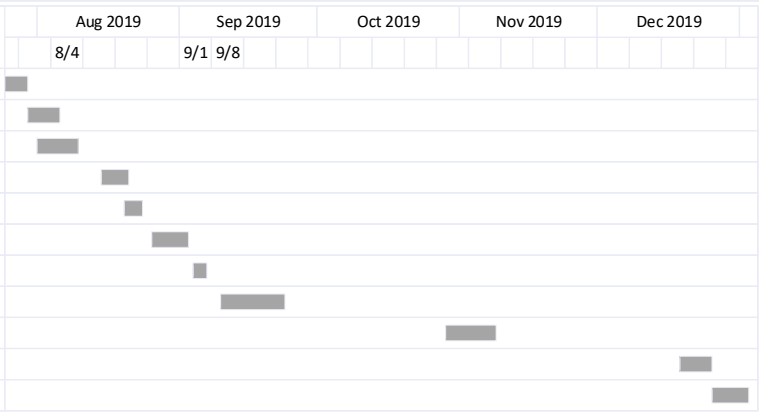

Table 6: Gantt Chart

6. Use Case Diagram

Here is the use case diagram of our application:

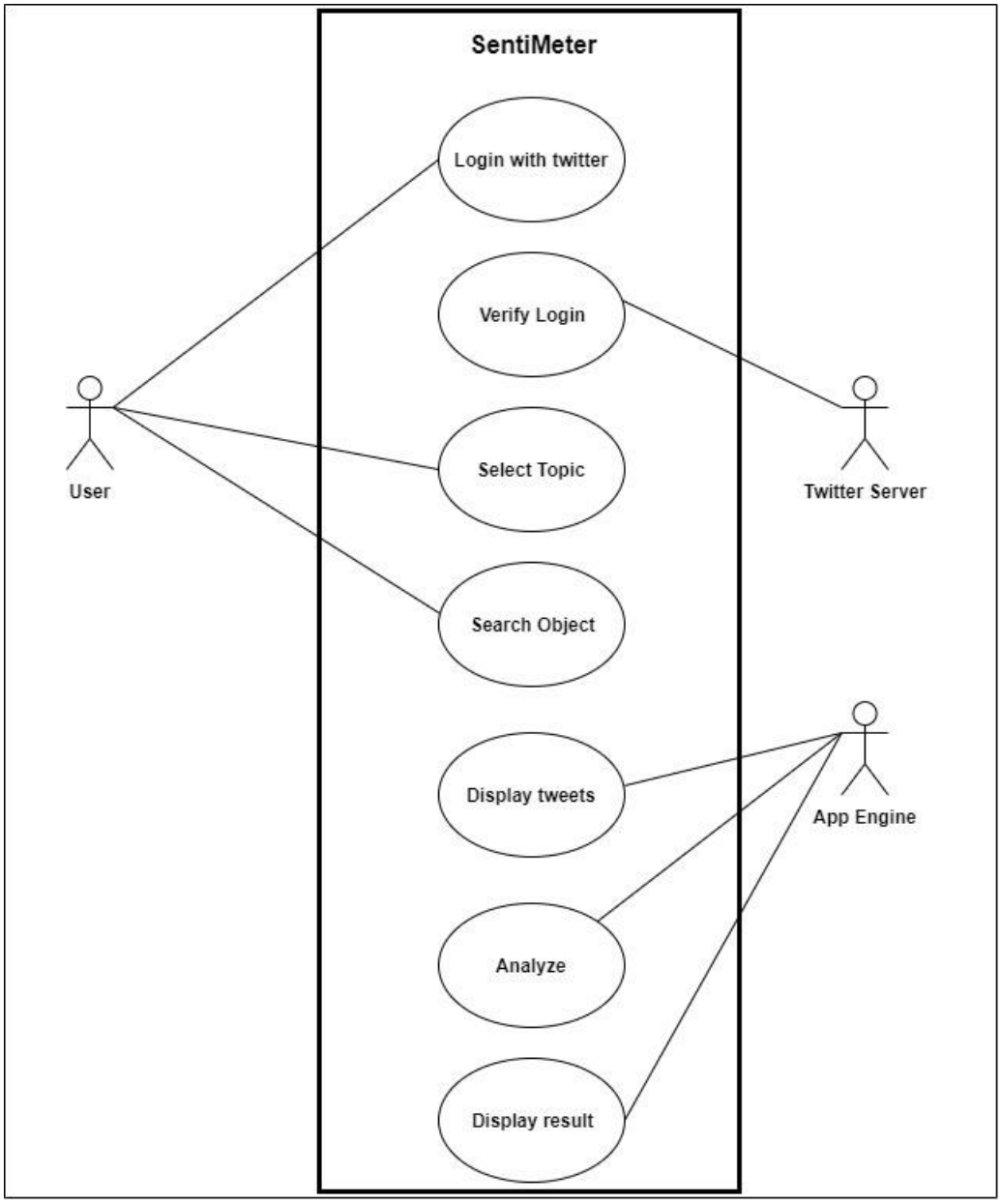

Figure 6: Use Case Diagram

7. Flowchart

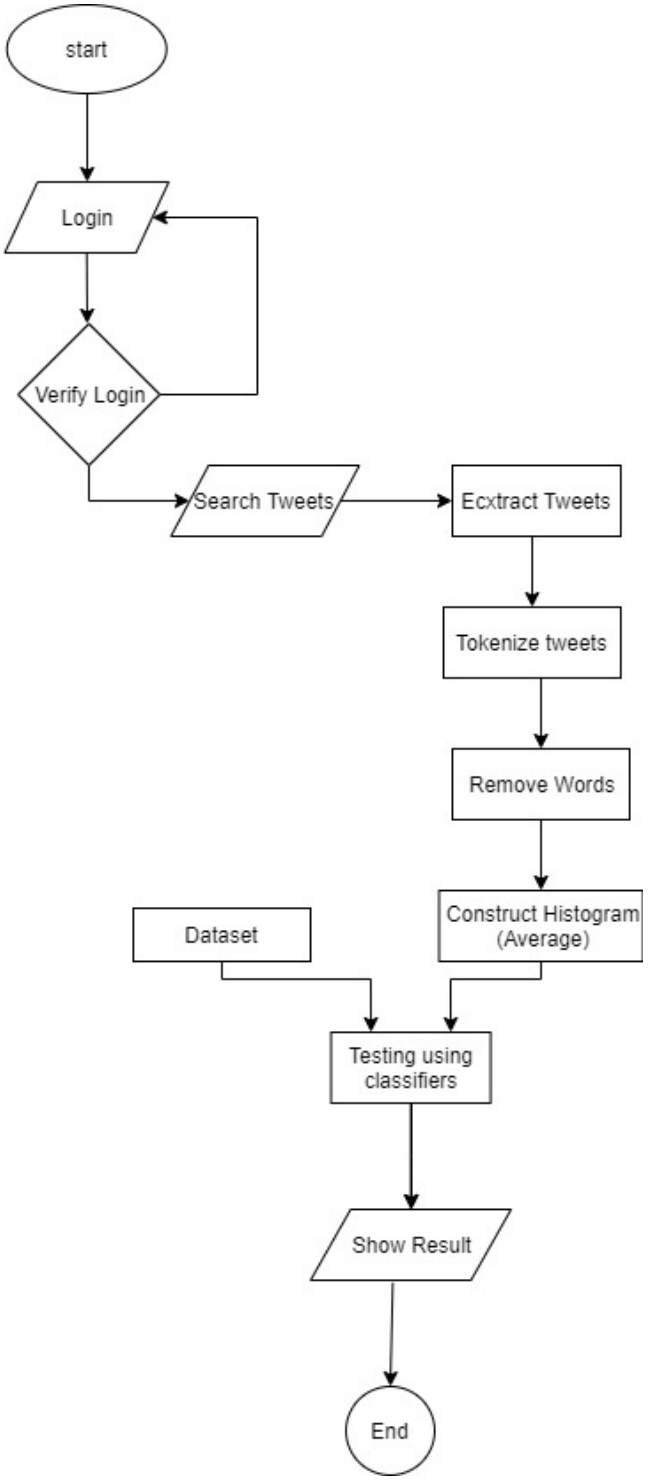

Figure 7: Flowchart

8. Final Outcomes

As proposed, an android application has been developed that analyzes the sentiment of the latest tweets using NBayes and KNN classifiers. The accuracy of the application as observed is 24.834%.

9. References

[1] Mtibaa, M. May, C. Diot and M. Ammar, "PeopleRank: Social Opportunistic Forwarding", 2010 Proceedings IEEE INFOCOM, 2010.

[2] J. Ramteke, S. Shah, D. Godhia, and A. Shaikh, “Election result prediction using Twitter sentiment analysis,” in Inventive Computation Technologies (ICICT), International Conference on, 2016, vol. 1, pp. 1–5.

[3] Dr. Khalid N. Alhayyan & Dr. Imran Ahmad “Discovering and Analyzing Important Real-Time Trends in Noisy Twitter Stream” n.p

[4] Li, H.-F. and Lee, S.-Y. (2009). Mining frequent itemsets over data streams using efficient window sliding techniques. Expert Syst. Appl. 36, 2, 1466–1477.

[5] Alexander Pak and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining". In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10), May 2010.

[6] https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighboursalgorithm clustering/ [Accessed: 9:00 AM 5th August]

[7] https://www.kaggle.com/blanderbuss/positive-and-negative-movies-reviews/[Accessed: 5:35 PM 29th December]

[8] www.wordart.com[For the word cloud in the application Accessed: 8:00PM 8th August]

[9] https://machinelearningmastery.com/ [For learning purposes Accessed in August]

[10] https://en.wikipedia.org/wiki/ [For Learning purpose Accessed in August]