Chapter 4.

4.1 System/Software Development

The scrum methodology was undertaken with regard to the following stages

4.1.1 Project Planning

The first step in the scrum methodology is the Planning phases. The project develops a sentiment analysis system that analyzes twitter sentiments. The Planning phase comprises of the following:

4.1.1.1 **Project scope**

The project implements social Media monitory in Kenya, a case study of Safaricom Company, being one of the leading telecommunication industries, it attracts huge following in the social media and especially twitter.

4.1.1.2 **Plan and Estimates**

The plan and estimates of the project are described in the project charter as below.

Table 1: Project Charter1

Budget Information: The project estimates Ksh. 461, 000 for this project. The majority of costs for this project will be internal labor. An initial estimate provides a total of 30 hours per week.
Project Manager: Collins Bunde 0713175471. collinsbunde@gmail.com.
Project Objectives: Develop a system for social media surveillance and monitory of mentions and posts on social media regarding corporate brands and classifies them as either positive or negative. Subscriber firms can, therefore, arrest the probability of negative remarks, or potential release of confidential info from going viral.
Approach: **Identify the existing threats of uncensored social media use in the enterprises** • **Identify how the lack of control has impacted different companies in Kenya.** • **Do a comprehensive literature review to** • **Design the system use cases and U.M.L diagrams.** • **Use twitter API to analyze a stream of tweets and prove the existing online brand damage.**

- **Suggest the uptake of the system methodology in implementing social media monitory to curb insider threat of narcissism and malevolence towards their own enterprises.**

Roles and Responsibilities:

Name	*Role*	*Responsibility*	*Position*	***Contact Information***
Dr. Samuel Liyala	Project Supervisor	Guiding and stewarding.	Lecturer	
Collis Bunde	Project Manager	Managing the project	Student	
Edward Kizito	**Project Assistant**	**Assisting in the development of the project**	**Student**	

Table 2: Project Charter 2

Project Title: Uncensored Social Media Utility

Date: October 18th Prepared by Collins Bunde, Project Manager, collinsbunde@gmail.com.

Project Justification: This project is necessary because it will help enterprises in mitigating social media risks that arise from data leakages, cyber bullying, and cybersquatting and social engineering effects of social media and maintain their privacy against the malicious online users. In addition, these institutions will be able to account for the existing employee productivity hours by putting up measures to prevent cyberloafing during working hours.

The budget for the project is ksh. 461,000. An additional Ksh. 140,000 will be required for operational expenses after the project is completed. Estimated benefits will be tremendous for enterprises. It is important to focus on the system paying for itself within two years.

Product Characteristics and Requirements:

1. **Machine Learning and AI techniques**
2. **Trends of unmonitored Social media use and side effects to the enterprises**
3. **Twitter Social Media platform.**
4. **Sentiment Analysis**

Summary of Project Deliverables

Project management-related deliverables: charter, scope statement, WBS, schedule, cost baseline, status reports, final project presentation, final project report, lessons-learned report, and any other documents required to manage the project.

Product-related deliverables:

1. **System Requirement and Specification document**
2. **Project Plan**
3. **Simulation of a sentiment analysis System as a local enterprise intelligent system.**
4. **Suggest social media policy formulation guideline for the companies.**
5. **Mitigation measures / Controls**

Project Success Criteria: Our goal is to complete this project within 3 months for no more than ksh.461, 000. We must also develop a method for capturing the benefits U.S.M.U implementations and its suggestion for uptake. If the project takes a little longer to complete or costs a little more than planned, the team will still view it as a success if it has a good payback and helps promote social media surveillance in Kenya.

4.1.2 Product Backlog Planning

In this phase we develop of a comprehensive backlog list. The backlog list contains:

- Project initiation
- Requirement analysis and conceptual designs
- Modeling and Implementing the classifier
- Closure

4.1.2.1 Analysis and Requirements

In analysis and requirements, we looked at the following important areas that characterized the functionality of the system

4.1.2.1.1 Machine Learning

Hackeling, (2014) describes, machine learning as the study of software items that makes future decisions from past experiences; it is the study of programs that learn from data. The primary goal of machine learning is to prompt an unknown rule from examples of the rule's application. The undisputed model of machine learning is spam filtering. Spam filters learn to classify new messages, by observing numerous emails previously considered as either spam or ham (Hackeling, 2014).

Why machine learning?

Factor such as growing sizes and varieties of accessible data, and the cheaper computational processing that is more powerful, and affordable data storage are the reasons machine learning is being embraced according to, Hackeling, (2014). It has facilitated the quick and automatic production of models that can analyze bigger, more complex data by delivering faster and accurate results. By building clear-cut models, businesses have a better chance of identifying profitable opportunities besides avoiding unknown risks. Machine Learning is categorized into the following:

1. Supervised Learning - A supervised learning program learns from labeled examples of the outputs that should be produced for an input. They make estimations using data. Our machine learning model for the project is based on supervised learning model.

Figure 3: showing an example of supervised learning for predicting whether an email is a spam

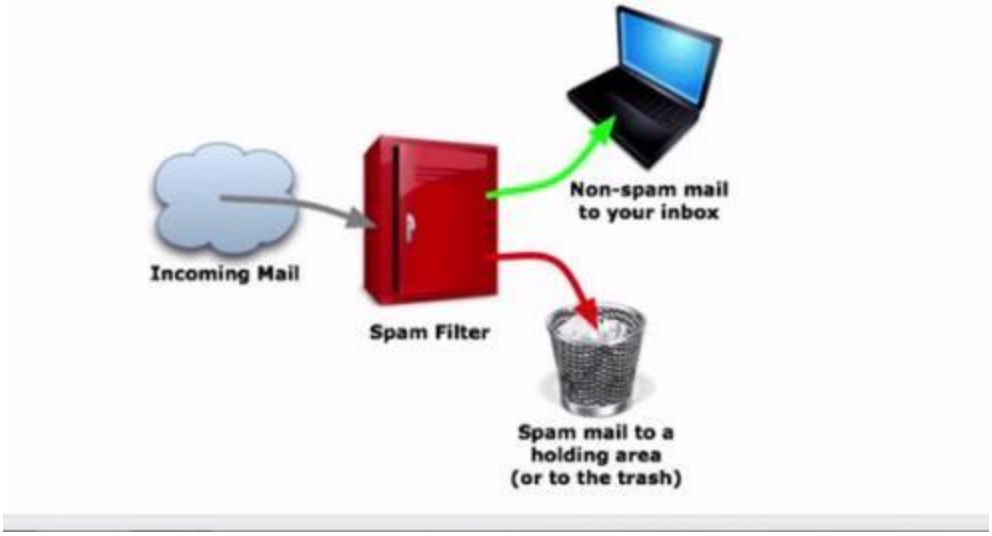

Figure 1 spam filter (supervised learning model)

2. Unsupervised Learning – In unsupervised learning, a program attempts to discover patterns in the data rather learning from labeled data.

The supervised Learning Model

The project used supervised learning, since it was learning from a preset data set of positive and negative text/ messages before being able to classify twitter data.

1. The first step is to train a machine learning model using labeled data. Labeled data is normally labeled with the outcome, the machine learning model then learns the relationship between the attributes of the data and its outcome.

2. The second step is to make predictions on new data for which the label is unknown

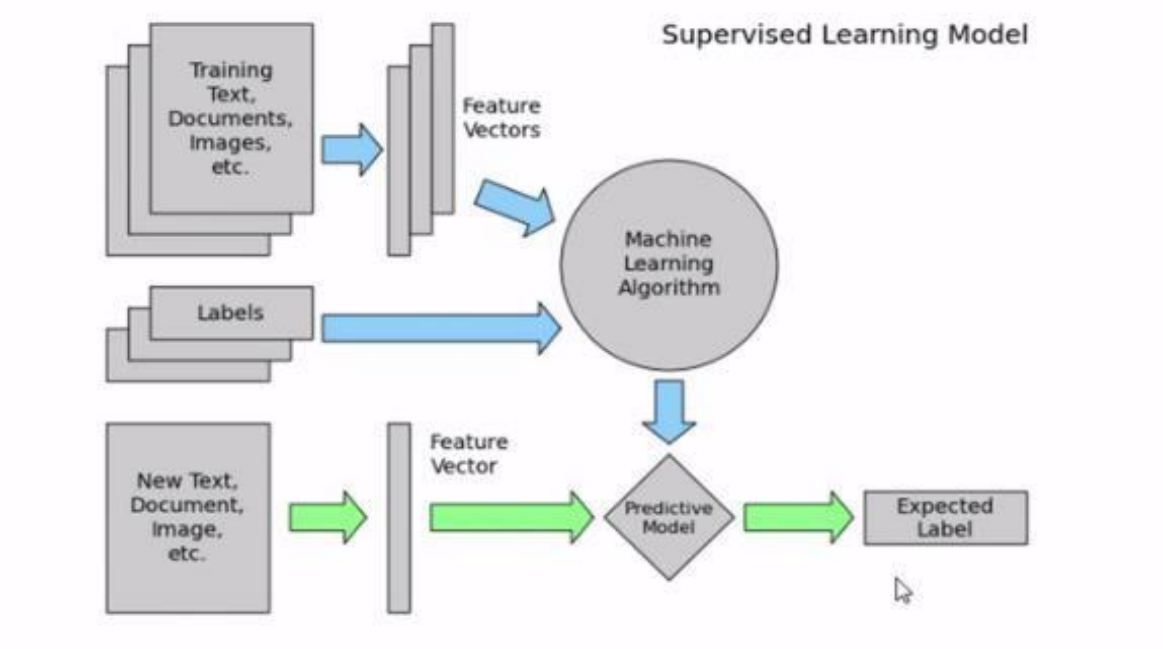

Figure 2: Supervised Learning Model

The primary goal of machine learning is to build a model that generalizes, It should accurately predict the future rather than the present.

Training data and test data

In training, observations comprise the particular capability the algorithm is using to learn. Supervised machine learning complications, is where each observation consists of an observed response variable and one or more observed explanatory variables (Hackeling, 2014). Hackeling, (2014), explains that the test set is an analogous collection of interpretations used to evaluate the performance of the model by the use of some performance metric. Importantly no observations from the training set should be incorporated in the test set. A test set that contains samples from the training set, is difficult to evaluate whether the algorithm has learned to generalize from the training set or has simply memorized it. To effectively perform a task with new data, a program should be able to generalize well. Of great importance to note is that the predictive power of many machine learning algorithms improves as the amount of training data increases (Hackeling, 2014)**.**

Our project uses supervised learning where we obtain a data set for both positive and negative sentiments and train the classifier to the data, after which we test the data set on a sample data before we eventually use it to classify twitter sentiments.

Performance measures, bias, and variance

To measure whether or not a program is learning to perform its duty more efficiently a number of metrics can be used. Many performance metrics measure the number of prediction errors for supervised learning problems. The two fundamental causes of error are bias and variance (Hackeling, 2014). A high biased model produces related errors for an input irrespective of the training set it was trained with; the model biases its personal assumptions about the real relationship over the one demonstrated in the training data (Hackeling, 2014). A high variant model, conversely, produces different errors for an input depending on the training set it was trained with. While a model with high bias is inflexible, a model with high variance can be quite flexible that it models the noise in the training set.

4.1.2.1.2 Natural Language Processing

Conferring with, Steven, Ewan, and Edward, (2009). Python toolkit and a fundamental arsenal for mining the social web. The Natural Language Toolkit (NLTK) is a suite of Python libraries designed to identify and tag parts of speech found in natural English text. Its development began in 2000, and over the past 15 years, dozens of developers around the world have contributed to the project, whose functionality provides tremendous.

NLP is inherently complex and difficult to work with reasonably well, and understanding it for large set of commonly spoken languages is seen as a problem of the century (Bird, S. 2006). The case of the rising interest in understanding of the web with such initiatives such as Google's Knowledge Graph that is being endorsed as "the future of search." Shows the existing interest in Natural language processing. A mastery of NLP is a reasonable strategy for passing the Turing Test, a computer program achieving that level of understanding, will have to demonstrate a weird amount of human intelligence.

How the classifier works

A classifier differentiates good and bad words and it has the following techniques:

Uni-grams

Keep track of consecutive sequences of words, the longer sequences of words are called n-grams.

Stemming

Stemming is getting rid of prefixes and suffixes in a word, it's an algorithm that takes words and strips out its suffixes and prefixes. An example of stemming applies in the following:

1. Watching, watched -> watch
2. Liked, liking -> lik
3. Cats, catlike, catty -> cat

Stop words

They help in preprocessing data by analyzing proper text, stop words are words that are typically pulled out since they don't have much meaning. They are referred to as filler words

Figure 3: Stop words

Figure 5: showing example of stop words

WordNet

The linguistic knowledge that tells one what are adjectives, and word synonyms, nltk enables one to leverage this language.

Part of Speech Tagging

A common preprocessing technique divides words into bigrams, trigrams or unigrams

It decomposes words into verbs adverbs adjectives nouns and pronouns**.** All nltk libraries have this functionality

4.1.2.1.3 Scikit Learn

Hackeling, (2014) introduces Scikit learn as one of the most popular open source machine learning libraries for Python. Providing algorithms for machine learning tasks such as classification, reduction, regression, and dimensionality and clustering. Additionally, it also provides modules for features extraction, models evaluation and data processing. Scikit-learn is popular for academic research due to its well-documented, easy-to-use, and adaptable API and the fact it is built on the popular Python libraries NumPy and Matplotlib (Hackeling, 2014). Developers can use it to experiment with different algorithms by altering only a few lines of the code. It also wraps some popular implementations of machine learning algorithms, such as LIBSVM and LIBLINEAR. Other Python libraries, including NLTK, have wrappers for scikit-learn. It also includes an assortment of datasets, allowing developers to put emphasis on algorithms rather than finding and cleaning data.

4.1.2.1.4 Twitter Platform

Russell, (2013), defines twitter as a real-time, vastly social microblogging facility that lets users post precise status updates (tweets) that appear on timelines. Tweets include one or more entities in their 140 letterings of content and reference, one or more places mapping to locations in the real world (Russell, 2013). An understanding of users, tweets, and timelines is predominantly vital to effective use of Twitter's API. Tweet entities comprise the user mentions, hashtags, URLs, and media be associated with a tweet, while places are locations in the real world. To make it all a bit more concrete, let's consider a sample tweet with the following text:

@KTNNews @Hassanjumaa @SMukangai @abullerahmed . Safaricom wanatuibia

The tweet is 83 characters long and contains two tweet entities: the user mentions @KTNNews @Hassanjumaa @SMukangai and @abullerahmed the text "Safaricom wananiibia." An API is largely abstract in that it specifies an interface and controls the behavior of the objects specified in that interface. The software that provides the functionality described by an API is said to be

“an implementation of the API”. An API is typically defined in terms of the programming language used to build an application (Russel, 2013).

Twitter uses the REST (Representation State Transfer Protocol), which is resource focused, and remote resources can be created, read, updated and deleted (Severance, 2013). The Twitter API requires a key and hence it is quite secure. However, it is no longer free. This API’s generally provide very valuable information. The data providers, limit the number of requests per day, demand an API key or even charge for the use.

4.1.2.1.5 Software requirements

- Python programming Language

Python programming language used because of its intuitive syntax, the amazing ecosystem of packages that trivializes API access and data manipulation, and core data structures that are practically json, make it an excellent tool that’s powerful yet also very easy to get up and running.

- Ipython Notebook

It’s a powerful, interactive Python interpreter that provides a notebook-like user experience from within your web browser and combines code execution, code output, text, mathematical typesetting, plots, and more. It’s difficult to imagine a better user experience for a learning environment because it trivializes the problem of delivering sample code that the reader can follow along with and execute with no hassles.

- Anaconda

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. Its package management system is conda.

4.1.2.1.6 Algorithms used

1. Multinomial Naive Bayes

The multinomial naive Bayes model is typically used for discrete counts. With a text classification problem, it takes the idea of Bernoulli trials one step further and instead of "word

occurs in the document" we have "count how often word occurs in the document", you can think of it as "number of times outcome number x_i is observed over the n trials"

2. Naïve Bayes Algorithm

This Classification is named after Thomas Bayes (1702-1761), who proposed the Bayes

Theorem. Bayesian classification provides practical learning algorithms and prior knowledge (Hassan, S., Rafi, M., & Shaikh, M. S, 2011, December). Some of the uses of the Naïve Bayes Classifier are:

I. Naive Bayes text classification. Naive Bayes classifiers are among the most successful known algorithms for learning to classify text documents.

II. Spam filtering is the best-known use of Naive Bayesian text classification. It makes use of a Naive Bayes classifier to identify spam e-mail. It has become a popular mechanism to distinguish illegitimate spam email from legitimate email

3 Gaussian Naive Bayes Algorithm

Assumes that the features follow a normal distribution. Instead of discrete counts, we have continuous features (e.g., the popular Iris dataset where the features are sepal width, petal width, sepal length, petal length).

4 Support Vector Machines algorithm

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis (Jordan, 2002). Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. SVMs can be used to solve various real-world problems:

• SVMs are helpful in text and hypertext categorization as their application can significantly reduce the need for labeled training instances in the standard inductive settings.

5. Logistic regression

It is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). In this case, the output of the classifier is binary data that is used later on the actual tweets (Jordan, 2002). The logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more metric (interval or ratio scale) independent variables.

We used more than one algorithm since two algorithms are not equivalent and will not necessarily produce the same accuracy given the same data. Since the results for each method/ classifiers are significantly different

4.1.1.3 Project Goals

The project goal is to monitor mentions of company brand names in social media taking a case study of Safaricom. This is with the objective of finding out company and brand reputation damage, and other social media risks that are imminent through Social Media Monitory

4.1.3 Sprints

In this phase we outline the sprints backlogs and we also define the Scrum Team.

4.1.3.1 Sprint backlog

Table 3: Product Backlogs and the sprint backlogs

	Product Backlog	Sprint backlog	
1	Initiate project	Project concept	
		S.Q.C.T targets	
		Formulate W.B.S	
		Project Charter	
2	Requirement analysis and conceptual designs	Requirement Analysis	
		Literature Review	
		Conceptual Framework	
		UML and Use cases	

3	Modelling & Classifier implementation	Obtain a dataset of (+ve) and (–ve) tweets	
		Train classifier	
		Obtain Twitter API	
		Do sentiment analysis on tweets	
		Graph live tweets (+ve/-ve)	
		Send email notification for tweets with high polarity	
		Scrape tweets to a database for analysis	
		Visualize most tweeting accounts in a word cloud	
	Closure	Final Project Presentation	
		Submission of project	

4.1.3.2 Scrum team

Table 3: shows the scrum team

Table 4: Scrum team

Scrum Master	Dr. Liyala	The project supervisor -Leads the team -Facilitates and coordinates -Helps removing the obstacles -Safeguards the process
Project manager(Owner)	Collins Bunde	- Defines initial content and Timing of the release - Manages evolution of project content

		- Deals with Backlog, risk and release content.
Development team	Edward Onyango and Collins Bunde	Edward – Documenter Collins – Lead Developer

4.1.3.3 Feasibility Study

1) Technical – We conducted a technical feasibility to know whether we have all the technical skills required for the project. What we considered are the skills in machine learning and python programming, we had the necessary resources and infrastructure to carry out the project up to the implementation bit.
2) Economic –the project work was facilitated despite, with financial constraints, this entailed using economically the resources available for the implementation of the system.
3) Operational – the operation feasibility of the system, gave consideration to using any enterprise in Kenya as a case study for the project.

Additionally, in the feasibility study we also considered the following:

4.1.3.3.1 Preconditions

Preconditions form the context within which the project must be conducted. This includes the legislation, working condition regulations, and approval requirements. Such kind of requirements are not influenced from within the project. Some of the existing preconditions for this project are:

1. The twitter Streaming API is rate limited to a certain number of requests per day, hence this should be adhered to or else, twitter will shut one (the twitter collector) out.
2. The enterprises, who would be willing to take such a system will have to fit it within their social media policy guideline, measures as to what an employee should or should not post and who are responsible for the social media responsibility in the company should be distinctly defined. Finally, how these persons are held accountable should well defined.
3. There might be privacy issues that arise from monitoring what others are posting hence there is a need for harmonization of this issue in the best way possible. This is because according to the constitution every individual has freedom of expression.

4.1.3.3.2 Design limitations

1. The design limitations of the system are that it only analyzes text messages, hopefully, in future, it can be advanced by doing image analysis of photos and images posted on most of this social media sites.
2. In addition, the implementation is only limited to one social media site, when most enterprises are using more than one social media tool such as Facebook, Instagram, WhatsApp and LinkedIn in their social business strategy.
3. The classifier was modeled with no consideration to the neutral tweets, this should with time be considered for future advancement. Neutral tweets are tweets that are neither positive nor negative.

4.1.4 Release Planning

The release Planning involve the procedure for release of accomplished sprints, once a sprint is accomplished a release planning is done. After training the classifier. A release planning was done whereby we considered using the Classifier for sentiment analysis not limited only to twitter data. Each user story is monitored here, in addition to planning and monitory which was done with the help of the supervisor.

The possible stakeholders of the project are:

1) Companies and Enterprises
2) Twitter Company -It has rate limited use to its API (must agree to terms and conditions)
3) Programmers
4) Project supervisor

This type of project is explicitly conceptualized on the basis of **a proof of concept.**

4.1.5 Design and Development

4.1.3.1 System Requirements

The following are the system requirements for the system:

Table 5: System requirements

	REQUIREMENTS	SPECIFICATIONS
Hardware	**Laptop/Desktop computer**	- 64 bit - 12 GB RAM - Windows 8,10 - Linux
		-
Software	**Python**	Python 3.5 or 2
	Anaconda (for data science)	64 bit version
	Database	MySQL or SQLite
	Mail server	Gmail, Yahoo or any other
	Twitter Account	For scraping and streaming tweets
	Tweepy library	
	Anaconda and Ipython	3.5

4.1.3.2 System Use Cases

During this phase, functional, support and training requirements are translated into preliminary and detailed designs. Decisions are made to address how the system will meet functional requirements. A preliminary (general) system design, emphasizing the functional features of the system, is produced as a high-level guide as the one below.

Figure 7: the functional features of the classifier

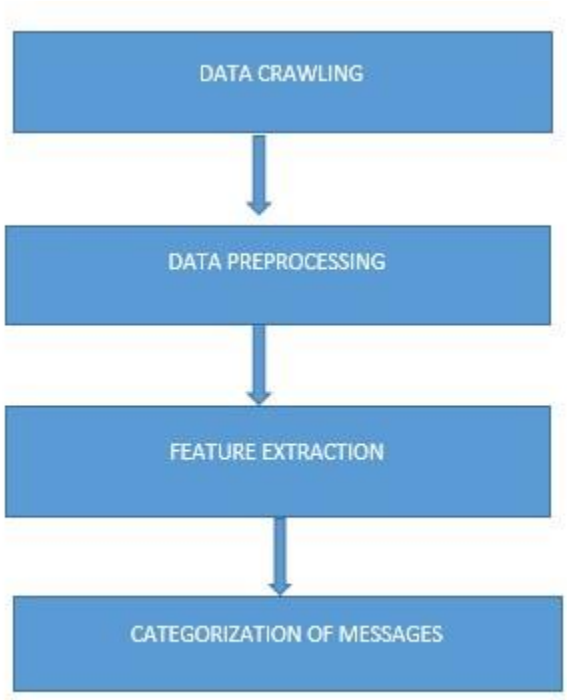

Figure 4: Classifier Functional Features

1. Data crawling- In this module, we captured data from Twitter by using our customized crawler written in Python. Data we got from twitter contain sentiments that are either positive or negative.
2. Data pre-processing - Data captured from twitter contains many missing fields, duplicate tweets. Pre-processing of the dataset involve following steps:
    - Missing fields are replaced by NULL.

- Stemming - The idea of stemming is a sort of normalizing method. Many variations of words carry the same meaning, other than when tense is involved. The reason why we stem is to shorten the lookup and normalize sentences

3. Categorization of Messages

In this module, we used a number of machine learning methods, which needed pre-labelled training data for automatic learning: Naive Bayes classifier, a classifier based on Decision trees, Support Vector Machines (SVM) and Multinomial naïve Bayes classifier.

The following diagrams are used to describe the functionality of the Twitter Sentiment Analysis System.

Figure 8: sentiment analysis use case diagram

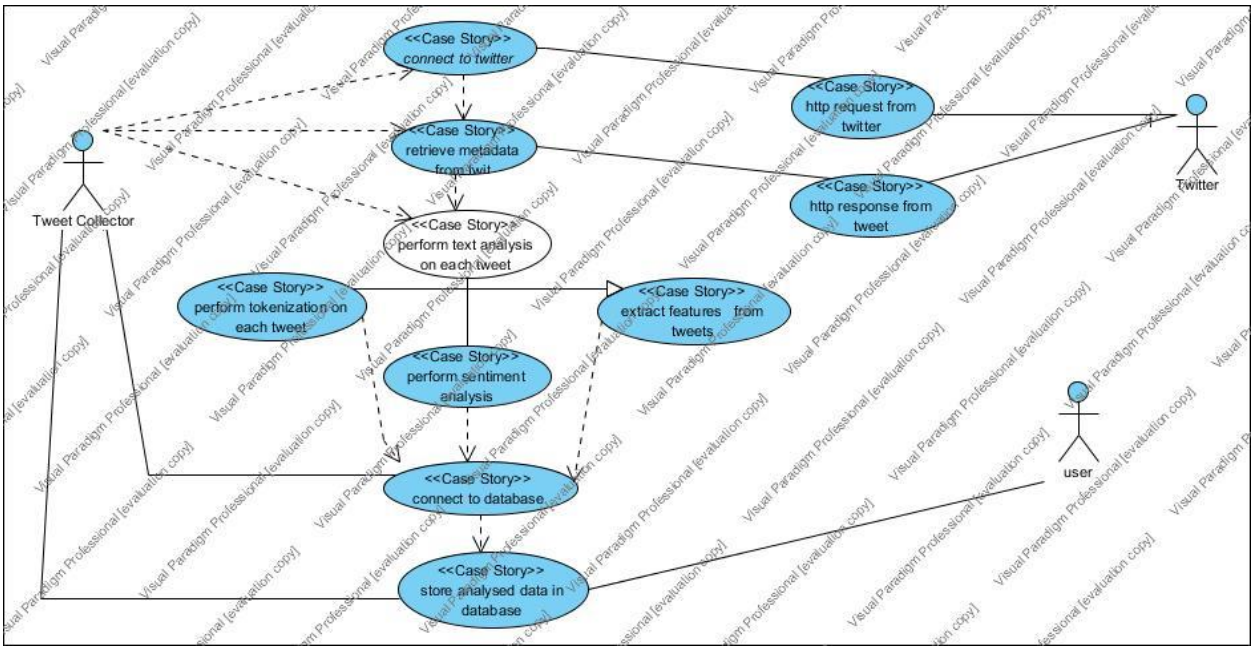

Figure 5:: Twitter Sentiment Analysis Use case

The following use case describes the interaction with the different entities with the twitter API to access tweets and do sentiment analysis on them.

Figure 8: Object diagram for the sentiment analysis.

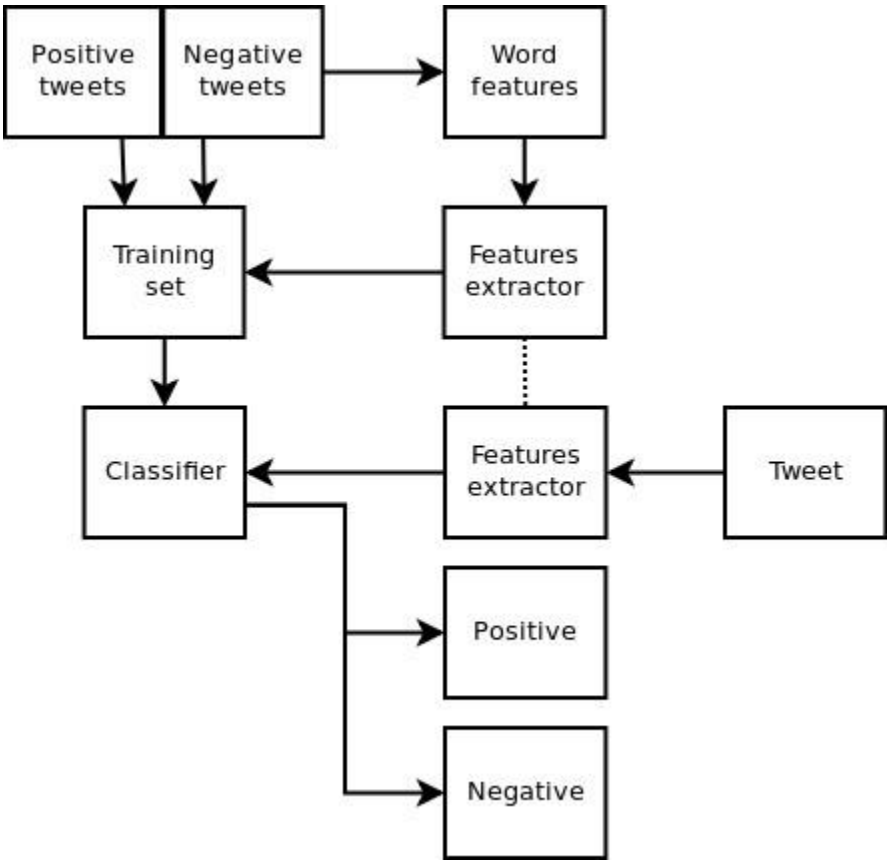

Figure 6: Sentiment Analysis, Object Diagram

4.1.3.3 System Architecture

Figure 10: system architecture

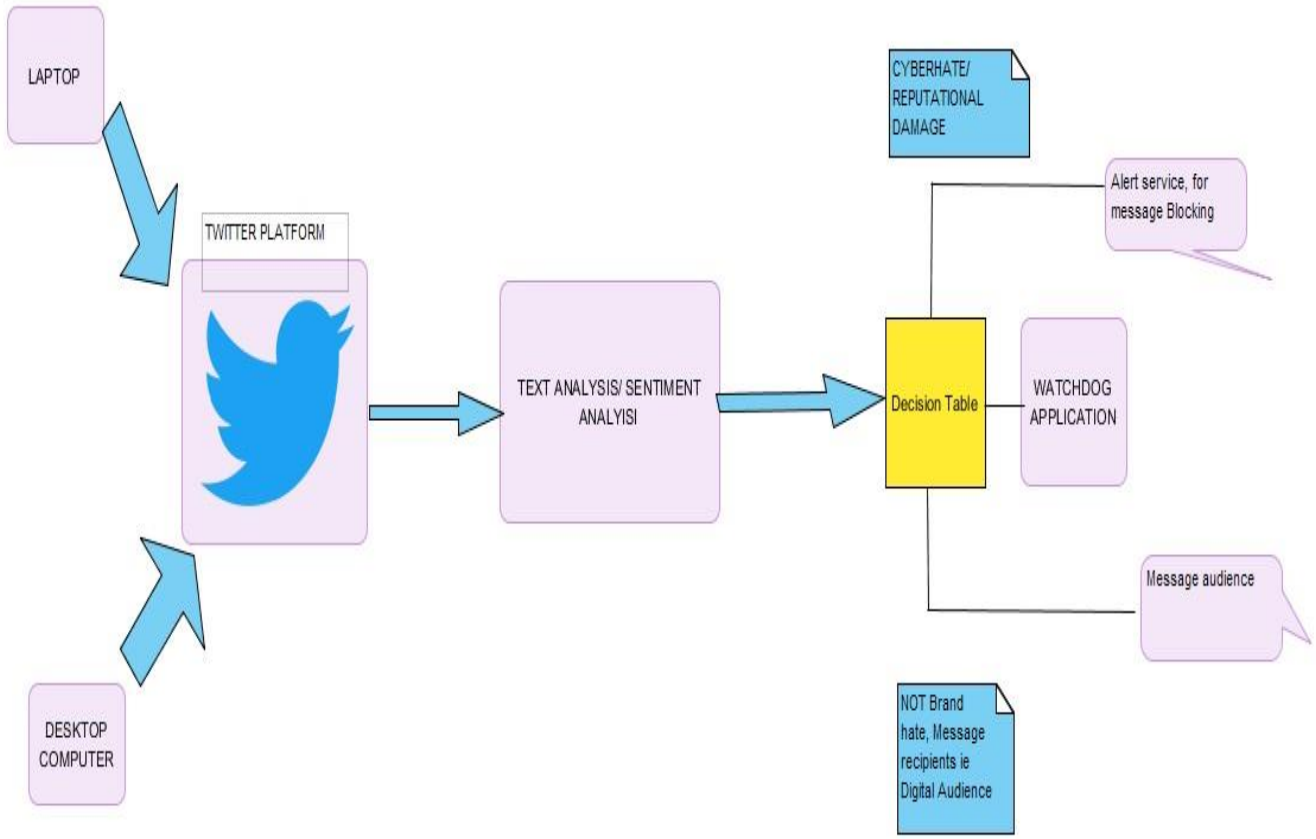

Figure 7: System Architecture

We can access the twitter API, from any laptop or desktop computer, then we can do a stream on the tweets for the mentions of Safaricom Company. Finally, we do text analysis on this tweets to classify them whether they are positive or negative and assign a polarity/confidence score. Eventually, for tweets with a higher polarity score, we send notifications via e-mail to the

personnel responsible for social monitoring from which, He can figure out the tweets to be flagged before they create an uproar on twitter if they are of effect to the company's reputation.

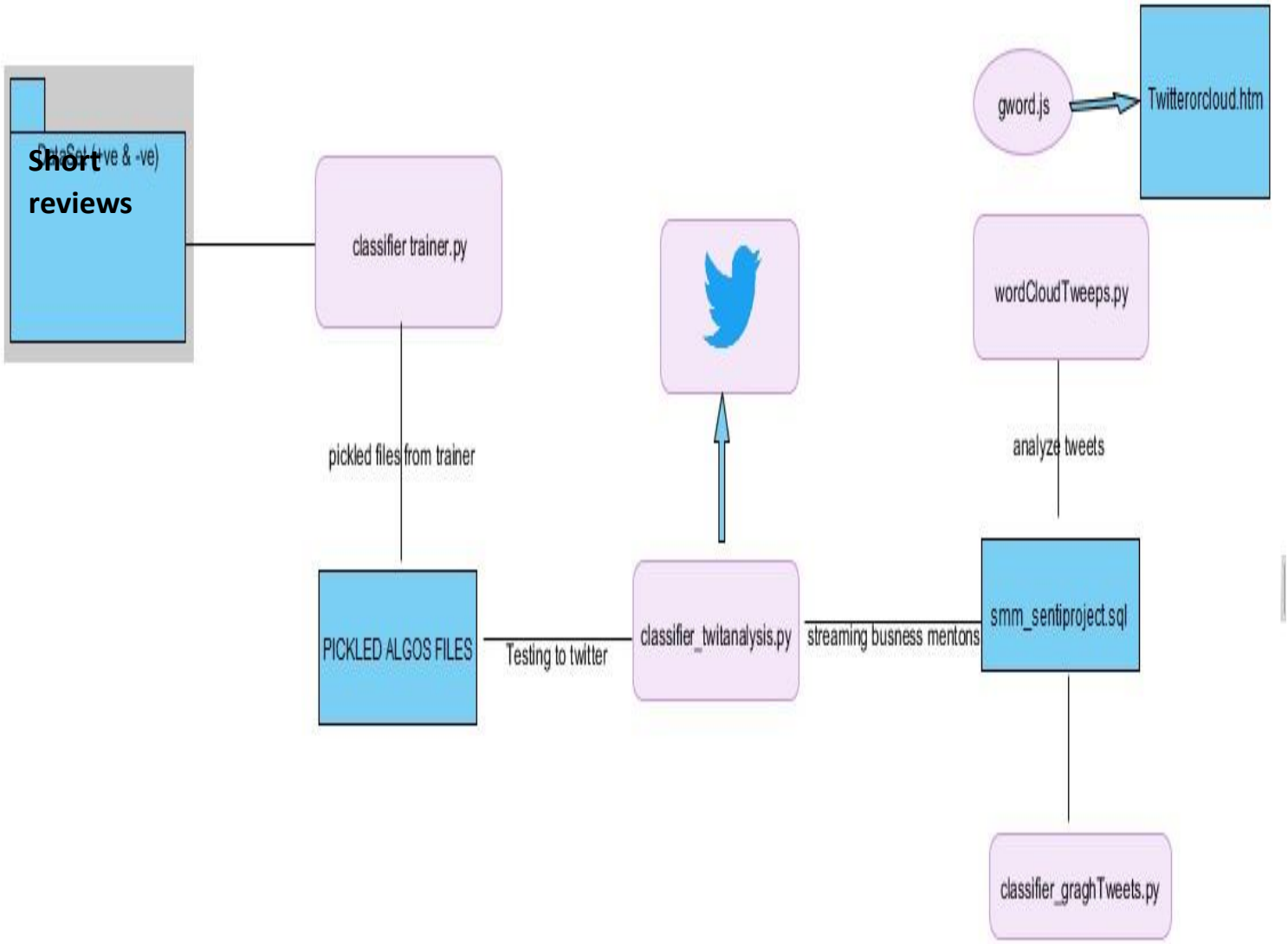

Figure 8: Module Interaction (twitter sentiment analysis)

Figure 11: The system module flow.

The flow of the system modules and code files, the classifier_trainer.py is trained against a dataset of positive and negative tweet data in the folder “Short reviews”. The trained data is saved in binary data as pickled files for each of the algorithm used into the folder “Picked algos file”. Consequently, the Classifier_twitanalysis.py uses the trained data and accesses twitter leveraging the twitter API. It then does sentiment analysis on the tweets containing only mentions of Safaricom. This is done by obtaining a stream of tweets for the mentions of Safaricom. This stream is classified into positive and negative tweets and they are assigned a polarity score. This data is then crawled into a database smm_sentiproject.sql, where it is stored for analysis. The analysis is done by the wordCloudtweeps.py. Which then writes the analyzed tweets to gword.js, which can later be viewed in a browser by opening the twittercloud.htm file to see the accounts tweeting the most negative tweets with high polarity score.

4.1.6 Implementation

Six algorithms were used in the classifier which can be shown in the figure below with their accuracy percent, then lastly a voted classifier is used to find an average accuracy that is overall used as the accuracy percentage of the classifier. Using more than one algorithm is meant to improve the accuracy of the classifier.

```
In [11]: run classifier_trainer.py
10662
Original Naive Bayes Algo accuracy percent: 71.29909365558912
Most Informative Features
              engrossing = True              pos : neg    =     20.8 : 1.0
                 routine = True              neg : pos    =     15.8 : 1.0
                 generic = True              neg : pos    =     15.2 : 1.0
                    flat = True              neg : pos    =     14.3 : 1.0
              refreshing = True              pos : neg    =     13.5 : 1.0
               wonderful = True              pos : neg    =     12.1 : 1.0
                    warm = True              pos : neg    =     12.1 : 1.0
                mindless = True              neg : pos    =     11.8 : 1.0
               realistic = True              pos : neg    =     11.6 : 1.0
                   stale = True              neg : pos    =     10.4 : 1.0
                tiresome = True              neg : pos    =     10.4 : 1.0
                  stupid = True              neg : pos    =     10.3 : 1.0
           extraordinary = True              pos : neg    =     10.2 : 1.0
             mesmerizing = True              pos : neg    =     10.2 : 1.0
                     wry = True              pos : neg    =      9.6 : 1.0
MNB_classifier accuracy percent: 71.90332326283988
BernoulliNB_classifier accuracy percent: 71.45015105740181
LogisticRegression_classifier accuracy percent: 73.1117824773414
LinearSVC_classifier accuracy percent: 71.6012084592145
SGDClassifier accuracy percent: 69.18429003021149
voted_classifier accuracy percent: 71.75226586102718

In [12]:
```

Figure 9: Classifier algorithms and their accuracy percentage

Figure 10: the graph of tweets with a polarity score higher than 80%

```
twitterStream.filter(track=["safaricom"])
```

```
('She_united', 'RT @saint_makaveli: Dear safaricom\nI slept with 1.5 GB of data then i wake up to find " your data is below 0.8
 MB\' sms jeeez was i streamin…')
('iKinuthia_', 'RT @PorkReebz: In my pants. https://t.co/UtyLgQN7Pl')
('Timberwolf___', 'RT @saint_makaveli: Dear safaricom\nI slept with 1.5 GB of data then i wake up to find " your data is below
 0.8 MB\' sms jeeez was i streamin…')
('Paulo_Adoop', 'RT @saint_makaveli: Dear safaricom\nI slept with 1.5 GB of data then i wake up to find " your data is below 0.
8 MB\' sms jeeez was i streamin…')
('njuechris5', 'RT @SafaricomCare: Please find assistance from the official Safaricom customer care twitter account @Safaricom_
Care')
('kaptila44silas', '@SafaricomLtd Send the mpesa menu to my phone 0724158285. It just disappeared in my phone with the safarico
m toolkit.')
('OketchDerrick2', 'RT @saint_makaveli: Dear safaricom\nI slept with 1.5 GB of data then i wake up to find " your data is below
 0.8 MB\' sms jeeez was i streamin…')
('RobahRdm', 'RT @saint_makaveli: Dear safaricom\nI slept with 1.5 GB of data then i wake up to find " your data is below 0.8 M
B\' sms jeeez was i streamin…')
('gikuyu254', 'RT @PorkReebz: In my pants. https://t.co/UtyLgQN7Pl')
('Lets_B_Real', 'After Launch in Nairobi, Safaricom’s Little is going to Nigeria https://t.co/FNAgb4qCH0 viaafrica #business #e
ntrepreneur')
('innov8tivmag', 'After Launch in Nairobi, Safaricom’s Little is going to Nigeria https://t.co/NkcDpdQyRB viaafrica #business #
entrepreneur')
('IBOMLLC', 'After Launch in Nairobi, Safaricom’s Little is going to Nigeria https://t.co/tJoCyrmEr4 viaafrica #business #entre
preneur')
('GuruAfrica', 'After Launch in Nairobi, Safaricom’s Little is going to Nigeria https://t.co/mSqtsNWQxJ viaafrica #business...
https://t.co/ZP01vOL59w')
('NBITLO', 'After Launch in Nairobi, Safaricom’s Little is going to Nigeria https://t.co/RO0ktPkIHP viaafrica #business... http
s://t.co/BQmp9mgZfU')
('IBOMLLC', 'After Launch in Nairobi, Safaricom’s Little is going to Nigeria https://t.co/tJoCyrEfiC viaafrica #business... htt
ps://t.co/P2hRR0AXjz')
('Safaricom_Care', '@jmurrayth browsing session on the Safaricom 4G network and the 4GB Data Bundle will be sent to you. ^NJ')
('TeddyLumidi', 'Safaricom planning to Release holistic M-Pesa API for Developers \nhttps://t.co/FLuJPtCq9g via @techweez')
('donxut6', 'jmurrayth browsing session on the Safaricom 4G network and the 4GB Data Bundle will be sent to you. ^NJ')
('TimKanya', "Experience the thrilling life with Safaricom 4G. If you are a virgin don't wait!")
('BensonM00985352', 'RT @SmileInvestClub: To follow us via sms and be able to receive live updates from us: \nSend sms to 8988
 (Safaricom) or 4040 (Airtel) \nMess…')
('Nyaoks', "@SafaricomLtd why won't https://t.co/ort7jy8Au5 work!?")
('SafaricomLtd', '@zecky_obonyo Hi, DM your mobile number or login to your Selfcare account https://t.co/YdsIu4T5wG in order to
 view your billing.^JM')
('SafaricomLtd', '@obvin56 Hi. Please share your number we check and advice. You may also view billing on Selfcare here http
s://t.co/SETkSQf0BW  ^WP')
```

Figure 10 twitter accounts with polarity score higher than 0.8%

Figure 11: the graph of tweets with a polarity score higher than 80%

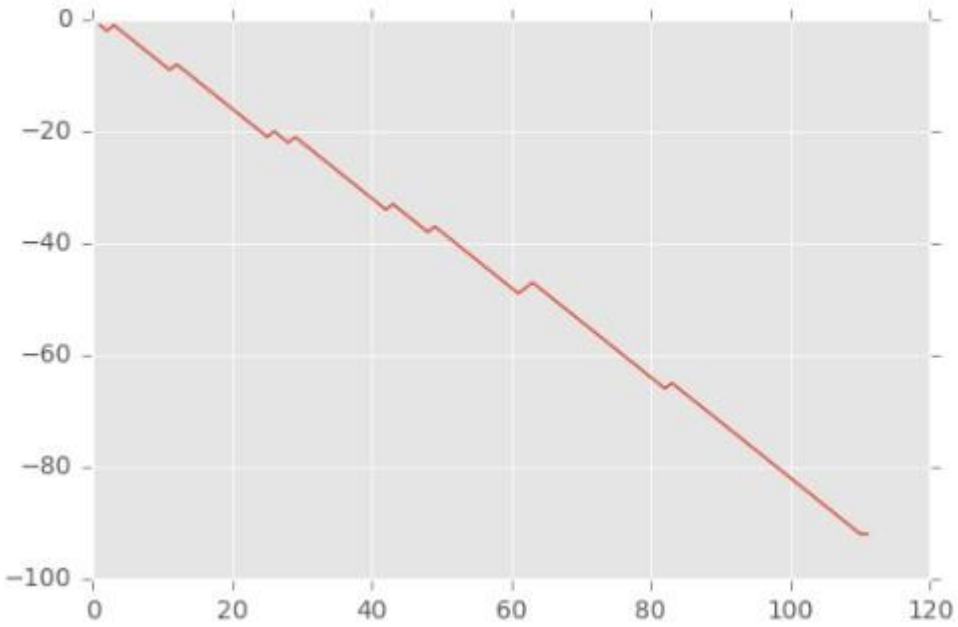

Figure 11: Graphing Tweets (positive against negative tweets)

Figure 12: showing accounts tweeting more about Safaricom

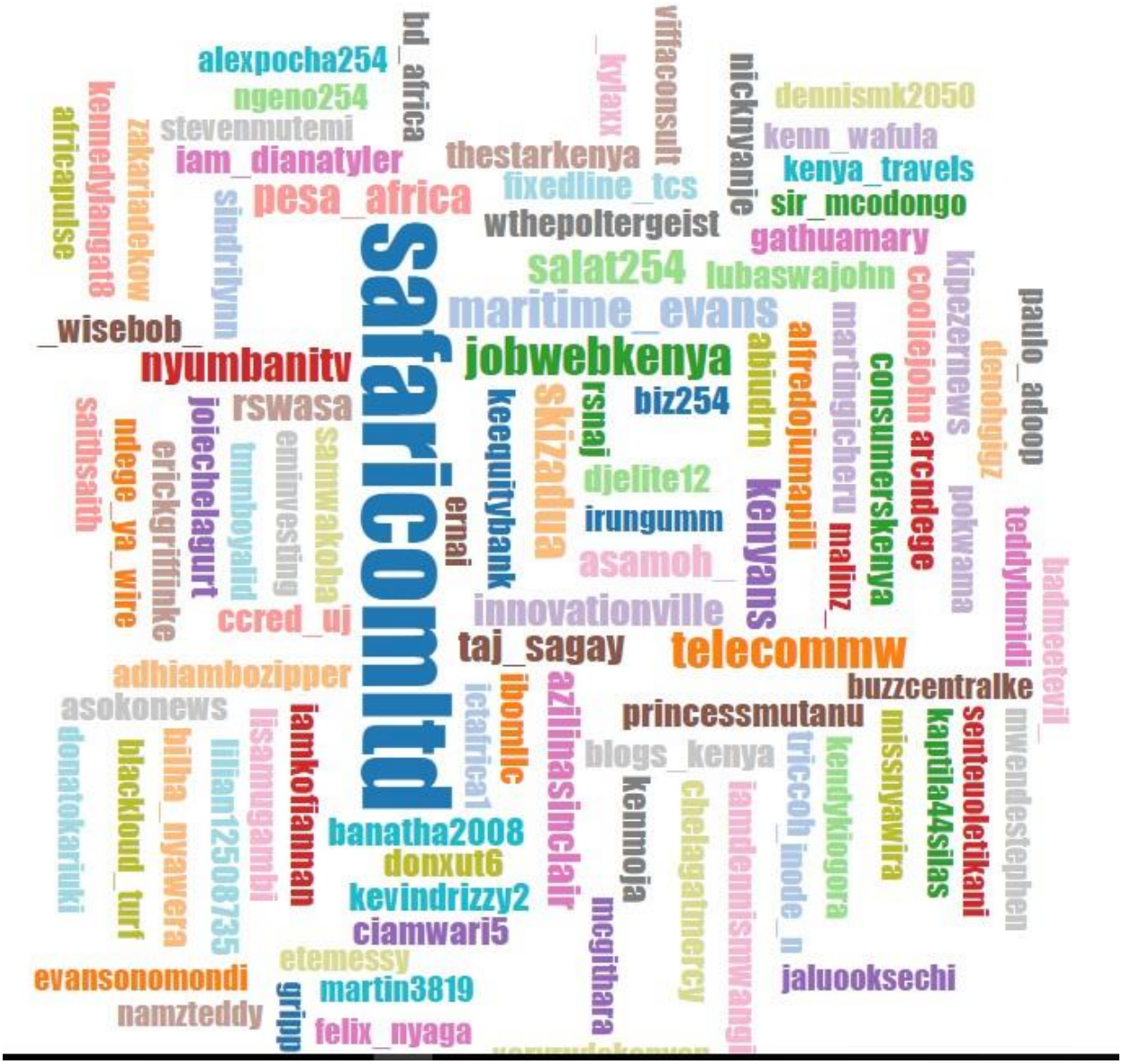

Figure 12: Word Cloud of Accounts with highest tweets about Safaricom

4.1.7 Testing and validation

Here we describe how testing and validation tasks were performed, by describing the plans and strategies used in unit testing, integration testing, and system testing. We also define the test plans and provide test procedures for testing the critical functions. Finally, we describe the test tools used.

4.1.7.1 Unit test

It involves wiring a set of tests that can be run, to guarantee that the code works as expected thus saving time and hence helps in the release of new updates.

Unit tests have a number of characteristics:

1. They do a number of thing for instance tests the module functionality, like ensuring that appropriate error notifications are thrown for instance when the system cannot retrieve tweets or if the API key has expired, or if the classifier is bugged. Unit tests are classified as per the components or modules.
2. They are mostly separated from the bulk of the code, since it's necessary to import and use the code being tested, this done by keeping them in different classes.

With the current trend towards test driven development, unit test have extremely become popular and because of their flexibility and length they are easily used by python.

4.1.7.2 Sampling

For this project, drawing from the knowledge of linguistics, one of the best models for conducting tests is by sampling people being modelled. This was done by asking people if they also deemed the classified texts as being negative or positive and when they also assigned the sentiment value the project model generated. We concluded that the classifier was accurate. This can be done by asking ordinary people their perceptions of the text.

4.1.7.1.1 Problems with sampling

It is susceptible to sampling bias, since people have contrary views. To overcome this bias more people should be able to review the sentences manually, it thus provides a broad representative of samples across the social demographic. In addition the viewers should be

selected randomly, this can eventually provide a mean score that is a representative of the true value.

A survey conducted on a broad sample, and a comparative of the mean score of the respondents and the model score is effective. That is one can be able to determine whether the values fall within one standard deviation of the respondents mean value. If they don't fall than probably the model is not as effective.

4.1.7.3 Using more than one Algorithm in the classifier

Using more than one algorithm for the classifier was also a test of the accuracy of the sentiment analysis. The results for each method/ classifiers are significantly different, most of the algorithms had an accuracy percent that fell within a very close range.

The table below shows the algorithms used for modelling the classifier and their accuracy percent.

Algorithm	Accuracy Percent	
Naïve Bayes algorithm	**71.299**	
Multinomial Naïve Bayes	**71.903**	
Linear Support Vector	**73.111**	
Bernoulli Naïve Bayes	**71.450**	
SGDClassifier	**71.601**	
Voted classifier(average)	**71.752**	**Mean of the classifier: taken for the sentiment analysis as voted classifier**

Figure 13 Testing Algorithms Accuracy