1. 
3. **INSTITUTE OF SOFTWARE TECHOLOGY**
4. PROJECT NAME
5. **COURSE**
6. **BY**
7. **NAME ADM**
9. A Project (proposal) submitted to the Department of Cyber Security and Ethical Hacking in Partial Fulfillment of the Requirements for the Award of Diploma in yber Security and Ethical Hacking at institute of software techology.
10. Date I.e june, 2024

# DECLARATION

1. We declare that this project is our original work and has been submitted to (name of school) for examination or award of a diploma.
2. Signed....................................................... Date....................................
3. Names
4. This project has been submitted for Examination with my authority as the College supervisor
5. Signed......................................................Date....................................
6. **Supervisor name**
7. Head of Department, Department of Cyber Security and Ethical Hacking
8. institute of software techology

# DEDICATION

# ACKNOWLEDGEMENT

# ABSTRACT.

Social networks are the most important communication channels in recent years, which popular among the different social groups. These networks affected the ideas and policies of individuals, groups and communities. Every day, millions of tweets on Twitter are being published. These tweets reflect opinions and beliefs of their publishers and affect others as well. Therefore, it is important to analyze these tweets and identify and classify trends of different users.

This research aims to classify social network to anomaly groups such as: Terrorist,and dissident; by analyzing tweets data on the Twitter; then identify an anonymous user’s affiliation to these groups. To address this problem, we first extract a set of features to characterize each group using different data mining techniques and store these features in the database. Text mining, sentiment analysis, and opinion mining techniques will be used to accomplish this extraction.

Furthermore, methods and techniques for the capture of Twitter timeline data, inclusive of first person and third party methods for data capture from personal accounts, public accounts, and keyword searches. These searches focuses on individual timelines, and small to medium scale search sets. The emphasis is on being able to obtain, examine, and convert Twitter data into knowledge quickly, and with limited requirement for technical skills.

The objective of data extraction is to measure the similarity of selected user tweets with respect to extracted features. It will enable to determine high percentage of similarity between the user tweets and group characteristics to expose his/her affiliation to this group.

The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Web sites contain millions of unprocessed raw data.

By analyzing this data new knowledge can be gained. Since this data is dynamic and unstructured traditional data mining techniques will not be appropriate. Data mining techniques, together with natural language processing are implemented in order to find context behind data. Moreover, sentiment score will be visualized on graph.

# TABLE OF CONTENTS

Contents

[DECLARATION ii](#__RefHeading___Toc1382_1347673152)

[DEDICATION iii](#__RefHeading___Toc1384_1347673152)

[ACKNOWLEDGEMENT iv](#__RefHeading___Toc1386_1347673152)

[ABSTRACT. v](#__RefHeading___Toc547_1601238674)

[TABLE OF CONTENTS vi](#__RefHeading___Toc1388_1347673152)

[LIST OF TABLES viii](#__RefHeading___Toc1390_1347673152)

[CHAPTER ONE 1](#__RefHeading___Toc549_1601238674)

[INTRODUCTION. 1](#__RefHeading___Toc551_1601238674)

[1.1 OVERVIEW. 1](#__RefHeading___Toc553_1601238674)

[1.2 BACKGROUND. 1](#__RefHeading___Toc555_1601238674)

[1.3 PROBLEM FORMULATION. 1](#__RefHeading___Toc557_1601238674)

[1.4 OBJECTIVES. 2](#__RefHeading___Toc559_1601238674)

[1.5 RESEARCH QUESTIONS. 2](#__RefHeading___Toc561_1601238674)

[1.6 SCOPE OF THE STUDY. 2](#__RefHeading___Toc563_1601238674)

[1.7 SIGNIFICANCE OF DATA MINING. 3](#__RefHeading___Toc565_1601238674)

[1.8 LIMITATIONS. 3](#__RefHeading___Toc567_1601238674)

[CHAPTER TWO 4](#__RefHeading___Toc569_1601238674)

[LITERATURE REVIEW. 4](#__RefHeading___Toc571_1601238674)

[2.1 Introduction. 4](#__RefHeading___Toc573_1601238674)

[2.2 Tweets and Retweets examiniation. 4](#__RefHeading___Toc575_1601238674)

[2.3 Tweet pre-processsing. 5](#__RefHeading___Toc577_1601238674)

[2.4 Tweet analysis. 6](#__RefHeading___Toc579_1601238674)

[CHAPTER THREE 7](#__RefHeading___Toc581_1601238674)

[METHODOLOGY. 7](#__RefHeading___Toc583_1601238674)

[3.2 Scrum Methodology. 7](#__RefHeading___Toc585_1601238674)

[3.2.1 Scrum methodology structure. 7](#__RefHeading___Toc587_1601238674)

[3.3 Basic Requirements Identification. 8](#__RefHeading___Toc589_1601238674)

[3.3.1 Product Requirements. 8](#__RefHeading___Toc591_1601238674)

[3.3.2 Development Requirements. 8](#__RefHeading___Toc593_1601238674)

[3.3.3 Application Requirements Analysis. 8](#__RefHeading___Toc595_1601238674)

[3.6 Scrum strengths and weakness. 11](#__RefHeading___Toc597_1601238674)

[CHAPTER FOUR 12](#__RefHeading___Toc599_1601238674)

[SYSTEM ANALYSIS. 12](#__RefHeading___Toc601_1601238674)

[4.1 Twitter Streaming API 12](#__RefHeading___Toc603_1601238674)

[4.1.1 Streaming API architecture. 13](#__RefHeading___Toc605_1601238674)

[4.2 Twitter Mining Application Setup. 13](#__RefHeading___Toc607_1601238674)

[4.2.1 Creating Twitter application. 13](#__RefHeading___Toc609_1601238674)

[4.2.2 Obtaining Twitter credentials. 14](#__RefHeading___Toc611_1601238674)

[4.3 Creating Streaming Connection. 14](#__RefHeading___Toc613_1601238674)

[4.4 Storing tweets in MongoDB. 16](#__RefHeading___Toc615_1601238674)

[4.5.1 Creating data frame 17](#__RefHeading___Toc617_1601238674)

[4.5.2 Twitter data domain. 18](#__RefHeading___Toc619_1601238674)

[4.5.3 Tweets mining and visualization. 18](#__RefHeading___Toc621_1601238674)

[4.6 Tweet Sentiment Classification. 18](#__RefHeading___Toc623_1601238674)

[6.8 Tweets Normalization. 19](#__RefHeading___Toc625_1601238674)

[SYSTEM DESIGN 20](#__RefHeading___Toc627_1601238674)

[5.1 Twitter and Tweets Handling. 21](#__RefHeading___Toc629_1601238674)

[5.2 Storing Tweets. 23](#__RefHeading___Toc631_1601238674)

[5.3 Tweets Pre-processing. 23](#__RefHeading___Toc633_1601238674)

[5.4 Sentiment Evaluation Methods. 23](#__RefHeading___Toc635_1601238674)

[Reference.. 25](#__RefHeading___Toc637_1601238674)

[APPENDICES 27](#__RefHeading___Toc1392_1347673152)

[APPENDIX 1 27](#__RefHeading___Toc1394_1347673152)

[APPENDIX 2 28](#__RefHeading___Toc1396_1347673152)

# LIST OF TABLES

list of Tables

Table1 12

Table2 17

Table3 18

Table4 19

list of Figures

[Shape1 i](#_toc47)

[Image1 13](#_toc561)

[Image2 14](#_toc584)

[Image3 15](#_toc592)

[Image4 15](#_toc594)

[Image5 16](#_toc598)

[Image6 17](#_toc612)

[Image7 20](#_toc750)

[Image8 22](#_toc780)

[Image9 22](#_toc784)

# 

# CHAPTER ONE

# INTRODUCTION.

## 1.1 OVERVIEW.

This study tries to use Social networks since it is not only the most important communication channels but also the most dynamic in the recent years, which popular among the different social groups. These networks affected the ideas and policies of individuals, groups and communities. Every day, millions of tweets on Twitter are being published. These tweets reflect opinions and beliefs of their publishers and affect others as well. Therefore, it is important to analyze these tweets and identify and classify trends of different users.

## 1.2 BACKGROUND.

The main thing that drove us to doing this project is rise in the number Cybercrime victims and this criminals have not been apprehended and a little or close to nothing has been done in trying to solve this cases. This made us to try and find a way to help in solving some of the crimes that can be committed over social media I.E Cyber bullying and Cyber terrorism this is by the use of mining of tweets from hashtags and retweets then analyzing them to obtain username of the bullies or criminals and the location they posted this tweets there by helping in building a case or solving a criminal activity that had earlier own been committed by the criminal in question. Data mining also can be used to predict how a certain product produced has been accepted by members of the public.

## 1.3 PROBLEM FORMULATION.

Tweets posted in social media platform could be from a terror group for instance Al-sha-baab militants, this tweet will be circulating on twitter getting retweeted and hashtag-ed no one knowing for sure whether this threat is for real or just a prank by someone and if this turns out to be true no one would be held accountable. Usually most people assume that twitter and other social media platform provide encryption for whatever they post in there social so that they cannot be back traced to them at any given time but through data mining everything you tweet and hash tag can be retraced back to you. Therefore, this is a wrong assumption we need to be careful with what we retweet, hash tag and share in our social media accounts since you don’t know where they originated from and the information they carry but you can be charged for facilitating a crime that you don’t even know that it occurred. Therefore, when retweeting and hashtagging be sure of the origin of what you are trying to retweet keeping in mind that there are forged accounts as explained by Keretna, Hossny, & Creighton, 2013.

## 1.4 OBJECTIVES.

* To gather information and knowledge from twitter.
* To develop a knowledge-driven data mining assistant.
* To construct useful information networks from raw Data sets.
* To develop new data mining and machine learning algorithms for social networks.
* To search for patterns in large batches of data.
* To evaluate trends and probability of future events.
* To analyze tweets which aid in making decisions.

## 1.5 RESEARCH QUESTIONS.

* What are the methods of gathering information analyzing them to obtain meaningful information?
* What are the stages for developing and API?
* What is the mechanism used to predict future trends?
* What are the technique for data analysis?

## 1.6 SCOPE OF THE STUDY.

* **Automated discovery of previously unknown patterns** - Data mining tools sweep through databases and identify previously hidden patterns in one step.
* **Artificial neural networks** -Non linear predictive models that learn through training and resemble biological neural networks in structure.
* **Decision trees -** Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
* **Genetic algorithms -**  Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
* **Nearest neighbour method** - A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called the k- nearest neighbour technique.
* **Rule induction** - The extraction of useful if then rules from data based on statistical significance.

## 1.7 SIGNIFICANCE OF DATA MINING.

* Helps in decision making by use of data obtained from different sources.
* Helps in analyzing tweets to obtain crucial information that can be used to build a case.
* Helps in Geo-locating Cyber bullies and Cyber terrorists.
* Reveals hidden profitability and helps in overcoming risk factors.

## 1.8 LIMITATIONS.

* Twitter Streaming API is limited to stream rates that allow to stream only small part of the total volume of tweets there by leading to a limited number of tweets to be crawled.
* Available Data Set with pre-labeled tweets with sentiment polarity classification would need political oriented tweets for training and testing.
* There is little knowledge on how to apply content-mining methods.
* Some of the Tweets may be missed in a search.
* Time consuming.

# CHAPTER TWO

# LITERATURE REVIEW.

## 2.1 Introduction.

We all know that everything trending in social media affects our lives in one way or another this Interpersonal communication affects everything from decision making, well being to the spread of ideas, the persistence of stereotypes, and the diffusion of culture. But although it is clear that social transmission is both frequent and important, what drives people to share? Why are some stories and information shared more than others? Traditionally, researchers have argued that rumors spread in the 3 C’s (conflict, crisis,and catastrophe) and the major explanation for this phenomenon has been generalized anxiety (i.e.,apprehension about negative outcomes). Such theories can explain why rumors flourish in times of panic, but they are less useful in explaining the prevalence of rumors in positive situations, such as a festival or other social event. Although recent work on the social sharing of emotion suggests that positive emotion may also increase transmission, why emotions drive sharing and why some emotions boost sharing more than others remains unclear.(J. Berger, 2011).

## 2.2 Tweets and Retweets examiniation.

According to Naveed, Gottron, Kunegis & Alhadi who studied Twitter by examining retweets, thereby determining what the Twitter community considers interesting on a global scale. They used this to construct a function of interestingness to generate a model to describe the content-based characteristics of retweets. They trained a prediction model to predict for a given tweet its likelihood of being retweeted based on its contents. From the parameters learned by the model they deduced what the influential content features were that contributed to the likelihood of a retweet (e.g., hashtag, username(@), URL, questionmark, arousal, dominance, negative emoticon). They found that general topics affecting many users (e.g., Christmas) are more likely to be retweeted than narrow, personal topics. Also, messages with hashtags, usernames, URLs, question marks, and exciting and intense tweets are more likely to be retweeted than messages with exclamation marks or positive emoticons.(Borgatti, S. P. 2009.)

According to Suh, Hong, Pirolli & Chi who studied retweeting as the key mechanism for information diffusion in Twitter, they gathered content and contextual features from 74M tweets and used this data set to identify factors that were significantly associated with retweet rate. They found that, a mongst content features, URLs and hashtags have strong relationships with retweet ability. A mongst contextual features, the number of followers and followee’s as well as the age of the account seemed to affect retweet ability, while, interestingly, the number of past tweets did not predict retweet ability. Further more, they state that retweets have quite different content characteristics from normal tweets: 56.7% of retweets have URLs in them while only 19.0% of regular tweets have URLs. We believe that this research would inform the design of sense making and analytics tools for social media streams. A possible drawback of the method of Suh et al. Is that it does not study the causal relation between, for instance, URLs and the probability a tweet is going to be retweeted: the presence of URLs can be a result of a third variable that was not measured.(Esmin, A. A. A., Jr., R. L. D. O., & Matwin, S. (2012).

## 2.3 Tweet pre-processsing.

Several studies have been published that aim to extract knowledge from massive data on the online environment. According to Kathy Lee and other colleagues they have used two different data classification models Network-based and Text-based to classify Twitter trending topics into 18 general classes. Aim to help searching on Twitter by aid users to look at smaller subgroups of trending topic which will improve information retrieval. The key contribution of their study is the use of the structure of the social network, rather than using social information exclusively. (Lee et al., 2011)

When mining for data in twitter one should consider these facts; tweets are not grammatically structured as regular document, and the limitation on the number of characters, no more than 140 characters in generated tweet. One should also know that text-based classification will provide fair results. But, it is still useful in cases where we cannot be able to apply network-based analysis. Twitter sentiments can be analyzed by use of an algorithm and according to Jurek and others Twitter sentiment contents can be analyzed by developing a lexicon-based sentiment analysis algorithm to estimate the level of disorder and disruption during public events. This algorithm differs from other existing model in the way in which it aggregates the values of positive and negative sentiment words within a Twitter message. They also increase the accuracy of the algorithm by proposing evidence based combination function to be applied in cases when mixed of positive and negative words appear in a message. The best aspect of this study is multi-dimensional sentiment analysis rather than just label positive and negative only, but further evaluation is still needed.(Jurek, Bi, & Mulvenna, 2014).

According to Alsaleh, Nayak and XU clustering in data mining is technique that improves and assists the matching process in social networks. These authors have used this technique for grouping users with same characteristics together into communities, then matching different users to these communities. There are two main drawbacks of matching users to users on social networks, namely, matching accuracy and computational complexity. The clustering process of users in social networks is divided into two phases. The first is the data pre-processing phase and the second is the data mining process. The proposed system reduces the computational complexity by limiting measuring the similarity to assigned communities instead of all users. But additional data techniques should be explored such as association rules that could be developed and assessed to improve the matching of users in social networks. (Alsaleh, Nayak, & Xu, 2011).

## 2.4 Tweet analysis.

When mining tweets there several factors to consider so that the information you obtain is from the real owner of that account there is a technique that analyzes Twitter short message to recognize the user identity in twitter. According to Keretna, Hossny, and Creighton who proposed a technique that authenticate real accounts versus forged accounts using text mining technique. The technique at first analyses short messages from the original accounts, then extracts a set of linguistic features that

characterize the writing style of the account owner. These extracted features will be used as input to the learning model to generate a classification model. Finally, this classification process can predict the author of the Twitter message taken as input. This study helps overcoming harmful activities against social network users and introduces a new method to recognize a user’s writing style that can be applied in different domains. (Keretna, Hossny, & Creighton, 2013)

Therefore with such information from mined tweets you can try and reduce incidents of Cyber bullying and Cyber terrorism.

According to Lerman & Ghosh they argue that the social network structure behind news sites such as Twitter and Digg plays a crucial role in the spread of information within these sites. Differences between Twitter and Digg are primarily due to Digg’s use of their front page which promotes articles and a relatively small number of top users that promote articles. Popularity of news stories and blog posts on Twitter grows smoothly until saturation, via spreading through the follower graph. The evolution of a story on Digg shows two distinct phases: an upcoming phase and a promoted phase (showing a tipping point), driven by Digg’s promotion mechanism. On both sites, it takes a day or less for the number of votes/retweets to saturate to their final values. On Twitter, many URLs never spread beyond the seed and its followers. A handful of URLs spread more than ten hops from the seed. On Twitter, far fewer URLs spread within a community than on Digg. Each retweeter is most likely to follow only one previous retweeter. On Twitter, cascades are more tree-like because the follower graph does not have significant community structure. On both networks, though information cascades spread fast enough for one seed to infect thousands of users, they end up affecting less than 1% of the follower graph. (K. Lerman and R. Ghosh, 2010).

# CHAPTER THREE

# SYTEM/PROJECT METHODOLOGY.

**3.1 SDLC - Scrum methodology.**

This methodology is more timeboxed and planned since the entire project is split up into timeboxes called Sprints, and all the team sit together and plan for each Sprint the list of tasks. Once the team agrees and commits to completing certain tasks in a given time frame, the development team is expected to stick to the commitment and complete all tasks within the Sprint.

A key principle of Scrum is the dual recognition that customers will change their minds about what they want or need (requirements volatility) and that there will. Scrum adopts an evidence-based empirical approach — accepting that the problem cannot be fully understood or defined up front, and instead focusing on how to maximize the team’s ability to deliver quickly, to respond to emerging requirements, and to adapt to evolving technologies and changes in market conditions.

## 3.2 Scrum Methodology.

Scrum is most suitable since it is most flexible methodology and more appropriate for long-run development with frequent changes to requirements. In addition it is effective where the cost of delay is high and deadlines should meet a minimal delay. Scrum is often used when the end product is unclear or the requirements have no proper feedback from the clients. Here the client is involved in the whole process and determine and focus on certain sprint product backlog items that need to be completed (along with the team).

Furthermore, Scrum is an iterative and incremental agile software development framework for managing product development. It defines a flexible, holistic product development strategy where a development team works as a unit to reach a common goal. This method enables teams to self-organize by encouraging physical co-location or close online collaboration of all team members, as well as daily face-to-face communication among all team members and disciplines involved.

### 3.2.1 Scrum methodology structure.

* **Product backlog -** a set of top-priority tasks allowing to build MVP as soon as possible.
* **Sprint backlog -** contains high priority features that developers are going to deal with following 2–4 weeks)
* **Sprint itself**
* This growth methodology is used for prompt development of software that happens to include a series of iterations to generate required software. It brings the deliberate progressing projects on track.

## 3.3 Basic Requirements Identification.

### 3.3.1 Product Requirements.

* Twitter API.
* Wing IDE.
* Tweets and retweets.
* JSON.
* Hashtags.
* Tweepy.

### 3.3.2 Development Requirements.

**Tools and Techniques.**

* Computer(Linux Operating System)
* Python(Programming)
* Tweepy(Mining tweets)
* Anaconda
* Textblob

### 3.3.3 Application Requirements Analysis.

**Tweets and Retweets.**

A “tweet” is a string of characters or simply a message on twitter restricted to 280 characters. A person needs to open up an account in twitter where he/she can send or receive tweets and also be connected with friends. The idea of tweets began when twitter was developed in 2006, it’s co-founder Jack Dorsey had imagined it to be an SMS-based communications platform. Friends could keep posting status updates known as tweets to keep tab of each other. The 140 character restriction

on tweets is because it was originally designed as an SMS mobile phone based platform. Due to growth of twitter tweets are no longer confined to 140 characters, one can think of as big step for twitter but only a few user have been access to this 280 arbitrary based choice. In twitter, connecting with friends mean having to follow people and have followers who can view the posts shared on the timeline and send or receive tweets, retweet a message and so on. People express their thoughts, opinions and emotions through tweets where one can not only write in plain text but also include URLs and pictures. Twitter users can share news and events and retweet to one another post. Retweet is a re-posting of a tweet, retweets can be found in one’s timeline, profile and other profile pages on Twitter. The retweet feature quickly helps share a tweet with users’ followers and one can retweet his/her own tweet or someoneelse’s tweet. Retweets look like normal Tweets with the author's name and username next to it, but are distinguished by the Retweet icon and the name of the user who Retweeted the Tweet. Some users can block the option of retweeting their posts so that others cannot retweet. One can see who have retweeted their tweets from the notification tab. There is no limit to the number of times a tweet can be retweeted but only top and most recent 100 people's’ tweets will be shown on the Home timeline who retweeted public posts. Currently the limit per day for each account is 1000 direct messages sent, 2400 tweets where retweets are counted as tweets. When hitting a limit, twitter sends an error message that limit has been reached to try messaging or sending tweets after the limit period has elapsed.

**Wing IDE**

In computer programming, wing IDE is an integrated development environment (IDE). It contains a base workspace and an extensible plug-in system for customizing the environment. Wing IDE smart code editor provides first-class support for Python, JavaScript, CoffeeScript, TypeScript, CSS, popular template languages and more. Take advantage of language-aware code completion, error detection, and on-the-fly code fixes!. Wing IDE offers great framework specific support for modern web development frameworks such as Django, Flask, Google App Engine, Pyramid, and web2py. WingIDE works on Windows, Mac OS or Linux with a single license key. With Wing IDE you can enjoy a fine-tuned workspace with customizable color schemes and key-bindings, with VIM emulation available. In addition to Python, Wing IDE supports JavaScript, CoffeeScript, TypeScript, HTML/CSS, Cython, template languages, AngularJS, Node.js, and more.

**Python**

Python is a widely used general-purpose, high-level programming language. Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of code than would be possible in languages such as C. The language provides constructs intended to enable clear programs on both a small and large scale. Python supports multiple programming paradigms, including object-oriented, imperative and functional programming or procedural styles. It features a dynamic type system and automatic memory management and has a large and comprehensive standard library. Like other dynamic languages, Python is often used as a scripting language, but is also used in a wide range of nonscripting contexts. Using third-party tools, such as Py2exe or Pyinstaller, Python code can be packaged into standalone executable programs. Python interpreters are available for many operating systems. CPython, the reference implementation of Python, is free and open source software and has a community-based development model, as do nearly all of its alternative implementations. CPython is managed by the non-profit Python Software Foundation.

**Twitter Rest API**

Twitter is a popular social network where users can share short SMS-like messages called tweets. Users share thoughts, links and pictures on Twitter, journalists comment on live events, companies promote products and engage with customers. The list of different ways to use Twitter could be really long, and with 500 millions of tweets per day, there’s a lot of data to analyse and to play with. The REST APIs provides programmatic access to read and write Twitter data. Author a new Tweet, read author profile and follower data, and more. The REST API identifies Twitter applications and users using OAuth; responses are available in JSON.

**Hashtags**

Anyone using twitter is familiar with the term “hashtag” which is actually used to simply categorize a tweet’s topic. Hashtags make it easier for users to identify a message under some specific theme or content. Hashtags are not case-sensitive so if a user searches for a specific tag on the search box for example if the user searches for “#NTV”, all the recent tweets or posts related to NTV or containing the tag #NTV will appear as the results. The result page has different filtering options for the given list of results, the default is Top. One can choose live feed as well as feeds of tweets that include links to current news stories, photos and videos. There are filter options such as from everyone, from people you follow, near you, etc. One of the most amusing things about hashtags is that it allows to create communities of people who are interested in similar topic by making it easy to share related information.

**Tweepy**

Tweepy is one of the python libraries which is open-sourced, hosted on GitHub and enables python to communicate with twitter platform and give the access to use its API. Tweepy can access Twitter using newer authentication method, OAUTH. This authentication method has consumer key and access tokens that are provided from the app created at dev.twitter.com for better security. It is possible to get any object with Tweepy that the official twitter API offers. Streaming twitter data is one of the main usages of Tweepy and the key component is the StreamListener object which monitors the tweets in real time and catches them.

**JSON.**

JSON (JavaScript Object Notation) it is a lightweight data-interchange format, designed to be easy for both humans and machines to read and write. It is easy to parse and generate, and easy for humans to understand. It is based on a subset of the JavaScript Programming Language.

JSON is built on two structures:

* A collection of name/value pairs. In various languages, this is realized as an *object*, record, structure, dictionary, hash table, keyed list, or associative array.
* An ordered list of values. In most languages, this is realized as an *array*, vector, list, or sequence.

A JSON object consists of key-value pairs (1 or more), where the value can be another JSON object. This means embedded JSON is possible.

**3.5 Twitter Data Analyses**

Next step in the methodology was dedicated to Twitter data collection and its analyses. As Twitter offers publicly accessible Application Programming Interface (API) our goal was to analyses different approaches to API, tokens, their implementation from theoretical side and finally collect tweets for sentiment analysis in order to develop classification method for training and testing. After collecting sample data set from Twitter, we started with data analyses in order to understand data structure of Twitter stream. Data analyses on tweet attributes are important in order to understand its format and prepare tweets for sentiment analyses, pre-processing phase and classification.

## 3.6 Scrum strengths and weakness.

**Strengths.**

* Decision making lies in the hands of the team.
* Business requirement document is considered insignificant.
* Lightly controlled method empathizing with constant updating.
* Scrum model is more flexible with disciplines which are open to last-minute changes.
* Teamwork, inspection, and transparency are key factors in the Scrum method.

**Weakness.**

* The processing method suffers because of wavering costs.
* Not suitable for big sized projects.
* Requires highly expert team, which has no place for novices.

# CHAPTER FOUR

# SYSTEM ANALYSIS.(Implementation and Testing)

Application Programming Interface (API) is an interface that enables interaction with web services. It gives developers and public to develop service products on top of API and thus implement it within own service solutions. Access to Twitter service is possible by two types of different methods: Streaming API and REST API.

## 4.1 Twitter Streaming API

Streaming API offers access to service, when tweets are retrieved as continuous stream of information. It is up to developer’s algorithm what kind of tweets are retrieved from API, such as keyword-based tweets, however in case of singular searches for user profile information, REST API will fit better.

REST API is stands for Representational State Transfer. Its architecture is based on network principles, which defines access methods to resource data. RESTful services communicate over request-response HTTP protocol, however in comparison with Streaming API, it does not require to keep persistent HTTP connection opened.

Such service applications make requests to API only when user explicitly requests data retrieval, for example information about followers or retweeted post on Twitter. Twitter service offers several types of streaming API endpoints, which has different capabilities, as shown by table 4.1. Twitter data are retrieved in JSON format.

**table 4.1 Twitter Streaming endpoints.**

|  |  |
| --- | --- |
| Public streams | Contain samples of public tweets,  recommended for data mining. |
| User streams | Contain all information about particular  Twitter user. |
| Site streams | For service applications that requires  handle streams from many users. |

From the above discussion, REST API will not serve sufficiently for purposes of this project, because of the HTTP server connection with user. On the other hand, Streaming API provides capabilities for real-time data mining with public stream endpoint, which is enabled by separation between streaming connection process and HTTP request process. Data mining technique, sentiment analyses will be implemented based on real-time stream from Twitter.

### 4.1.1 Streaming API architecture.

Hence, Streaming API design architecture enable to collect necessary data for pre-processing stage in methodology. Architecture of Streaming API is showed on Figure 4.2.

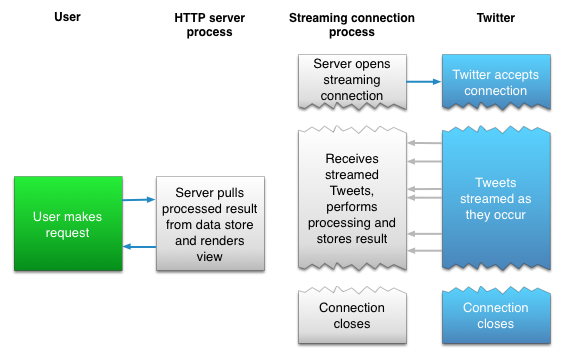


Figure 4.1: Twitter Streaming API dataflow.

The above architecture main advantage is that after streaming process retrieve tweets from stream, it can manipulate data before store the results. User requests to access data are then handled with HTTP process, which query requested data from data storage.

## 4.2 Twitter Mining Application Setup.

In order to initiate authorized calls to Streaming API and collect data for preprocessing phase, we need to create Twitter application that will obtain access token. Open Authentication (OAuth) is standard for authentication that provides capabilities for applications to access data from other service without revealing credentials. In our solution for data mining we want to establish connection to Streaming API and thus Twitter control panel for developers, apps.twitter.com offers

generating our access token.

### 4.2.1 Creating Twitter application.

Firstly, the process of getting access token starts with creating new application from Twitter Application Management panel by filling up required attributes for application, such as: Name, Description and Website.

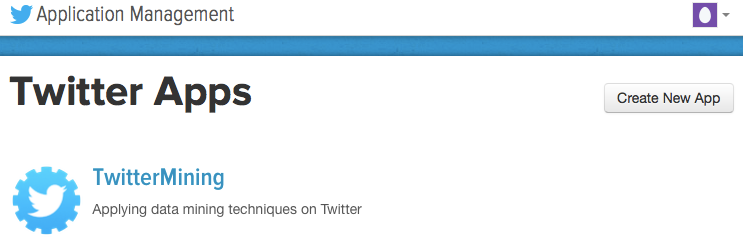


Figure 4.2 Twitter Application Setup.

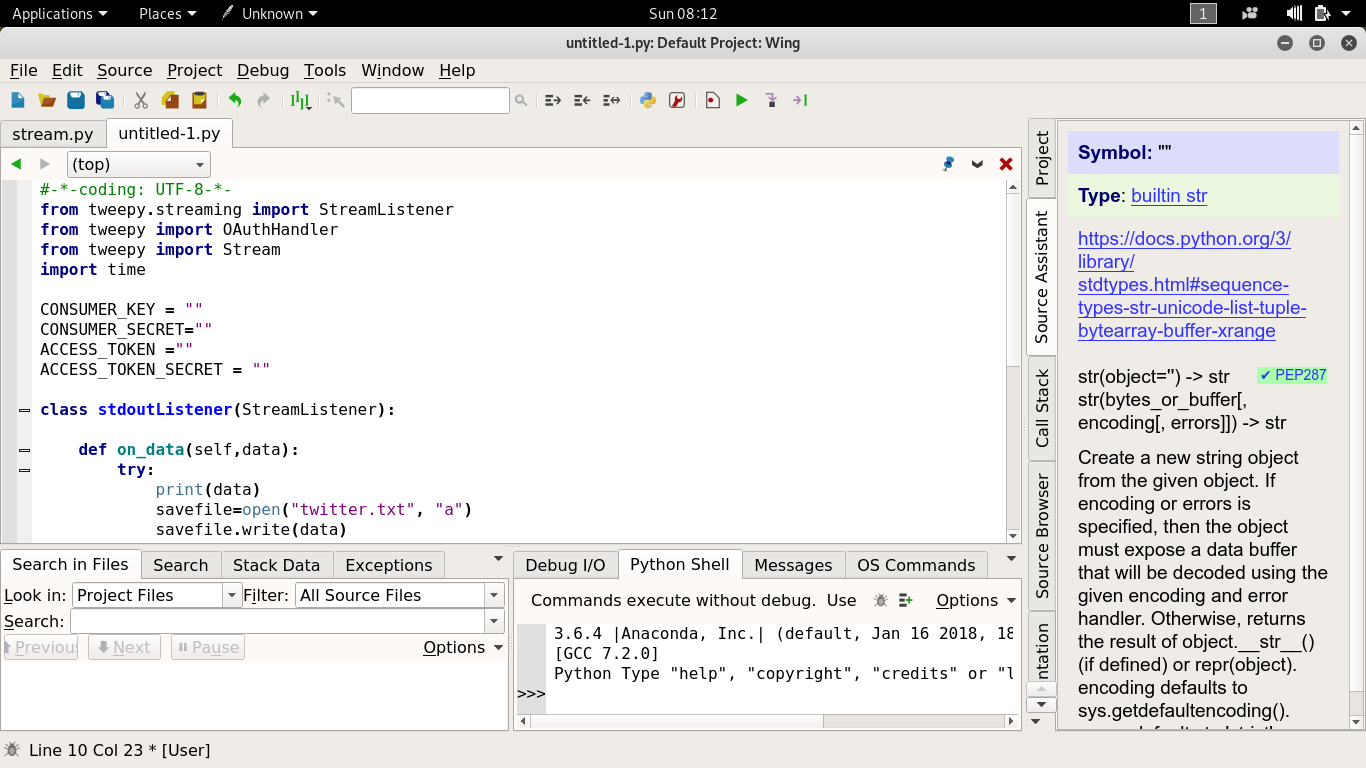
### 4.2.2 Obtaining Twitter credentials.

For the development and Streaming API access are important credentials located under Keys and Access Tokens tab in newly created TwitterMining application settings. There are four credentials to note: Consumer Key (API Key) and Consumer Secret (API Secret), Access Token and Access Token Secret. These credentials provide everything for TwitterMining application to authorize itself and make API requests on its Owner behalf. For our case at the moment we are still awaiting our developer’s account. Access tokens are used to make API requests to Twitter service from owners account. Moreover, access level is set to Read and write for purpose of this project, however if necessary it can be changed according to permission settings of application. Furthermore, tokens can be regenerated or revoked at any given time.

### 4.3 Creating Streaming Connection.

After successfully obtaining credentials for Streaming API, we need to initialize connection and collect sample tweets in order to analyze structure of each tweet and attributes they contain. API key and API secret have to be passed to OauthHandler that will create object auth in order to setup authentication while function set\_access\_token will setup Access Token and Access Token Secret. Code is included in the attachment file that comes with this project. Figure 4.3 Shows the

approach.

Figure 4.3 Twitter Streaming API authentication.

Data analyses require some sample partition of tweets to get familiar with its content, therefore the following code will stream near real-time tweets based on the user input. All tweets are printed in console, but for better overview they are piped also into text document. As example for user input: Donald Trump, tweet stream looks as on Figure 4.4.



figure 4.4 Twitter stream in JSON.

Idea of analysing tweets as part of pre-processing before sentiment analyses is to understand attributes within JSON format for each tweet and normalize them according to natural language processing paradigms. Following class is listener instance for Twitter streaming that will print tweets and save them into document TwitterMining.txt and inform user with message if any error occur with streaming.

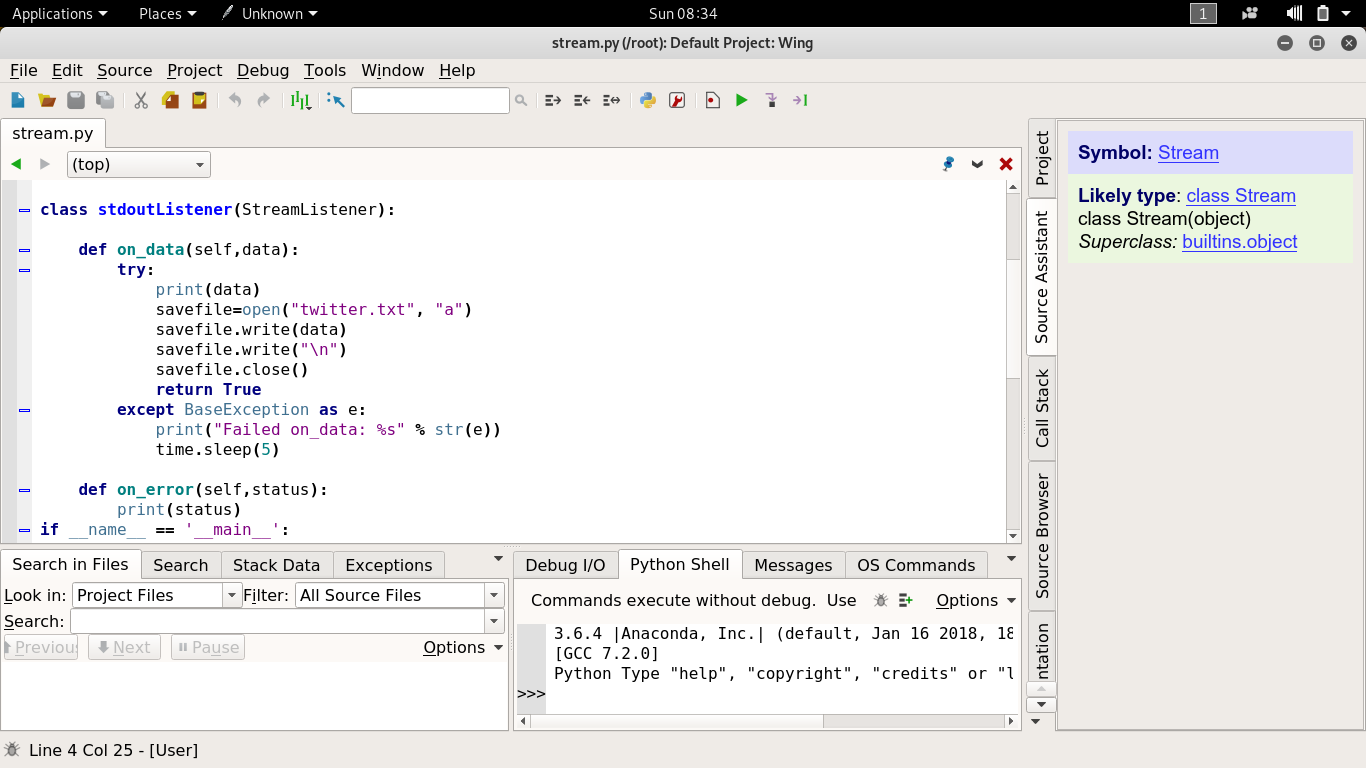


Figure 4.5 Listener class

Twitter streaming is ready after auth and listener() instances are passed to Stream object, that will now contain credentials for authentication and information about tweets, which have to be streamed. Moreover, the filter() method called on twitterStream will take input stream word from user and collect only tweets in English language. Process of collecting and analysing tweets is an important part of data mining, because it prepares data in format needed for processing. Collected Twitter stream of sample tweet dataset is now capable to store data in JSON format. JavaScript Object Notation (JSON) is format for data interchange based on comma-separated key-value pairs. It is human readable data format that is easy to parse.

### 4.4 Storing tweets in MongoDB.

In order to store dataset from API, not only in text document but also into database, we have implemented MongoDB. It is NoSQL open-source database that is capable of storing massive volumes of data and supports effective data integration in dynamic formats such as JSON. Developing in Python enables to implement library for MongoDB called pymongo. Anaconda is the middleware between Twitter Streaming API and MongoDB. It is a tool for interactive Python development, which serves in this project for Twitter data analyses. Following code will save tweets into MongoDB database, “”.

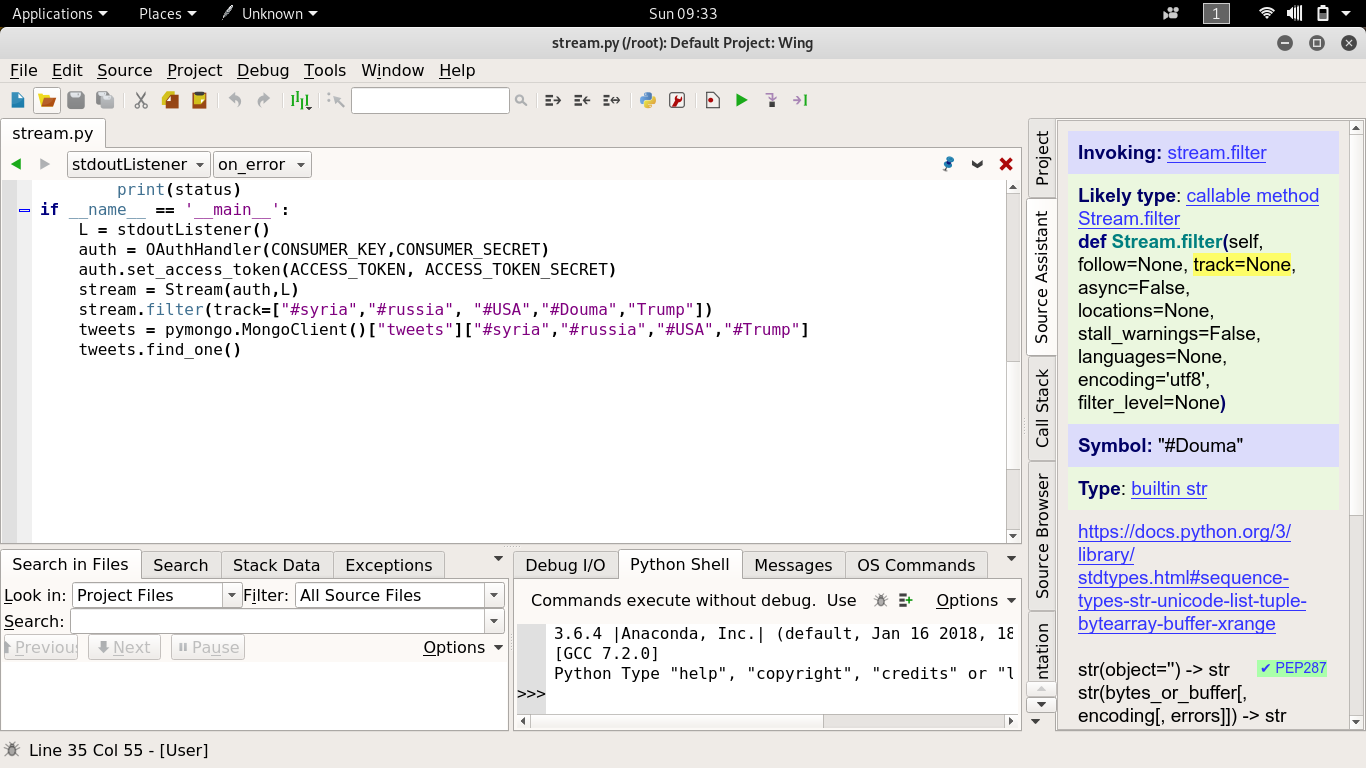


Figure 4.6 MongoDB connection.

**4.5 Tweets Attributes.**

Anatomy of the tweets stored in the database can be analyzed from the Anaconda. Following are some interesting attributes that can be contained in each tweet.

Table 4.2 Tweet attributes overview.

|  |  |
| --- | --- |
| Id | Unique tweet ID, that serves as identifier |
| Created\_at | Date when tweet was created |
| Favourite\_count | Number of favourites this tweet have |
| Lang | Language |
| Retweet\_count | Number of retweets |
| Text | Text of the tweet |
| Entities | Url, hashtags #, user mentions |

From the attributes overview we can implement different data mining algorithms to perform processing, such as the most favourited tweets in comparison to users profile or analyses towards most used hashtags. However, for the sentiment analyses and classification with machine learning, text attribute matter the most. Based on this tweet text will perform algorithm implicitly decisions if it contains positive or negative statement. Collecting tweets from Streaming API are stored in database, however there is capability to view near real-time streaming output also in text file, which is used for purpose in pre-processing.

### 4.5.1 Creating data frame

In data analyses, an important role is lead towards data selection based on data structures. Effective organized data into tables gives better overview for data scientists about its structures. In our tweet data analyses model, we use Data Frames enabled with Python Pandas. Data frame is capable to store data into tables and label its rows and columns. Hence, data mining can process data-driven solutions and deliver meaningful insights for companies, marketers, researchers, or brand management. Code is included in the attachment that comes with this project. Tweets that are stored in database are organized into Data Frames, we will have to write a code that will create Data Frame from collected tweets and displays in tabular view for each tweet in MongoDB database. Moreover, width of columns is configured for proper text displaying within its full range.

### 4.5.2 Twitter data domain.

Text attribute contains statements from Twitter Streaming API about USA, syria, Trump, Douma and Uhuru Kenyatta. The Reason for this particular study domain is because of their attention from worldwide media and social networks. They feature in almost all highly discussed events, which is necessary for obtaining near real-time streaming data. Capability of implementing algorithm on dataset has to be within domain, which outputs opinions or statements about particular event, product, brand, etc. Attention from users on Twitter on the event such as presidential elections provides necessary background for data mining and sentiment analyses, which distinguishes tweets based on their sentiment into positive, neutral and negative category.

All tweets are streamed from public API and therefore the opinions that come from random Twitter users.

### 4.5.3 Tweets mining and visualization.

In social media can have huge impact on perception, aim is to deliver solution that classifies streaming tweets based on sentiment and results into public opinion, and visualized with graph. Twitter data can be stored under assigned variable. Therefore, different analyses methods can be applied directly on this data frame. In order to explore time zones and get geographical insights about Twitter users, which posted tweet with keyword “Donald Trump”, “Syria”, “Douma”, “Uhuru Kenyatta” we can apply count method and visualize the output.

### 4.6 Tweet Sentiment Classification.

Sentiment classification is part of machine learning techniques, which are further analysed on top of Twitter dataset. Table 4.3 is an example of tweet sentiment classification. Text was manually taken from tweets that are stored in database to show sentiment classification capabilities.

Table 4.3 Sentiment categories.

|  |  |
| --- | --- |
| Positive | Uhuru kenyatta fights corruption. |
| Positive | I just want to say, "Donald Trump will be the GREATEST PRESIDENT EVER!!! |
| Neutral | NTV Releases Statement About Uhuru Kenyatta. |
| Neutral | Freedom and protection for all. |
| Negative | New report says Trump wants Chinese parts out American weapons. |
| Negative | Waste, Fraud and abuse Don jr cost tax payer’s nearly 100,000$ for a trip to India to sell Trump properties. |

### 6.8 Tweets Normalization.

Twitter Streaming API stream its data in JSON format as discussed earlier. Tweets contain various key-value attributes about its context and information about user profile. For sentiment analyses is most important the text of actual tweets on which base is deployed algorithm that will classify them into sentiment categories by their polarity.

In order to perform data cleaning algorithm on dataset with streamed tweets, initial code has to specify which segments of JSON attributes to include. Attribute under which is tweet sentence located is in JSON called – text. Hence, splitting up tweet data by text will give us the ability to clean tweets from other unnecessary attributes, such as: “created\_at”, “id\_str”, “source”, etc.

Table 4.4 Twitter data cleaning.

|  |
| --- |
| Twitter Data Analyses. |
| Stream keyword: Donald Trump |
| **Twitter Streaming API data format** |
| {"created\_at":"Sun Nov 18 04:53:37 +0000 2018","id":1064018766244261893,"id\_str":"1064018766244261893","text":"Cheer up Democrats. President Trump has tweeted 5,559 times since his inauguration. At his current pace, you can on\u2026 https:\/\/t.co\/5ypmZLOz7v","source":"\u003ca href=\"http:\/\/twitter.com\" rel=\"nofollow\"\u003eTwitter Web Client\u003c\/a\ |
| **Splitting by text tweets cleaning** |
| "Cheer up Democrats. President Trump has tweeted 5,559 times since his inauguration. At his current pace, you can on\u2026 https:\/\/t.co\/5ypmZLOz7v" |
| **Pre-processing tweet format** |
| Cheer up Democrats. President Trump has tweeted 5,559 times since his inauguration. At his current pace, you can on\u2026 |

Pre-processing phase of Twitter data mining shows Table 4.4 . Tweets that are streamed from Twitter Streaming API in JSON format are now cleaned from redundant information and normalized for purpose of sentiment analyses. Above table depicts the difference between tweets format before and after data cleaning, which is requirement for further classification.

# SYSTEM DESIGN

This project's goal is to build a system that is able to retrieve Tweets from twitter, filter them, normalise them, give them polarity scores and visualise the produced results for the user. Activity diagram in Figure 5.1 illustrates the flow of the application to be built and each main step will be explained throughout the subsections of this chapter. All design decision were taken considering that Python is the programming language that the system to be built is written in.

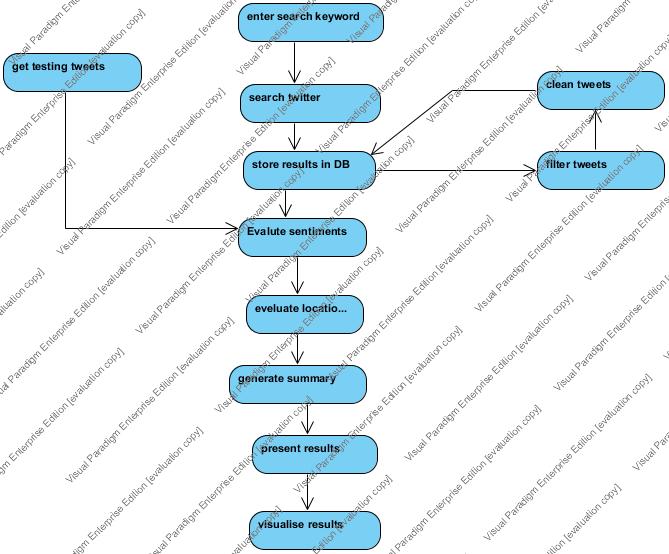


figure 5.1 Activity Diagram.

## 5.1 Twitter and Tweets Handling.

The first thing our software requires is the ability to retrieve Tweets from Twitter, so we can later classify them. However, we do not want to collect every Tweet from Twitter, since 500 million are sent every day. We can assume that Tweets expressing Cyber Bullying or Cyber Terrorism will form a small subset of the total Tweets sent every day, so it would be computationally intensive and wasteful to collect all of them; most would be irrelevant (noise). It would be an unachievable goal anyway, because Twitter limits the total amount of Tweets that can be accessed through their APIs to about 1% of the total, which also means that if we tried to collect every Tweet being created and then tried filtering through them at a later time, we cannot make sure that the Tweets Twitter allowed us to collect contained any which expressed social isolation.

Tweets are not only associated with their textual content, these are believed to be among the most important attributes of a Tweet needed by this project:

• **Twitter user's ID** which is used for calculating the number of Tweets about the investigated suspecting highlighting their victims.

• **Time/Date** which are used when the Tweets are being analysed depending on posting time or plotted with sentimental values over time. Also, it can be used for filtering.

• **Tweet's ID** which is used for identifying duplicate Tweets when the system is used to retrieve and store the Tweets automatically i.e. if the Tweets are already stored they will be discarded from the storing process.

• **URLs** which might be used in any future work to retrieve their content and analyse it or for presenting some multimedia for positive, negative and/or neutral Tweets.

**• Retweets count** this can be used for ranking purposes and for identifying the most important Tweets which can be used in computing the overall sentiment of the investigated topic.

• **Place** which might be used in order to locate where the tweet originated from.

Once Twitter API is used to search Twitter over a given query, then the resulting Tweets are used for the purpose of this project such as Tweet's text after preprocessing and polarity score. After they are retrieved from Twitter, Tweets are stored in a database known as MongoBD(MBD) database. Storing and retrieving Tweets to and from database will be done using Vim. In Addition, manipulating the Tweets and their attributes functionality should be supported by the DB interfaced that will be designed. Also, when the user issue a query, the system should look for already available Tweets in the to be retrieved and evaluated along with the new Tweets that were retrieved from Twitter.

We need a way of preventing as much noise as possible from being collected. The solution is filtering, which is functionality provided by the API.

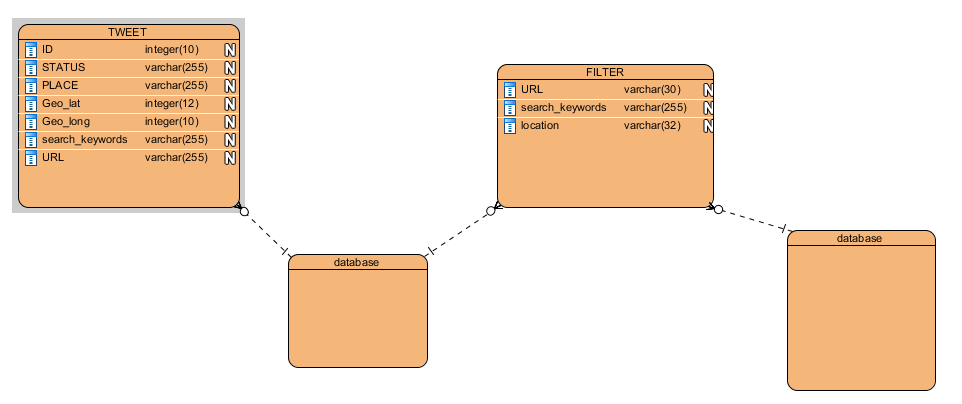


Figure5.2Entity relationship diagram (ERD)

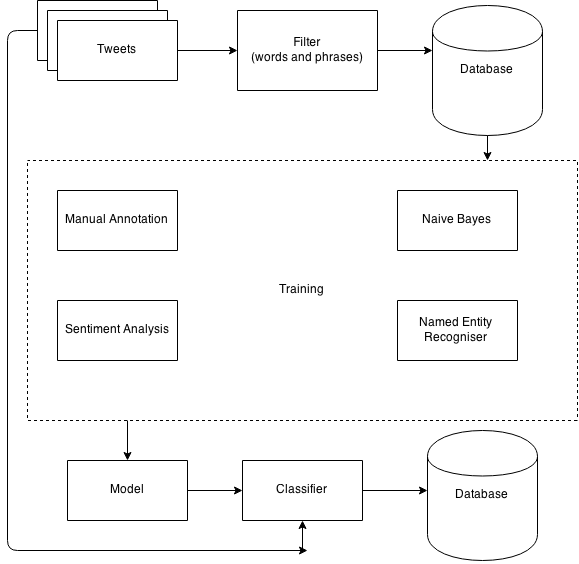


figure 5.3 The architecture design of the system

## 5.2 Storing Tweets.

Once we have got the Streaming API set up, we need to store the collected Tweets. As previously mentioned, for this project we use the MongoDB database. MongoDB is a non-relational database, based around what are referred to as documents. It differs from a traditional relational database in that instead of inserting rows into a table to represent items, we insert documents. Documents are simply optimised JSON files, known as BSON. The main advantage of this is that every entry in the database does not need an entry for each field. I.e. one entry can have a value for a “name” field, while other entries in the database wouldn’t even need to have a “name” field let alone an entry for it. You can also add new fields at will. MongoDB is a good choice for our database, because the Twitter API returns Tweets in the JSON format. This means that entering data returned by the API into our database is seamless. It also means that later when we wish to start using the Named Entity Recognisers to label data, we can add extra fields to data that matches a specific NER, without having to add an empty version of that field to every single piece of data we own.

## 5.3 Tweets Pre-processing.

Attributes of Tweets are beneficial when the Tweets are processed. They can be used to filter Tweets e.g. distinguishing Tweets that are written in English from those written in other languages. Furthermore, Spam Tweets such as Tweets with a URL only or hashtags only should be identified and therefore excluded from the Tweets to be processed. Tweets that are processed need to be cleaned from @mentions, #tags that are not necessary and Twitter specific abbreviations such as RT(retweet) and PRT(previous retweet). Also, URLs are not useful for sentimental analysis and should be removed prior to processing. There are multiple issues with Tweets to be solved when a Tweet is normalised. Han and Baldwin mentioned some of them including duplicating letters within words e.g. gooooood, typographical errors, Slangs and abbreviations. They can be found easily using a POS tagger with <unk>(unknown) option or using a spellchecker . Furthermore, emoticons contains strong sentiments and they should be either normalized or used for sentimental evaluation

## 5.4 Sentiment Evaluation Methods.

There are many methods that can be used in evaluating sentiments of Tweets. Firstly, using baseline classifier which classify Tweets depending on the occurrence of previously known words that have some given sentiments. This baseline classifier might be implemented based on Multilabel classification. This classifier can be improved by employing modifiers and negations which might increase or decrease polarity scores. Also, different lexicons can be used noticing their effectiveness. For both methods, there are some prerequisites such as tokenization and splitting sentences. Two corpora are required for implementing such classifier.

The main design approach will be using different evaluation methods on Tweets and presenting the results from each method to the user. Otherwise, the user should be able to choose the method prior to processing because different methods give better performance depending on the context of the Tweets used and the way each method is designed. Also, the average of evaluation for all of them can be presented to the user who can then decide of the accuracy of these methods. There should be statical results of the number of Tweets in each polarity class i.e. positive, negative and neutral.

# Reference..

* J. Berger, “Arousal increases social transmission of information,” Psychological science, vol. 22, no. 7, pp. 891–893, 2011.
* Borgatti, S. P. 2009). 2-Mode concepts in social network analysis. Encyclopedia of Complexity and System Science, 8279–8291
* (Esmin, A. A. A., Jr., R. L. D. O., & Matwin, S. (2012). Hierarchical Classification Approach to Emotion Recognition in Twitter (pp. 381–385). Presented at the 2012 11th International Conference on Machine Learning and Applications, IEEE.

<http://doi.org/10.1109/ICMLA.2012.195>)

* Jurek, A., Bi, Y., & Mulvenna, M. (2014). Twitter Sentiment Analysis for Security Related Information Gathering (pp. 48–55). Presented at the 2014 IEEE Joint Intelligence and Security Informatics Conference, IEEE. <http://doi.org/10.1109/JISIC.2014.17>
* Alsaleh, S., Nayak, R., & Xu, Y. (2011). Finding and Matching Communities in Social Networks Using Data Mining (pp. 389–393). Presented at the International Conference on Advances in Social Networks Analysis and Mining, IEEE. <http://doi.org/10.1109/ASONAM.2011.90>
* Lee, K., Palsetia, D., Narayanan, R., Patwary, M. M. A., Agrawal, A., & Choudhary, A (2011). Twitter Trending Topic Classification (pp. 251–258). Presented at the 2011 11th IEEE International Conference on Data Mining Workshops, IEEE. <http://doi.org/10.1109/ICDMW.2011.171>
* Keretna, S., Hossny, A., & Creighton, D. (2013). Recognising User Identity in Twitter Social Networks via Text Mining (pp. 3079–3082). Presented at the IEEE International Conference on Systems, Man, and Cybernetics, IEEE <http://doi.org/10.1109/SMC.2013.525>
* K. Lerman and R. Ghosh, “Information contagion: An empirical study of the spread of news on digg and twitter social networks,” in Proceedings of the Fourth International Conference on Weblogs and Social Media, Association for the Advancement of Artificial Intelligence. AAAI Press, Menlo Park, California, 2010.
* Mohammad Al-Fayoumi, Soumya Banerjee, Jr., and P. K. Mahanti, “Analysis of Social Network Using Clever Ant Colony Metaphor”, PROCEEDINGS OF WORLD ACADEMY OF SCIENCE, ENGINEERING AND TECHNOLOGY VOLUME 41 MAY 2009 ISSN: 2070-3740.
* W. McKinney, “Python for Data Analyses,” Published by O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, in 2013, ISBN: 978-1-449-31979-3.
* “Tweepy, An easy-to-use Python library for accessing the Twitter API,” [Online]. Available: http://www.tweepy.org. [Accessed: 18-11-2018].
* Twitter, About Twitter [Online], URL: <https://about.twitter.com/> company, [Accessed 18 November 2018]
* Adam Green, Aggregating tweets: Search API vs. Streaming API, [Online],URL: <http://140dev.com/twitter-api-programming-tutorials/aggregating-tweets-search-api-vs> streaming-api/, [Accessed 10 November 2018]
* The Stanford Natural Language Processing Group. “Stanford CoreNLP: A Suite of Core NLP Tools”. (2013). Available at: <http://nlp.stanford.edu/software/corenlp.shtml> [Accessed on 9th November 2018]
* Socher, S. et. al. “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank”. (2013). Available at: <http://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf> [Accessed on 11th November 2018]

# APPENDICES

## APPENDIX 1

**Code to stream for tweets.**