A Minor Project Proposal 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

Date

11 August, 2019

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 Neighbours) 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: Sentiment Analysis, Twitter, KNN, NBayes, Java.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

1.1 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 more prompt 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.

1.2 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 project 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.

1.3 Problem Statement:

This project aims to extract the features of tweets and analyze the opinion of tweets as positive or negative. It aims to classify the tweets on certain people or objects as positive or negative for a set of latest tweets by people around the globe.

A major benefit of twitter is that we can see the good or bad thing people say about the particular brand or object or personality. Using such data we are performing sentiment analysis to calculate if the majority of tweets are of positive sentiments or a negative one.

1.4 Project Overview:

Our project is designing an android application which performs sentiment analysis of twitter data in real-time. Using the KNN (K-Nearest Neighbours and NBayes(Naive Bayes) Classifiers, we will analyze the tweets and provide the result as, if it is positive or negative. We are going to extract the realtime twitter data by importing twitter SDK using twitter developer keys.

This document provides the scope and context of the project to be undertaken. It also provides a schedule for the completion of the project, including a list of all the deliverables and presentations required.

At first, the user has to login to this application via their twitter account then, this application extracts the latest tweets about some specific topic chosen by the user and performs sentiment analysis on those group of tweets and display the result as, how much positive they are.

Then the user can again select different topics and search for people or objects as per their choice to perform sentiment analysis on the tweets where that object or people are mentioned.

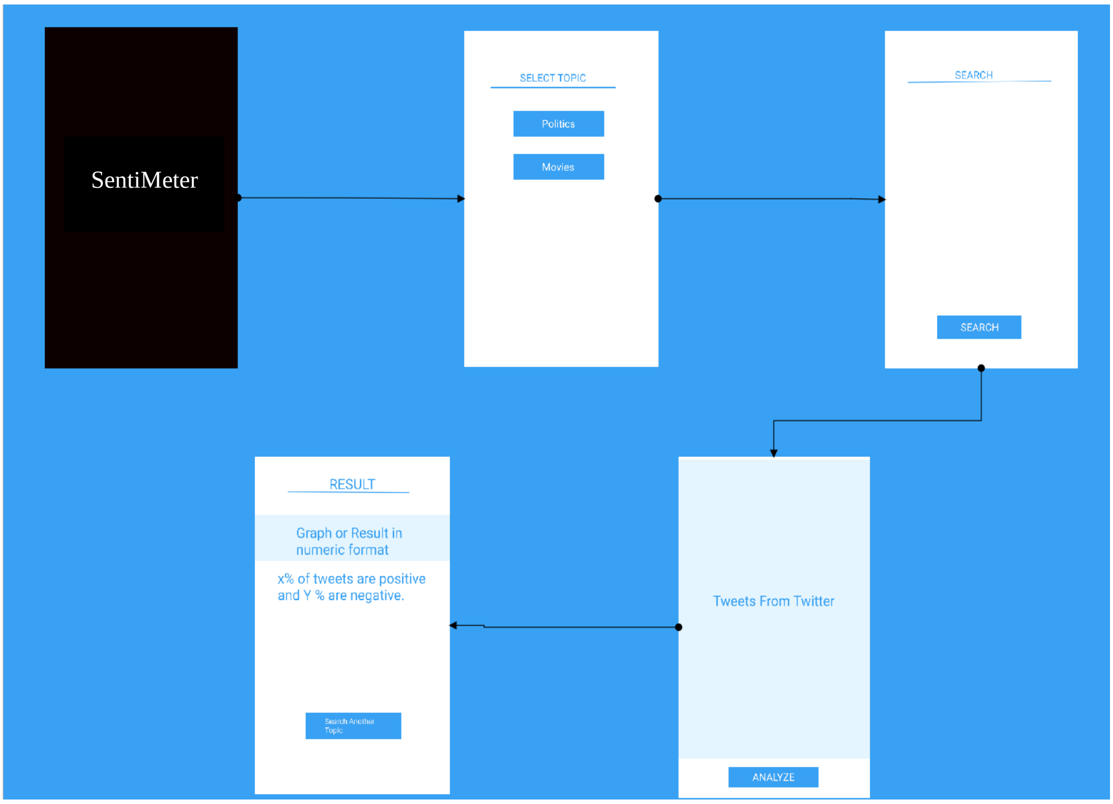


Figure 1: Overview of Project UI(Designed in Figma)

Here is the use case diagram for our project.

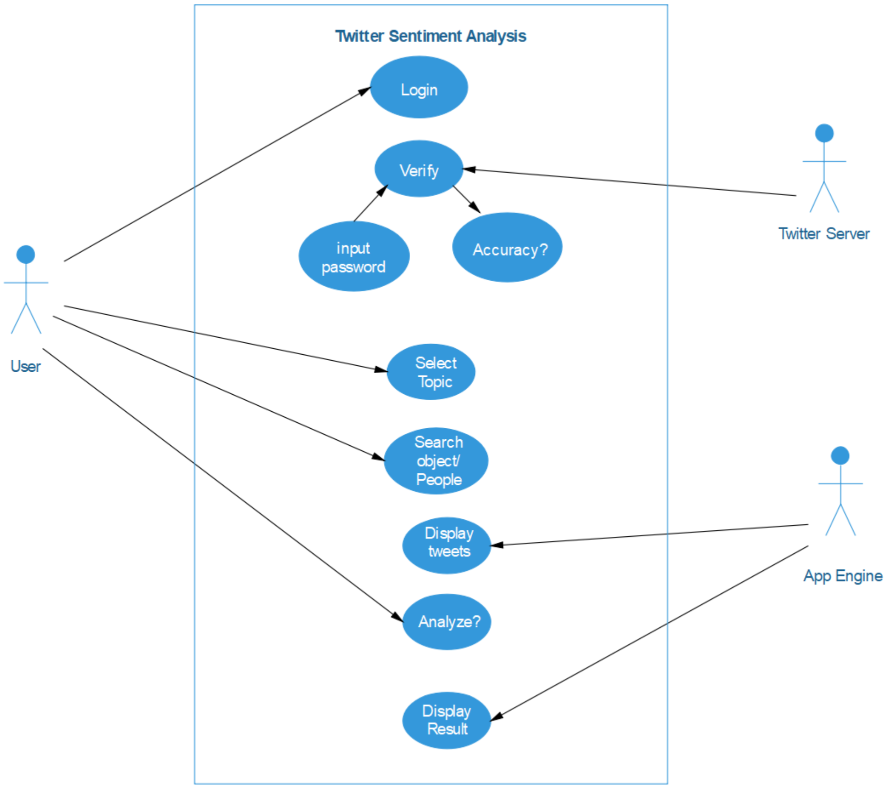


Figure 2 Use Case Diagram of Twitter Sentiment Analysis Project

1.5 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 take action 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.

1.6 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 take action accordingly.

The limitation of this application are:

- Cannot identify humor and sarcasm, So sometime might be wrong.
- Since we use KNN we cannot assure 100% accuracy but the result is closest to human thoughts.
- For now, this project is limited to the English language.

1.7 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 StreamSensing 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. Parallel computation of TDM, TF-IDF score and determining top 5 keywords generated from TDM in each minute as the sliding window moves are one of the highlighting features of this research work. Thus, it leverages the fast computation power of Apache Spark.

3 Methodology

We have planned to work 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

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. 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.

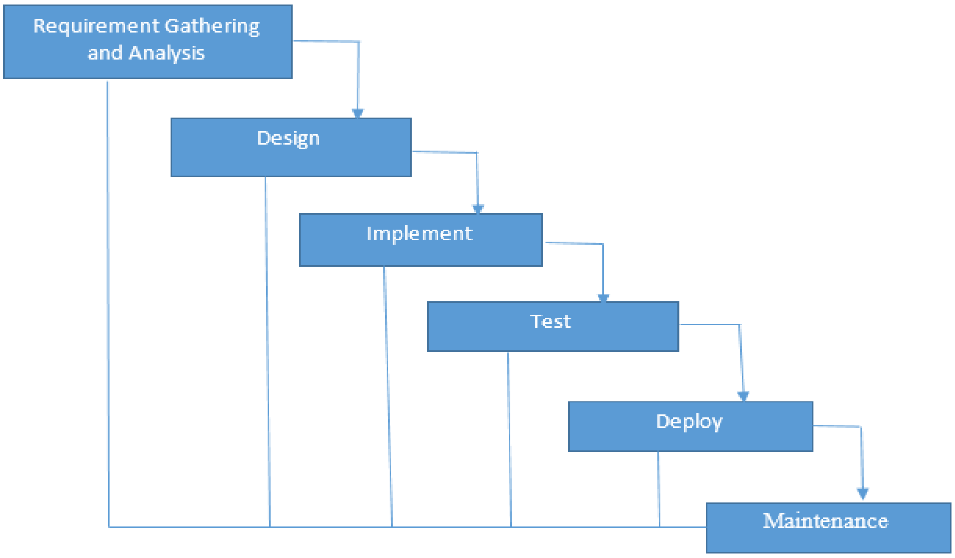


Figure 3: Waterfall Model

The sequential phases in the Waterfall model are −

- **Requirement Gathering and Analysis** − 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 are 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.
- **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.

4 Technical Description

This application will be implemented in android using Java programming language. We will use twitter's developers 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 to be used:

- Java for Android development
- Twitter Developer API to extract twitter data
- Java for KNN and NBayes
- XML for Android UI

4.2 Tools to be used:

The tools used for documentation, designing and developing UI/UX, testing are listed below table:

TOOLS	PURPOSE
Figma	Designing UI/UX
Edraw	Design Charts
Github	Control versioning and managing teamwork
Android Emulator	Testing app real-time
Intellij	IDE to develop the app
Microsoft Word	For Documentation
Google Chrome	For learning and research

Table 1: Tools to be used

4.3 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.

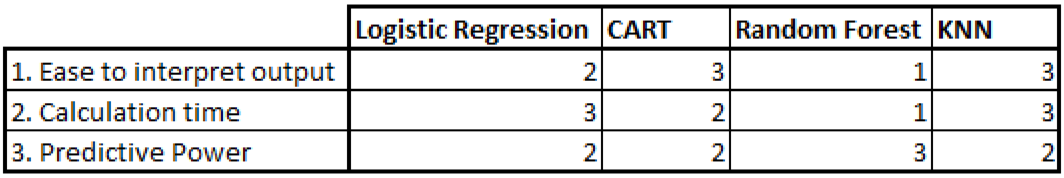
4.4 K-Nearest Neighbours(KNN):

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 2: KNN and other classifier comparisons

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

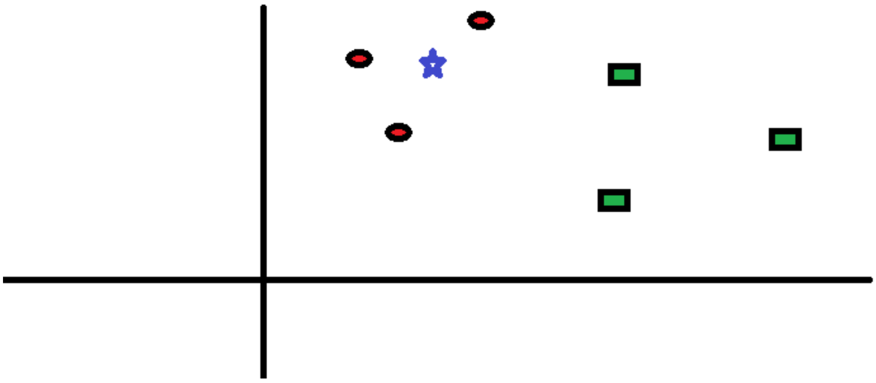


Figure 4: How KNN Works[6]

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:

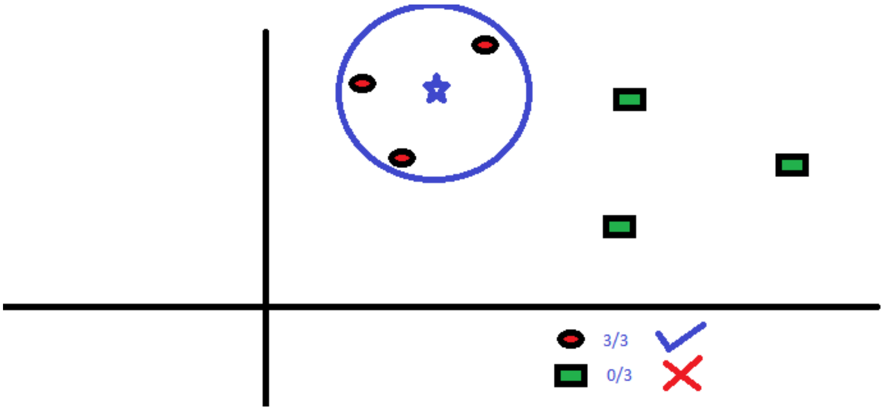


Figure 5: How KNN Works?

The three closest points to BS is 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. Next, we will understand what are the factors to be considered to conclude the best K

5 Proposed Results

Our project at its final phase will able to provide users with an android application that is able to perform sentiment analysis with real-time twitter data. The end product will have the following end results:

- Users log in with their twitter account
- Users can search a particular topic according to their interest
- User can see all the latest tweets related to that particular topic.
- User can retweet tweets from that list.
- User can analyze all those latest tweets and find out if they are positive or negative.

6 Project Tasks and Time Schedule

The project tasks involve idea selection, research, and analysis, UI/UX designing, Proposal Documentation, Android design, twitter data extraction, applying sentiment analysis, product testing and final documentation.

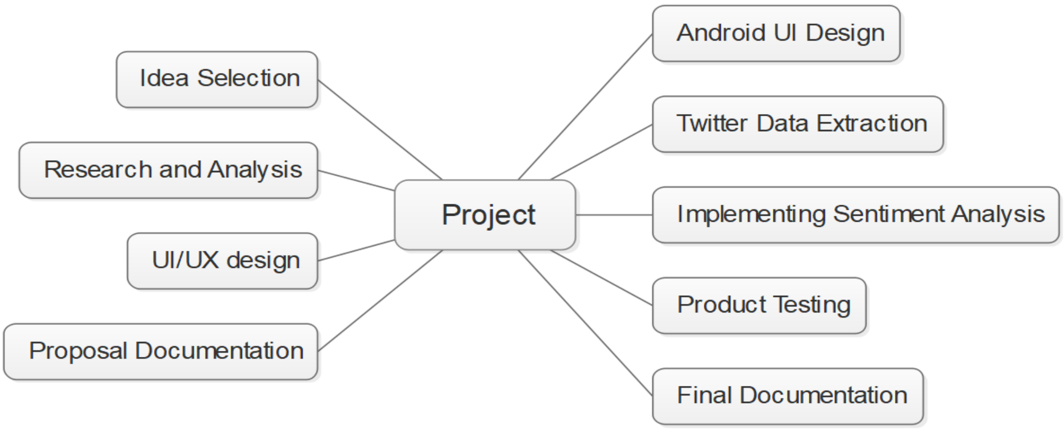
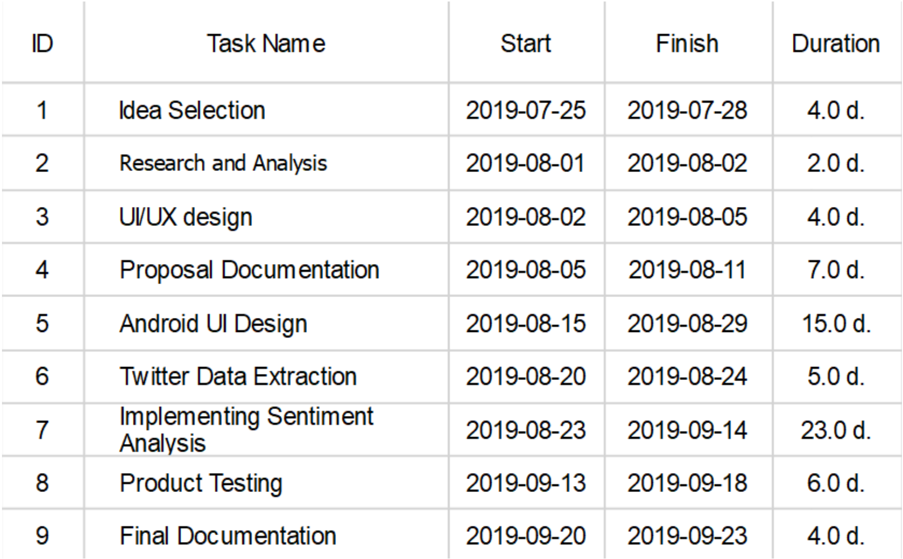


Figure 6: Project Tasks

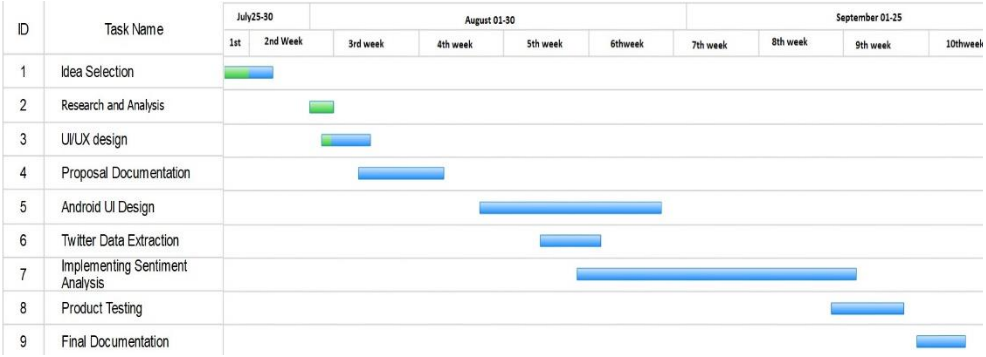
The project schedule has been designed as per the requirements and constraints involved. This project is scheduled to be completed in about 2 months. We **put** an emphasis on *requirement analysis* and *Testing*. Documentation is evidence of good project management.



ID	Task Name	Start	Finish	Duration
1	Idea Selection	2019-07-25	2019-07-28	4.0 d.
2	Research and Analysis	2019-08-01	2019-08-02	2.0 d.
3	UI/UX design	2019-08-02	2019-08-05	4.0 d.
4	Proposal Documentation	2019-08-05	2019-08-11	7.0 d.
5	Android UI Design	2019-08-15	2019-08-29	15.0 d.
6	Twitter Data Extraction	2019-08-20	2019-08-24	5.0 d.
7	Implementing Sentiment Analysis	2019-08-23	2019-09-14	23.0 d.
8	Product Testing	2019-09-13	2019-09-18	6.0 d.
9	Final Documentation	2019-09-20	2019-09-23	4.0 d.

Table 3: Project Task and Schedule

Here is the Gantt chart for our project. We will be completing our project by September 25



ID	Task Name
1	Idea Selection
2	Research and Analysis
3	UI/UX design
4	Proposal Documentation
5	Android UI Design
6	Twitter Data Extraction
7	Implementing Sentiment Analysis
8	Product Testing
9	Final Documentation

Table 4: Gantt Chart

7 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-neighbours-algorithm-clustering/ [Accessed: 9:00 AM 5th August]