2020

A Public complaints data mining and visualization tool for social media: a case of Nairobi City County.

Olweru, Stephanie
Faculty of Information Technology
Strathmore University

A Public Complaints Data Mining and Visualization Tool for Social Media: A case of Nairobi City County

By

Stephanie Olweru

100933

A Thesis Submitted to the Faculty of Information Technology in partial fulfilment of the requirements for the award of Master of Science in Information Technology.

Master of Science in Information Technology

Strathmore University

April 2020

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Name: Stephanie Olweru
Signature: …….
Date: .....

Approval

The thesis of Stephanie Olweru was reviewed and approved by the following:

Dr. Vincent Omwenga,
Faculty of Information Technology,
Strathmore University

Dr. Joseph Orero,
Dean, Faculty of Information Technology,
Strathmore University

Dr. Bernard Shibwabo,
Director of Graduate Studies,
Strathmore University

Abstract

The government of Kenya has been experiencing low levels of public complaint feedback with regards to the current complaint reporting methods. Citizens consider the organized public meetings and office visits as time consuming and opt out of the complaint reporting process. However, citizens have been taken to complaining on social media to express their dissatisfaction with infrastructural service delivery by government. Recent efforts to improve low levels of interaction between governments and their citizens by mining data from social media have proven that citizen sourcing from social media is a suitable approach to solving this issue. Social media posts, comments, likes, favourites, shares, retweets and similar features are clear indicators of public feedback if the posts are related to public service. This study proposed the development of a tool that extracts and visualizes public complaints relating to Nairobi county from social media. Roads were the focus infrastructure category for the study. For classification of collected tweets, a Support Vector Classifier (SVC) with TF-IDF features was trained and tested using labelled tweets. The classifier was able to label tweets collected with an accuracy of 77.52%. Location information was retrieved from the tweets classified as complaints using the Stanford Core NLP named entity recognizer. The locations identified in the complaint tweets using the Stanford NER were reverse geocoded using the Google Geocoding API and the resulting geographic coordinates plotted on a static google map. Frequent words in the complaint tweets were represented using a WordCloud. The county government can use this information for decision making, planning and to give feedback to the public on resolution of the same.

Keywords: Public complaints; Social media; Data mining; Natural Language Processing;

Table of Contents

List of Tables

List of Figures

List of Abbreviations

API- Application Programming Interface

JSON-JavaScript Object Notation

NLP- Natural Language Processing

NLP- Natural Language Processing

OAuth- Open Authentication

RAD- Rapid Application Development

RFID-Radio Frequency Identification

TAM- Technology Acceptance Model

Chapter 1 Introduction

1.1 Background

Owing to the challenges of managing centralized service delivery, governments worldwide have moved towards decentralized governance in the past quarter century (Ahmad, Devarajan, Khemani, & Shah, 2005). Centralized governments were failing to deliver basic services including education, health, water and sanitation (Ahmad et al., 2005). The promulgated constitution of Kenya 2010 brought with it devolution from a centralized government to county governments. Functions relating to the provision of infrastructural services which were previously under the mandate of the national government were cascaded down to county administration. This was to allow for improved service delivery to citizens at lower administrative levels (Government of Kenya, 2010).

Urban areas lie within the jurisdiction of the counties in which they reside. In this regard, the constitution is supported by the County Government Act 2012 and the Urban areas and cities act 2012. Residents of a city, town or municipality have the right to take part in the governance of the urban areas or cities in which they reside (Government of Kenya, 2012). Urban areas are classified into cities, towns, municipalities and market centers based on the population and facilities that are available in the areas (Government of Kenya, 2012). Nairobi is the capital city of Kenya, also identified as Nairobi county. Nairobi county government is responsible for providing the following infrastructural services to its residents; Physical Planning, Social Services and Housing, Public Health, Inspectorate Services, Primary Education Infrastructure, Environment Management, Garbage Collection, and Public Works (UK AID, 2015).

However, the Nairobi county government is yet to satisfactorily address provision of these services according to its residents (Kibanya, 2015). Nairobi has experienced flooding during rainy seasons due to poor drainage, inadequate garbage disposal mechanisms, poor roads in residential areas, poor street lighting, inefficient public health services, inadequate education infrastructure and water rationing. This is despite the fact that Nairobi county has been allocated the highest budget among all counties in the years following devolution up to the financial year 2017-2018 (Republic of Kenya,National Treasury, 2018).

Residents rely mainly on organized meetings or visits to the county offices to have their complaints regarding these infrastructural issues addressed. Complaints are defined as statements about expectations that have not been met (Aziz, 2015). Complaints have also been defined as dissatisfaction with products or services which occurs when an organization fails to meet customer expectations (Trappey, Lee, Chen, & Trappey, 2010). In the citizen-government scenario, the citizen is the customer and the government is the client. These complaints made by citizens regarding the delivery of public services by government are therefore known as public complaints.

With regards to the current complaint reporting methods, a great number of citizens consider the current mechanisms time consuming. They believe this time would be better spent trying to make a living for themselves (Ronoh, Mulongo, & Kurgat, 2018). This has resulted in low levels of citizen interactions with the government. This translates into the government being unaware of the needs of their citizens and therefore not fulfilling these needs as required (Aziz, 2015).

1.2 Problem Statement

Nairobi county government currently relies on organized public meetings with citizens and citizen visits to county offices to enable citizens notify the government of their infrastructure related complaints. These methods are considered time consuming and inconveniencing by citizens who then opt out of the complaint reporting process. This has resulted in low levels of public complaint reporting which translates into the government being unaware of the concerns of its citizens and therefore not being able to address them. As much as provision of infrastructural services is the responsibility of the government, the citizen is the ultimate driver of their own development by holding the government accountable through civic engagement (Ahmad et al., 2005).

Civic technology refers to the use platforms and applications that enable citizens to connect as well as collaborate with each other and the government. An example of civic technology is SeeClickFix, a web and mobile application that enables citizens communicate non-emergency issues to their governments (SeeClickFix, 2020). Another example of civic technology is governments use of social media like Twitter and Facebook to communicate and interact with citizens (Mergel, 2013). According to a study by Pade-Khene, Thinyane and Mwazvita (2017), civic technology has significantly improved levels of interactions between governments and their citizens. This is because civic technology platforms are easy to use, offer increased transparency and accountability

(Pade-Khene et al., 2017). The use of mobile technology based social media applications support the interactive engagement between parties that was once considered impossible without physical interactions. Smartphone penetration in Kenya is also widespread with all these users accessing the internet on their phones (Jumia Kenya, 2015). This greatly facilitates the use of social media as a platform to enable citizens raise complaints related to their living areas. For example, citizens often take to social media to voice their complaints related to infrastructure within their living areas. This data can be collected and analysed to enable county administration to facilitate planning and resolution of the same.

This study proposes the development of a tool will facilitate the extraction and visualization of public complaints relating to Nairobi county from social media. This will assist the county government extract complaints from social media to support communication of complaints from citizens to the government.

1.3 Aim

The aim of this study is to collect and derive useful information on infrastructure related public complaints for Nairobi county government through the use of a tool that extracts and visualizes data sourced from citizens on social media using data mining techniques.

1.4 Specific Objectives

(i) To evaluate existing frameworks relating to public sector complaint handling.

(ii) To analyse the current models of social media interactions in the public sector.

(iii) To analyse the current techniques, approaches and architectures used for data mining aspects in social media

(iv) To develop a domain specific tool for extraction and visualization of public complaints.

(v) To validate the functionality of the developed tool.

1.5 Research Questions

(i) What are the existing complaint handling frameworks relating to the public sector?

(ii) What are the current models being employed in the public sector to facilitate social media interactions?

(iii) What are the current techniques, approaches and architectures used for data mining aspects in social media?

(iv) How can a new system be designed and developed?

(v) How can the functionality of the proposed tool be validated?

1.6 Justification

Public complaints reporting is a citizen's civic right and a key facilitator of government accountability. This translates to better service delivery with regards to the infrastructure related public services handled by county governments. Well-functioning infrastructural facilities improve the livelihoods of citizens. Kenya currently relies on organized meetings and visits to county offices to facilitate complaint reporting by citizens. The Nairobi county government lacks a platform to facilitate the transfer of complaints regarding county services from citizens to the county government that is accessible to citizens at all times. Citizens often take to social media to voice their complaints relating to their living areas. This research attempts to analyse how social media can be used to extract and visualize the complaints relating to the delivery of public services in Nairobi county. This ensures that the government can effectively rate the performance of its county services and plan resolutions taking into consideration the input from citizens. Citizens also have the opportunity to contribute toward sustainable living within their communities by contributing towards the improvement of public services. In addition, this research forms a basis for future research on citizen complaint mining, analysis and visualization for resolution planning.

1.7 Scope and Limitation

This study is restricted to Twitter as a social media platform. The data mined is that which is restricted to the mandate of the Nairobi County government with roads being the focus infrastructure category for this study. The study relies mainly on citizen sourcing as a technology for data mining on social media.

Chapter 2 Literature Review

2.1 Introduction

This chapter covers the relevant literature and previous studies relating to public complaints. It covers frameworks relating to complaint handling in public service, civic technologies, data mining on social media and previous studies. This chapter will conclude with a description of the research gap identified which is to be covered in this study.

2.2 Theoretical Frameworks

2.2.1 Technology Acceptance Model

The Technology Acceptance Model (TAM) was developed by Davis (1986). TAM explains computer use behaviour based on two factors: perceived usefulness and perceived ease of use. Perceived usefulness is defined as the user's subjective likelihood that using a specific application system will enhance his or her job or life performance. Perceived ease of use is the degree to which the prospective user expects the target system to be effortless. In TAM, a user's belief towards a system may be influenced by other external variables. These external variables which may have an effect on system usage behaviour include user training, user participation in the development of the system, the nature of the implementation process and system characteristics (Venkatesh & Davis, 1996). Figure 2.1 illustrates the TAM model.

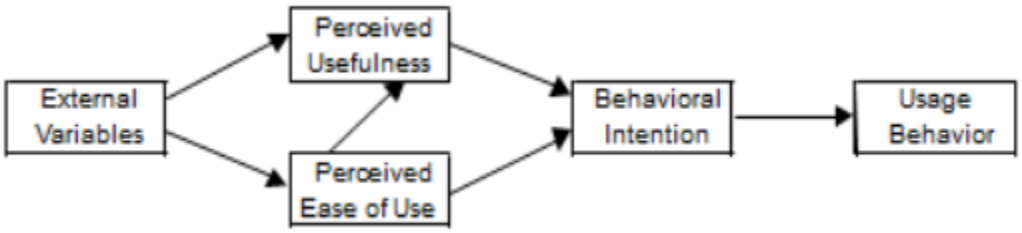


Figure 2.1: The Technology Acceptance Model (Venkatesh & Davis, 1996)

TAM has been widely in previous research to predict the acceptance and use of information technologies by users. A study by Shafeek (2011) attempted to use TAM to evaluate the acceptance of eLearning systems by teachers. Zhou, Dai, and Zhang (2007) developed a new model called online shopping acceptance model (OSAM) based on TAM to study online shopping behaviour.

A study by Pavlou (2003) developed a model for predicting the acceptance of e-commerce by adding the external variables trust and perceived risk in TAM. Müller-Seitz, Dautzenberg, Creusen and Stromereder (2009) used the Technology Acceptance Model to understand how with security concerns affect user acceptance of Radio Frequency Identification (RFID).

Perceived ease of use has been shown to be key in IT acceptance in general. However, a study by Davis (1989) indicates that perceived usefulness directly links to increased acceptance of systems as compared to ease of use. This means that users are more likely to accept a system simply because of the functions that it provides them. As much as difficulty in use can affect the usage of an efficient system, ease of use cannot make up for a system that does not perform its intended functions. This means that system designers and system developers should not overemphasize ease of use at the expense of perceived usefulness of the system in designing and implementing successful systems (Davis, 1989).

TAM is key when designing a data mining tool for extraction and visualization of public complaints on social media. This theory facilitates an understanding of the relationship between perceived usefulness, perceived ease of use and acceptance of information technologies by users. Perceived usefulness and perceived ease of use should be included in the design of a system to promote user acceptance Lack of acceptance of systems by users is an indication of system implementation failure. However, perceived usefulness contributes more to the increased acceptance of systems as compared ease of use. This is because ease of use cannot make up for loss of perceived usefulness in a system. This study should ensure that the design of the tool majorly supports perceived usefulness to ensure its acceptance by the intended users.

2.2.2 Participative Management Theory

Participative management refers to a process in which subordinates share a considerable level of decision making powers with their leaders (Shagholi, Abdolmalki, & Moayedi, 2011). Rensis Likert came up with the participative decision making theory which states that highly productive leaders usually involve their subordinates in the decision making process (Likert, The human organization: Its management and value, 1967). Based on this finding, Likert came up with four unique management styles discussed below.

i. Exploitative authoritative management

In this style, the leader has absolutely no trust in the subordinates and their decision making capabilities. Decisions are simply made at the top of the hierarchy and imposed on the subordinates. Communication is unidirectional and subordinates are managed by fear (Likert, The human organization: Its management and value, 1967).

ii. Benevolent authoritative management

With this management style, the leader believes that decision making should be done solely at the managerial level. Subordinates are expected to comply with the leader's decision as the leader is considered as the appropriate decision maker. Communication is unidirectional and subordinates are motivated using rewards (Likert, The human organization: Its management and value, 1967).

iii. Consultative management

In consultative management, the leader laces some level of trust in their subordinates. However, the leader does not fully trust the judgement of the subordinate to make decisions on their own. The leader therefore seeks input from them and uses their input to come to a decision. Communication is therefore bidirectional and subordinates are motivated by the feeling of inclusion (Likert, The human organization: Its management and value, 1967).

iv. Participative management

In this style, decision making is spread across the organizational hierarchy. Leaders fully trust that their subordinates can make appropriate decisions. Subordinates are therefore involved in the decision making process and their inputs directly influence the decision made by the leader.

Communication is bidirectional and subordinates are highly motivated by the trust placed on them (Likert, The human organization: Its management and value, 1967).

Participative management results in better decision making and greater efficiency (Likert, 1967), identification and solving of problems (Blase & Blase, 2001) and open communication (Blase & Blase, 2001). These advantages have seen participatory management being incorporated in public administration in the governance interactions with their citizens (Likert, 1981). Participative management theory is therefore key to providing a guideline in the development of a tool to extract and visualize public complaints.

2.3 Frameworks Relating to Complaint Handling in Public Service

Farrell (2010) analysed two frameworks that can be implemented in the domain of public complaint handling; the 'ladder of participation and involvement' and 'voice, choice and exit'. Both frameworks intend to illustrate the involvement of citizens in public service delivery by detailing the level of involvement and influence of citizens when it comes to voicing their concerns and complaints (Aziz, 2015).

The ladder of participation defines the level to which citizens can participate in service delivery. The ladder was developed by Arnstein (1969) and divides participation into levels. The lowest level is 'Non participation'. It comprises of 'Manipulation' and 'Therapy'. The aim on non-participation is to educate the public to facilitate public support. The next level is 'Tokenism'. Tokenism comprises on 'Informing', 'Consultation' and 'Placation'. With tokenism, citizens are consulted but they lack the power to alter the status quo. The final level is 'Citizen control'. Citizen control comprises of 'Participation', 'Delegated Power' and 'Citizen Control'. Partnership allows citizens to negotiate and be involved in decision making. Delegated power and citizen control give citizens the power to make decisions. Figure 2.2 illustrates the ladder of participation as described by Arnstein.

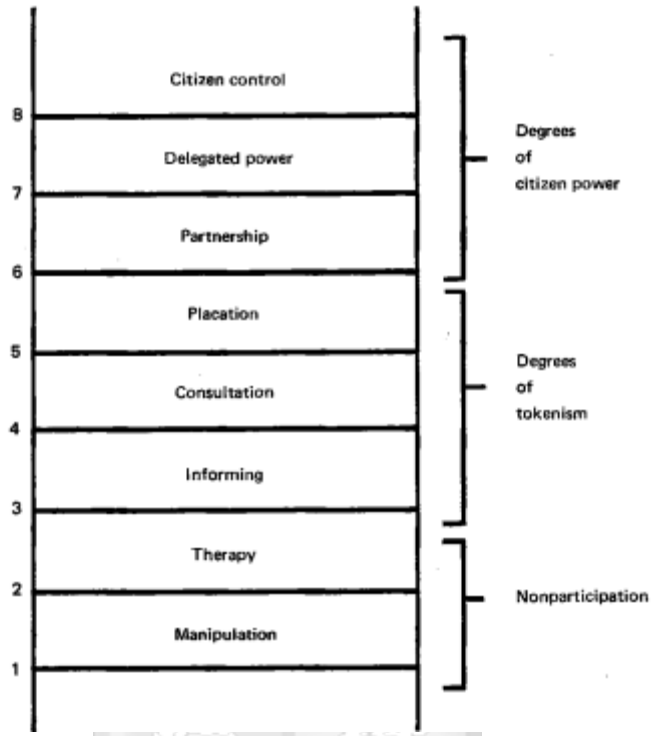


Figure 2.2: The Ladder of Participation (Arnstein, 1969)

The 'Voice, Choice and Exit' framework is used to review the level of power that users of a public service have. Hirschman (1970) came up with the framework which focuses on the customer's choice when faced with dissatisfactory services; voice to make public administration more responsive to their needs or exit. Voice is described as an attempt to change an unfavourable state of affairs by individually or collectively petitioning management with the intention of enforcing a change (Hirschman, 1970). The same has been applied when dealing with public service complaints. Citizens who are dissatisfied with public service delivery may choose to voice their complaints and opt for exit options if the complaints are not satisfactorily addressed, for example, vote to change the current leadership. Hirschman concludes that the citizen should express their concerns to the government so that the government is aware of their wants, however, decision making should be left to government (Hirschman, 1970).

2.4 Civic Technology

Civic technology refers to the use platforms and applications that enable citizens to connect as well as collaborate with each other and the government (Saldivar et al., 2018). This area has been

continuously growing at an estimated rate of 23% between 2008-2012 (Lukensmeyer, 2017). Civic technology is mainly focused on two themes open government and community action.

Open government platforms focus on enhancing transparency by giving the public access to government data. Activities that are handled by these platforms include; transparent access to data as well as analysing, visualizing and mapping of provided data to give resident feedback on how to improve service delivery.

Community action platforms major on collaboration. They facilitate peer-to-peer sharing of information, civic crowdfunding and crowdsourcing to address civic issues. Activities here are more involving; citizen crowdsourcing to fund civic projects, organizing communities to manage civic campaigns, crowdsourcing information from residents in order to inform and act on issue, promoting peer to peer sharing of goods and services and neighbourhood collaborations (May & Ross, 2018). Figure 2.3 illustrates the civic crowdsourcing cycle between governments and citizens.

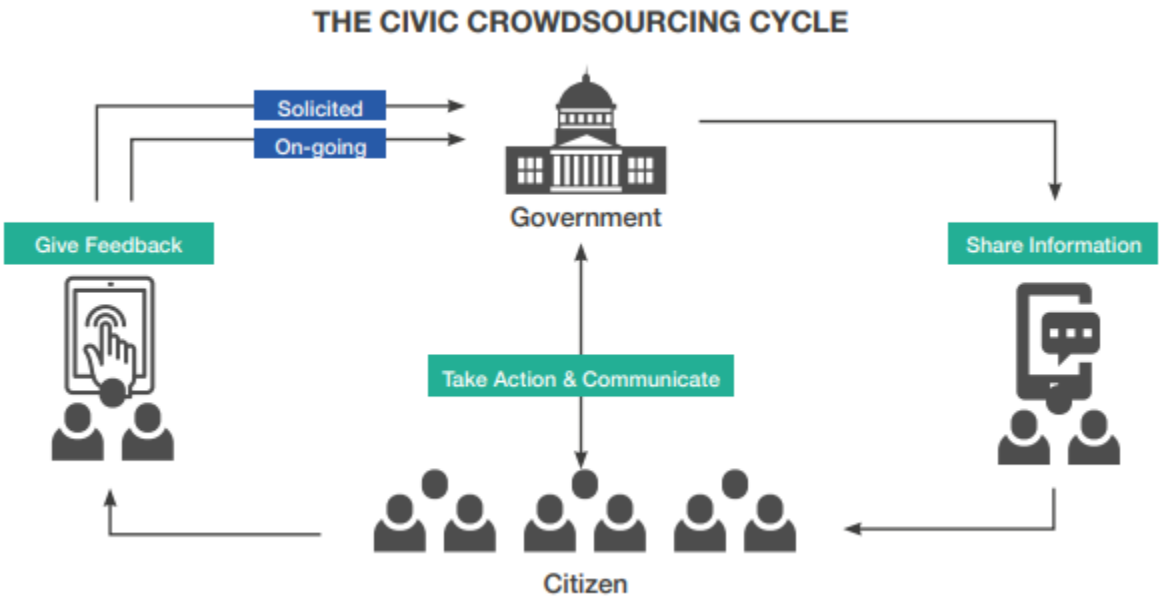


Figure 2.3: The Civic Crowdsourcing Cycle (Ghorayeb & Soule, 2017)

2.4.1 Citizen Sourcing as a Civic Technology Mechanism

Citizen sourcing is a field within crowdsourcing. Crowdsourcing refers to a business solution that harnesses the power of distributed individuals to gather creative solutions or fresh ideas from a

large number of individuals (Leukis, 2018). Citizen sourcing is the public service branch of crowdsourcing that is concerned with "the act of taking a task that is traditionally performed by a designated public agent (usually a civil servant) and outsourcing it to an undefined, generally large group of people in the form of an open call" (Lönn, Uppström, & Nilsson, 2016, pp. 3-4).

According to Leukis (2018), governments embracing citizen sourcing have incorporated the use of information communication technologies to enable them tap into the vast knowledge base that is the public. This has been facilitated by the growing digitization of government systems and Web 2.0 technologies. Web 2.0 refers to the version of the worldwide web that focuses on the ability of people to collaborate and share information over the internet. The emphasis here leans more towards the web facilitating human interactions than the technology itself. These applications which are accessible via the web and on mobile phones, have been used by citizens for reporting infrastructure complaints using the functions of information communication technology.

2.5 Social Media as a Citizen-Sourcing Platform

Social media refers to a group of internet- based applications which incorporate the use of Web 2.0 technologies to enable users to communicate, collaborate and share content, ideas and information across the globe (Susilawati, 2016). Examples of social media applications include Facebook, Twitter, Instagram, YouTube, Flicker and Snapchat. Facebook and Twitter are the two most popular social media networks (Hu, 2019). Social media analytics refers to the process by which data is gathered, analysed and visualized to support insightful for decision making.

Mergel (2013), conducted a study to evaluate how governments are using social media to facilitate interactions with the public.The result was a framework that models government interactions on social media using three aspects namely, transparency, participation and collaboration. Transparency refers to providing the public with information regarding what their government is doing. Governments are constantly releasing information via newsfeeds on their accounts which the public has access to. The goal of transparency is voluntarily provide information to reach a maximum number of citizens. The main measure here is the number of views or likes on a post and the number of followers on a government account (Mergel, 2013).

Participation which is the second approach focuses on citizen engagement. Participation goes beyond the government providing access to information to the public to them diligently seeking

feedback from the public through their social media platforms.The feedback is then used to better the final decision or policy. Participation is incorporated via specific social media analytics methods (Mergel, 2013). Other parameters are monitoring retweets on content posted by government and mentions related to government social media pages.

Collaboration gears towards maintaining a reciprocated relationship between governments and citizens to enable citizents directly interact with the government and contribute towards governance innovation. A passive strategy is preferred by most social media managers in the implementation of collaborative measures.Collaborative engagement can be monitored via direct messages and chats between both parties (Mergel, 2013). Table 2.1 describes the framework for Government social media interactions as described by Mergel.

Table 2.1: Framework For Measuring Social Media Interactions in Government (Mergel, 2013)

Mission	**Goal**	**Tactics**	**Social media Mechanisms**	**Outcome**
Transparency	Information education	One-way push	• Number of followers & likes/friends (change from start) • FB likes • Twitter followers • Unique visits to blog • Time spend on page <30 • Visits only home page • Views on YouTube & Flickr • "Read more"	Accountability trust
Participation	Engagement	Two-way pull	• Click-throughs from social media sites • Reach: demographic	Consultation, deliberation, satisfaction

			data (gender, location, cities) • Bookmarking & digging content • Twitter retweets, hashtags • Posting ratings & reviews • Spend more than 1min on site • Comments on blog & Facebook • Ratings on YouTube • Number of links & trackbacks • Frequency of check-ins on Foursquare	
Collaboration	Cross boundary action Two-way interactive	Networking Co-design of services	• Request for membership in a LinkedIn group • Subscriptions to blog, YouTube channel • Facebook shares • Twitter direct messages • Creating their own content • Downloads of videos, documents	Community building Creation of issue networks

			• Conversations • Volunteering, donations • Offline actions	

2.5.1 Twitter

Twitter has been noted as one of the best platforms to support social media analytics (Fatkhurrochman, Noviandha, & Setyanto, 2018). Twitter is a popular microblogging tool that has been growing steadily in popularity since it was launched in October 2006. As of 2017, Twitter allows users to post messages known as Tweets that are a maximum of 280 characters long. Twitter users can follow other users to receive tweet updates from them, mention other users in tweets, reply to tweets and retweet other user's tweets. These actions above can capture the interactions among Twitter users (Mergel, 2013). All of these attributes of Twitter make it a great platform for researchers to collect and conduct big data research. Twitter has an estimated 330 million active users per month and over 500 million tweets per day. Such magnitude of data has proved to be a great resource to organizations that are seeking to maintain a positive reputation in regards to their economic, social or political status (Jianqiang & Xiaolin, 2017).Twitter has been used by governments to facilitate better planning and decision making.

Twitter as a social media platform has a number of unique features that have led to its popularity:

(i) Limited message length. Tweets were limited to a maximum of 140 characters up until 2017. Users can currently post tweets that are up to 280 characters long. However, users are still maintaining the brevity that they were used to previously. Only 12% of tweets are longer than 140 characters (Rosen, 2017). This makes the data on Twitter short, precise and comparatively easy to process.

(ii) Mentions. Twitter makes use of the @mention feature that allows a user to make directly communicate with another Twitter user whether they follow that account or not. This gives citizens direct access to government offices and officials (Garg et al., 2017).

(iii) The tweet data can be searched via keywords as hashtags, denoted by the "#" sign.

2.6 Data Mining on Social Media

2.6.1 Data Mining on Twitter

Tweets can be accessed by developers with registered accounts using two APIs; the Search API and the Streaming API. Access to the twitter APIs is supported via Open authorization (Oauth). The search API allows for queries on popular and recent tweets. It allows for searching of tweets dating back to 7 days. The Streaming API uses one connection to return public tweets based on a given set of filters in JSON (JavaScript Object Notation) format.

However, both the standard Twitter APIs are limited to accessing only tweets that are 7 days old or less. Access to earlier tweets can be obtained on premium or enterprise developer accounts which come at a cost. A study by Jefferson Henrique provided an open source code that can be used to extract data from Twitter beyond the 7day limit in the standard Twitter API account (Henrique, 2017). A improvement to Jefferson Henrique's code was developed by Dmitry Mottl to fix known bugs and ensure compatibility with Python version 3 (Mottl, 2019). Tweets were collected from posts mentioning the Nairobi County Government official page '@NairobiCityGov' and from tweets mentioning 'Nairobi county' in combination with the words 'road', 'roads' or 'rd'.

2.7 Natural Language Processing

Natural language processing (NLP) is a field of computer science and artificial intelligence which incorporates the use of algorithms and techniques to enable interactions between computers and human (natural) languages. Specifically, it is the process of a computer extracting meaningful information from natural language input and/or producing natural language output (Batrinca & Treleaven, 2015). Language is defined as a set of rules or symbols. NLP programs must be able to understand the structure of the language comprising of what words are and how they are combined into sentences. This is especially useful in regards to social media data which is usually unstructured in its raw format.

NLP can be classified into Natural Language Understanding and Natural Language Generation (Khurana, Koli, Khatter , & Singh, 2017). Natural language understanding is concerned with mapping a given input in natural language into useful internal representations for example converting natural language into a SQL query. Natural language generation is concerned with

producing meaningful phrases and sentences in the form of natural language from some internal representation, for example deriving information that is of interest in Natural language from data. This study will focus on Natural language understanding with an aim to process and derive insightful information from complaints relating to Nairobi county roads sourced from Twitter.

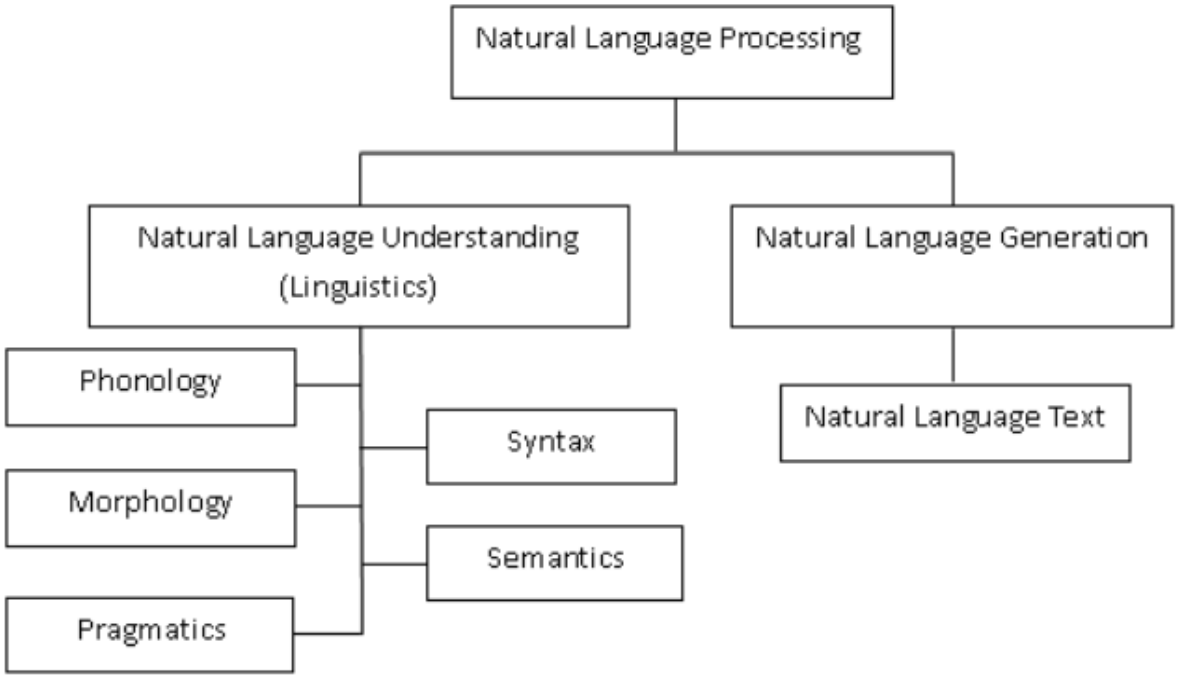


Figure 2.4 Classification of Natural Language Processing (Khurana et al., 2017)

2.7.1 Natural Language Understanding Processing Steps

NLP programs must be able to understand the structure of the language comprising of what words are and how they are combined into sentences (Maldonado, Alulema, Morocho, & Proaño, 2016). This is especially useful in regards to social media data NLP provides components for language structure as pragmatic, morphological, syntactic, phonological, discourse and semantic (Khurana et al., 2017).

2.7.1.1 Phonological Analysis

Phonological relates to the use of sounds in a language. These sounds are systematically arranged to derive meaning. English makes use of about 45 phonemes also known as contrastive sounds, for example /p/ and /b/ are contrastive because pan and ban have different meanings (Briscoe, 2010).

2.7.1.2 Morphological Analysis

Morphological is concerned with lexical knowledge. Words are constructed from basic units called morphemes. For example, the word send is atomic but the word resend is made up of the morphemes re and send. The interpretation of a morpheme is the same across all words. A good example is ed which when added to a verb signifies an action in the past tense. Parts of speech are assigned to each word at this stage too (Khurana et al., 2017).

2.7.1.3 Syntactic Analysis

Syntactic relates to how words are organized to construct meaningful and correct sentences. Both grammatical analysis and parsing are carried out at this level. The order of words in a sentence could portray different meanings. For example, “The snake bit the goat.” and “The goat bit the snake.” convey a different meaning owing to the position of words (Khurana et al., 2017).

2.7.1.4 Semantic Analysis

Semantic refers to how morphology and syntax are combined to derive meaning from a sentence (Briscoe, 2010). Semantic disambiguation derives possible meanings from a sentence by associating the parts of speech within it. For example, the word nail can either mean a body part but it is also a construction item.

2.7.1.5 Discourse Analysis

Discourse focuses on the properties of text that conveys meaning by analysing and making connections between component sentences. The meaning of a sentence depends on the meaning just before it. It also affects the meaning of the sentence preceding it. This is done via Anaphora resolution which replaces the words such as pronouns, which are semantically stranded, with the entity which they refer to and discourse/Text Structure Recognition which sways the functions of sentences in the text thereby adding to the meaningful representation of the text (Khurana et al., 2017).

2.7.1.6 Pragmatic Analysis

Pragmatics is concerned with the use of language in context and how different contexts affect the meaning of sentences. The output of this level is represented as a structural dependency of words within the sentence.

2.8 Information Extraction in Natural Language Processing

Information extraction is the process by which relevant information is obtained from unstructured text. The information that needs be extracted is predefined as per the information use interests and is made up of entities, relationships between entities and events. According to the Message Understanding Conference, five tasks have been defined as key to executing the process of information extraction. These tasks are named entity recognition, coreference resolution, template element, template relation and scenario templates. Of these, named entity recognition and information extraction are the most important tasks for the information extraction process.

2.8.1 Named Entity Recognition

Named entity recognition is the process by which entity mentions in text are identified and classified according to predefined classes (Anggareska & Purwarianti, 2014). Entity mentions represent some form of real life objects. For example, nouns, names and pronouns.

Named entity recognition can be done using two methods: rule based approach and statistical approach (Appelt, 1999). The rule based approach, relies on rules formulated by experts to identify named entities. This approach works really works for domain specific entity recognition. However, it does not perform well when the annotation is moved to a different domain (Appelt, 1999). Another disadvantage of this approach is that it takes a lot of time to manually annotate entities. The statistical approach is learning based and makes use of training data. The annotated training corpus is then run on a training algorithm and the results are used in a system for annotating future texts. The statistical approach is straightforward and portable across domains.

2.8.2 Relation Extraction

Relation extraction refers to the detection and classification of relationships between entities in a text (Anggareska & Purwarianti, 2014). Relation extraction can be performed through supervised methods and unsupervised methods. The most commonly used supervised learning method is feature based methods. The model works with a predefined set of relation examples, to derive syntactic and semantic features from text. The derived features are used to determine whether entities in the given text are related or not. Therefore, feature selection is key as this affects the accuracy of the results (Hasby & Khodra, 2013).

2.8.3 Information Filing

The result from the information extraction step were stored in a structured format. This will make it easier to query data and retrieve information. The data will contain identified entities and their relationships (Anggareska & Purwarianti, 2014).

2.9 Previous Studies on Social Media Citizen-Sourcing for Government

2.9.1 A Prototype for Mapping of Tweets On State Services for Decision Support: A Case of Huduma Kenya

The researcher mapped tweets on state service tweets to aid in decision support. The organisation under study was Huduma Kenya which is a "one stop shop" for service delivery for government services in Kenya. The main aim of this research was to show that tweets can be analysed based on keywords specific to Huduma services (Ng'ang'ira, 2018). Data was collected using the Twitter public streaming API. The dataset that was used was derived from Huduma Kenya official Twitter pages from all regions within the country. Using a wordlist, tweets were classified as negative and positive and mapped according to their corresponding tweet coordinates. The main objective was to map tweets according to their geo-coordinates (Ng'ang'ira, 2018). However, most twitter users do not enable geolocation on the application which means that this approach filters out a significant number of tweets in the mapping that could be useful in the data visualisation (Ng'ang'ira, 2018).

2.9.2 Real-Time Government Policy Monitoring Framework Using Expert Guided Topic Modelling

Majeed, Malik, Khan and Khalid (2016) conducted a study was aimed at streaming tweet data which was analysed to produce real-time analysis about the top government policies. The research monitored government policies through public tweets related to government policies on twitter with the intention of providing insights into public needs, foster interactions between the government and the public and driving policies according to public needs rather than the government's own approach. This research applied semi-supervised topic modelling where experts were consulted to enable the construction of a domain specific group of top words that normally represent a topic. A one-time activity was carried out to assign words to specific topics and a model was trained on a collection of documents containing government policies. This model was used to monitor streamed tweets regarding government policies on Twitter. NLP techniques are then applied with part of speech tagging of tweets to derive opinion from tweets with the help of an

external dictionary. At any given time, top five policies are visualised and they keep changing according to current events (Majeed et al.,2016). The main limitation of this research was that there was no visualization of identified topics by location.

2.9.3 Other Studies

The researchers conducted a study on the social media pages of municipalities of North Rhine-Westphalia, Germany with regard to the topic "open government" in support of e-government (Born, Mainka, Meschede, & Siebenlist, 2019). The study focussed on mining data from Facebook Twitter and YouTube. The data collection was done via each respective application's official supported API. Posts were collected together with their metadata, namely likes, comments, time, and shares. Data was analysed and visualised according to social media activities denoting citizen participation and topics. A corpus was developed based on literature reviews and interviews with municipality to identify terms associated with open government. The corpus was used to perform topic detection with Latent Dirichlet Allocation (LDA) as a method with the software MALLET (Born et al., 2019). The main limitation of this research was on the data retrieval limitations for Twitter, the collection was restricted to the last 3200 tweets of a user. This could result in leaving out of data that could potentially be crucial to the information retrieved after the analysis process.

2.10 Research Gap

Previous works have attempted to mine social media data to facilitate collection of citizen feedback with regards to government services. Based on previous studies, it is possible to mine and analyse data from social media. An analysis of current studies reveals that most of them do not focus on public service complaints. There is also lack of a mechanism for the government to analyse and visualize the data collected based on locations that are mentioned in text format.

This research aimed to develop a domain specific tool that extracts and visualizes public complaints within the mandate of the Nairobi county government from Twitter. The focus infrastructure category is roads. The data was visualised via a graphical interface where the county government can view which locations have the most complaints and using a WordCloud of frequent words appearing in the complaint tweets.

2.11 Conceptual Framework

The conceptual framework illustrates a simplified approach on how the complaints identified in this research shall be addressed. This study proposes the conceptual framework below to extract and visualize social media data with regards to infrastructural services in Nairobi county. Data is extracted from Twitter using the standard APIs. The results are visualised on a graphical user interface to make it easier for the county government to evaluate the complaints received to enable them plan service delivery improvements.

The conceptual framework is illustrated in Figure 2.5 below.

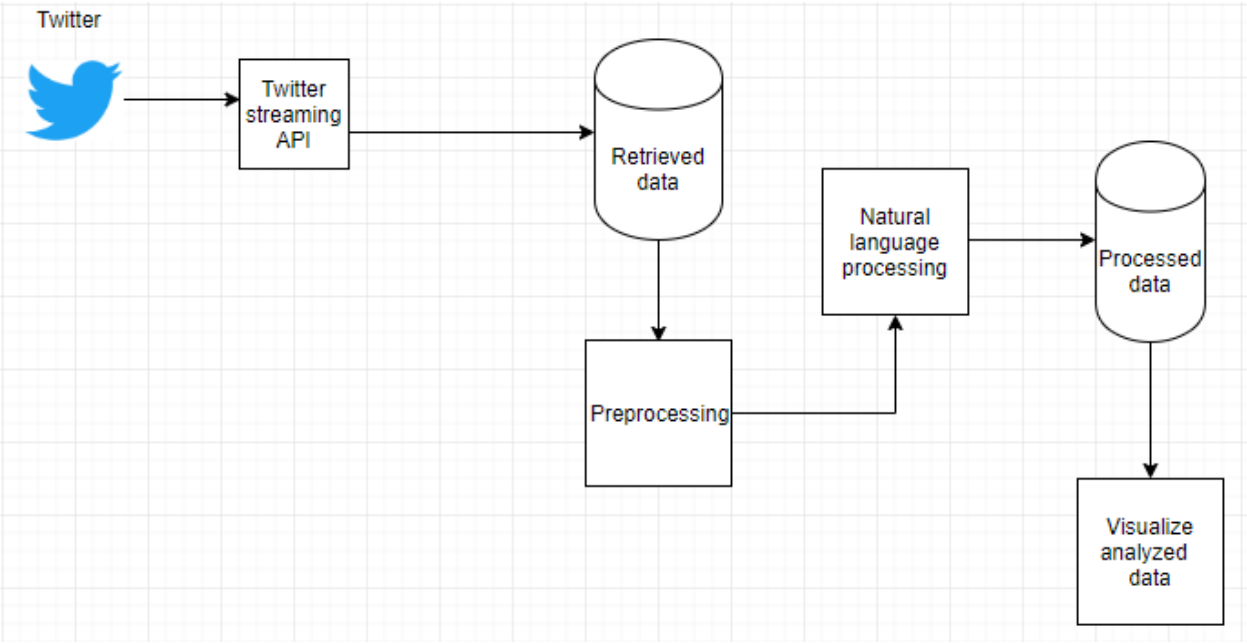


Figure 2.5 Conceptual Framework

Chapter 3 Research Methodology

3.1 Introduction

This research aimed to find out how public complaints extraction and visualization can be implemented via citizen sourcing on social media. Rapid prototyping system development methodology was used in conducting this study is described in this chapter. The target location, target population, sampling, data collection and analysis procedures, and programming tools are also discussed. This chapter concludes with a description of the research quality and ethical considerations.

3.2 Research Design

Research design refers to the overall approach that is used to conduct the research in terms of data collection, measurement and analysis. Research design guides in the methods that are most suited to collect, measure and analyse data (Kothari & Garg, 2014) . This research was conducted using an experimental design. The researcher formulated research objectives which guided how the research was to be conducted. The independent variable 'text' in the collected tweet was used to predict the dependent variable 'classification_id' of the tweet as either complaint, appreciation or neutral.

3.2.1 Prototyping

A prototype refers to an initial version of a system that is used to demonstrate concepts and design options in order to provide possible solutions to the problem at hand. Prototyping is iterative to ensure that costs are controlled and a feel of the proposed system is experienced by the stakeholders early enough in the development process (Sommerville, 2011). Any changes that are noted in the evaluated prototypes are incorporated in subsequent releases. The final prototype is accepted as the working system.

3.2.2 Rapid Application Development

This study adapted the Rapid Application Development (RAD) methodology. RAD is a set of methods designed to reduce software development costs while being flexible and adaptive to changing user needs. RAD was designed to deliver faster development of software. RAD is commonly implemented in four phases which are Planning, Analysis, Design and Implementation (Delima, Santosa, & Purwadi, 2017). This study implemented the system prototyping approach of

RAD. System development is done concurrently in the analysis design and implementation phases to ensure that the system is developed quickly. The first prototype of the system contains only minimal features that are needed. The prototype is analysed and the feedback is fed back into the analysis, design and re-implementation of the subsequent prototype. This is done until the final prototype is realized.

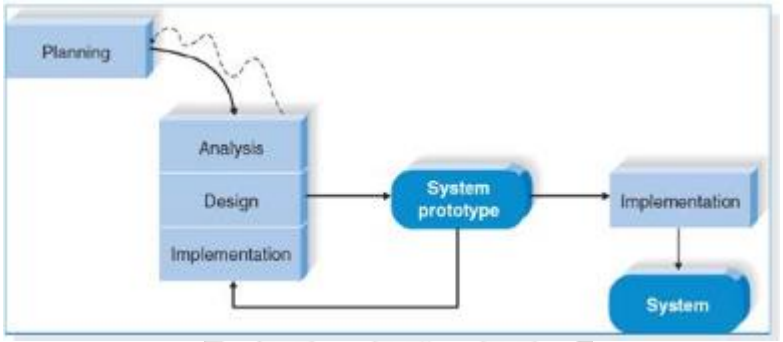


Figure 3.1: Phases of RAD Model Using Prototyping (Delima, Santosa, & Purwadi, 2017)

3.2.3 Planning

The main aim of this phase is to identify the problem that the study aims to solve. During the planning phase, the scope of this study was established, research objectives formulated and the feasibility of using social media as a public participation platform analysed.

3.2.4 Requirements Analysis

In this phase, the requirements for developing a public complaint tool were obtained from previous works and government publications. This was done in an aim to understand current frameworks relating to the handling public complaints, similar systems and the challenges being experienced in the use of these approaches. After gathering, the requirements were analysed and documented in the form of a context diagram.

3.2.5 Design

The structure and architecture of the tool were defined in this phase. Unified modelling language was used to represent the proposed system in form of, data flow diagrams, data models, entity relationship diagrams and system wireframes.

3.2.6 Implementation

After the design phase, the tool was developed using Python as a programming language due to its vast number of available libraries. The Stanford Core NLP were used as the platform natural language processing. MySQL database were used to store data from the retrieved tweets.

3.3 Target Population and Sampling

Target population refers to the entire group of objects or individuals that the researcher considers during the study (Kothari & Garg, 2014). Twitter posts relating to roads in Nairobi county were chosen as the population for the research. The study was conducted via purposive sampling on social media. A sample was selected from Tweets mentioning Nairobi county Twitter page, '@NairobiCityGov' or 'Nairobi county' in combination with the words the word 'road', roads or rd.

3.4 Data Collection

Data collection refers to the process of collecting relevant information from the relevant sources in order to find answers to a research problem. This will enable the researcher to later test the hypothesis and evaluate the outcomes (Kumar, 2011). The accuracy of the outcome of the study relies on how well data collection is carried out.

Primary data collection was done via mining on Twitter using the standard APIs provided by the application. Twitter APIs are limited to the extraction of tweets which are seven or less days old. Mining of older tweets were done using GetOldTweets3 open source code. Posts were collected from the Nairobi County Government twitter account as well as from tweets mentioning the same account. The search was conducted using a combination of keywords as illustrated in table 3.1.

Table 3.1: Keywords to be Used for Twitter Search During Data Collection

Words	**Search combination**
@NairobiCityGov, road	@NairobiCityGov and road
@NairobiCityGov, roads	@NairobiCityGov and roads
@NairobiCityGov, rd	@NairobiCityGov and rd
Nairobi county, road	Nairobi county and road

Nairobi county, roads	Nairobi county and roads
Nairobi county, rd	Nairobi county and rd

3.5 Data Pre-processing

Data directly mined from social media sources like Twitter is highly unstructured. This makes immediate analysis challenging. It is therefore important to clean the raw data. Pre-processing of data was done to convert the unstructured Tweets into a structured set. The clean tweets were stored in a Pandas dataframe for further processing. The tweet preprocessing included removal of URLs, hashtags, special characters and some Twitter specific texts.

3.6 Natural language processing

This study incorporated a hybrid model of classification and named entity recognition. Named entity detection classes predefined in Stanford Core NLP were used. Classification labels relevant to the domain of public service participation were defined; negative (0), positive (1) and neutral (2).

The result from the classification step was stored in a structured format in a MySQL database. Stored data comprised of identified tweets and their classification. This made it easier to query data and retrieve information. Processed data contains identified tweets and their classification. Information was then retrieved from the tweets labelled as complaints. Locations were extracted from the complaint tweets using Stanford NER. The extracted locations were then reverse geocoded using the google geocoding API and plotted on a map. Frequent words in the complaint tweets were visualized using WordCloud in Python.

3.7 Research Quality

Research quality is the extent to which the research results portray correctness and how the researcher ensured that quality was upheld. This is done via controlling two criteria; Validity and Reliability. Validity refers to the extent to how appropriate the tools, techniques and data used are for the research problem. Reliability refers to the ability to replicate the exact processes and achieve the same results. The data collected was divided into a training test and a test set. The test set that had not been used for training and therefore unknown to the classifier was used to evaluate

the model. The test set was fed into the model for prediction and the quality of the classifier evaluated using the metrics described below.

A confusion matrix contains information about actual and predicted classifications done by the classification model. True positives (TP) is the number of correct tweet predictions for each classification label. True negatives (TN) is the number of incorrect tweet predictions for each classification label. Table 3.2 below illustrates a confusion matrix that will be used in the evaluation of the classifier.

Table 3.2 Confusion Matrix

	Predicted 0	Predicted 1	Predicted 2
Actual 0	TP	FN	FN
Actual 1	FN	TP	FN
Actual 2	FN	FN	TP

The accuracy of the (SVC) with TF-IDF features complaint classifier was evaluated using the metrics accuracy, precision, recall and f-score.

Accuracy refers to the ratio of correct predictions to the total predictions made (Feldman & Sanger, 2007). Accuracy is calculated as

accuracy = total correct predictions / total predictions made * 100

Precision refers to the ratio of correctly classified documents to the total number of documents classified within a particular category (Feldman & Sanger, 2007). Precision is a measurement of false positives calculated as:

Precision = True Positives / True Positives + False Positives

Recall is defined as the number of correctly classified documents among all documents belonging to that particular category (Feldman & Sanger, 2007). Recall is a measurement of false negatives calculated as:

Recall = True Positives / True Positives + False Negatives

F-Score is defined as a mean of precision and recall (Feldman & Sanger, 2007). This is calculated as:

$$F - Score = 2 * Precison * Recall / Precision + Recall$$

3.8 Ethical Considerations

Ethical approval for the research was obtained from the Strathmore University Institutional Ethics Review Committee (SU-IERC). The SU-IERC is accredited by the Commission for Science, Technology and Innovation (NACOSTI) to conduct ethical reviews of research proposals in the human and behavioural sciences. The research proposal was submitted to the SU-IERC for review. Issues that arose during the review processed were addressed in consultation with the review committee. An ethical review approval was awarded to the researcher upon satisfactorily addressing the issues raised.

The researcher handled any data provided by the participants in confidentiality to preserve privacy rights according to the law. No sensitive data was made available to third parties unless the participant authorized it. This was done by ensuring that the participants personal identifiers such as National Id numbers and telephone numbers are not linked to their data. Any previous research referenced during this study was duly acknowledged. This included citation and referencing of the source of the content, tables and images.

Chapter 4 System Design and Architecture

4.1 Introduction

The purpose of this study was to bring public complaints to the attention of government through the use of a tool that extracts and visualizes public complaint data sourced from citizens on social media using data mining techniques. This chapter describes the architecture and in detail of the proposed tool by based on defined requirements. An initial description of data analysis and requirements analysis for the tool is provided. UML diagrams were used to illustrate: the system architecture, description of system components and interaction between users and system components. This was achieved by using various design diagrams comprising; system architecture, context diagram, data flow diagram, data model and system wireframes.

4.2 Data Analysis

Data was collected GetOldTweets3 open source code by Dmittry Mottl (Mottl, 2019). A total of 1023 tweets were collected using the keywords specified in section 3.4. The collected data had a total of were composed of the fields; date, username, to, replies, retweets, favorites, text , geo, id, mentions, hashtags and permalink. The field ‘date’ contains the date and time when the tweet was posted. ‘username’ contains the account name of the user who created the tweet. ‘to’ field contains the account name of the user that the tweet is directed at. ‘replies’ field holds the value of the number of responses that a tweet received. The ‘retweets’ field holds the number of reposts that a tweet got from other users. ‘favorites’ hold the number of times that a tweet was favorited. The ‘text’ field contains the actual tweet content posted by the user. The ‘geo’ field contains geographic coordinates for user accounts that have enabled geolocation for their tweets. The ‘id’ field contains a unique numeric tweet identifier for each tweet. The ‘mentions’ field contains the usernames for accounts that appear in the tweet text. The ‘hashtags’ field consists of hashtags that are present in the tweet text. The ‘permalink’ contains a URL link of the tweet. Figure 4.1 illustrates a sample of the raw data collected.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
date	username	to	replies	retweets	favorites	text	geo	mentions	hashtags	id	permalink					
07/02/2020 14:04	Ma3Route		0	0	3	17:04 @TrishEmbenzy @NairobiCit		@TrishEmbenzy @N		1.23E+18	https://twitter.com/Ma3Route/status/1225782486829015040					
07/02/2020 14:04	OK_Onyatta	TrishEmbenzy	0	0	0	Who do you think is to blame for the mess? The boda bc				1.23E+18	https://twitter.com/OK_Onyatta/status/1225782391068856321					
16/01/2020 16:49	Di_Mystro	SakajaJohnso	1	1	2	That should mean fixing the #drain		@Ma3Rou	#drainage	1.22E+18	https://twitter.com/Di_Mystro/status/1217851479345844224					
16/01/2020 15:45	Sir_Labz		0	0	5	Our mental health is affected by tr		@KenyanTraffic @M		1.22E+18	https://twitter.com/Sir_Labz/status/1217835334077161473					
14/01/2020 18:57	Ma3Route		2	0	5	21:57 @NairobiCityGov @Paul_Fro		@NairobiCityGov @		1.22E+18	https://twitter.com/Ma3Route/status/1217158788186353670					
14/01/2020 18:56	shafiee65	Ma3Route	0	0	0	He should be arrested, no need to tolerate people who				1.22E+18	https://twitter.com/shafiee65/status/1217158673320906752					
13/01/2020 16:15	thedunne		0	13	21	These rains have revealed the fake		@NairobiCityGov		1.22E+18	https://twitter.com/thedunne/status/1216755542628208642					

Figure 4.1 Raw Data Collected Using GetOldTweets3

After analysing the collected data in relation to the aim of the study, four fields were seen to be the only ones required. These were id, date, username and text. The rest of the fields were dropped from the data used in the following stages of this study. The 'id' field which was in scientific format was also converted to numeric format to allow for easy processing in the following stages.

4.3 Requirement Analysis

This research aimed at developing a tool for extraction and visualization public complaints from Twitter. Based on this, the requirements that should be met by the proposed tool are defined in this section.

4.3.1 Functional Requirements

The following functional requirements were identified as key to ensuring for the tool to efficiently extract and visualize public complaints from Twitter.

i. The system should be able to retrieve data from using Twitter based on keywords specified by the user.
ii. The system should be able to classify complaint and non-complaint tweets from tweets collected by the user.
iii. The system should be able to identify the locations mentioned in the complaint tweets using named entity recognition
iv. The system should be able to identify frequently mentioned words in the complaint tweets collected
v. The system should display to the user a map of identified locations and frequently mentioned words within complaint tweets.

4.3.2 Non- Functional Requirements

i. Usability

The intended users of this system are the Nairobi County staff in charge of handling citizen complaints. The solution should therefore be easy to use to enable the users mine data from Twitter and view the analysed results. The visualised data should also be easy to read and understand.

ii. Scalability

The solution should be able to handle increased number of simultaneous users without downgrading its performance

iii. Availability

The system should be available when needed for use and should not suffer inconveniencing downtimes.

iv. Security

The system should be able to prevent any unauthorized access or manipulation to the data stored.

4.4 System Architecture

The system architecture illustrates the input, interactions and outputs between various components of the system. The solution here illustrates the complaint classification, analysis and visualizing components of the system. Figure 4.2 illustrates the architecture of the system that will mine and visualize public complaints from Twitter.

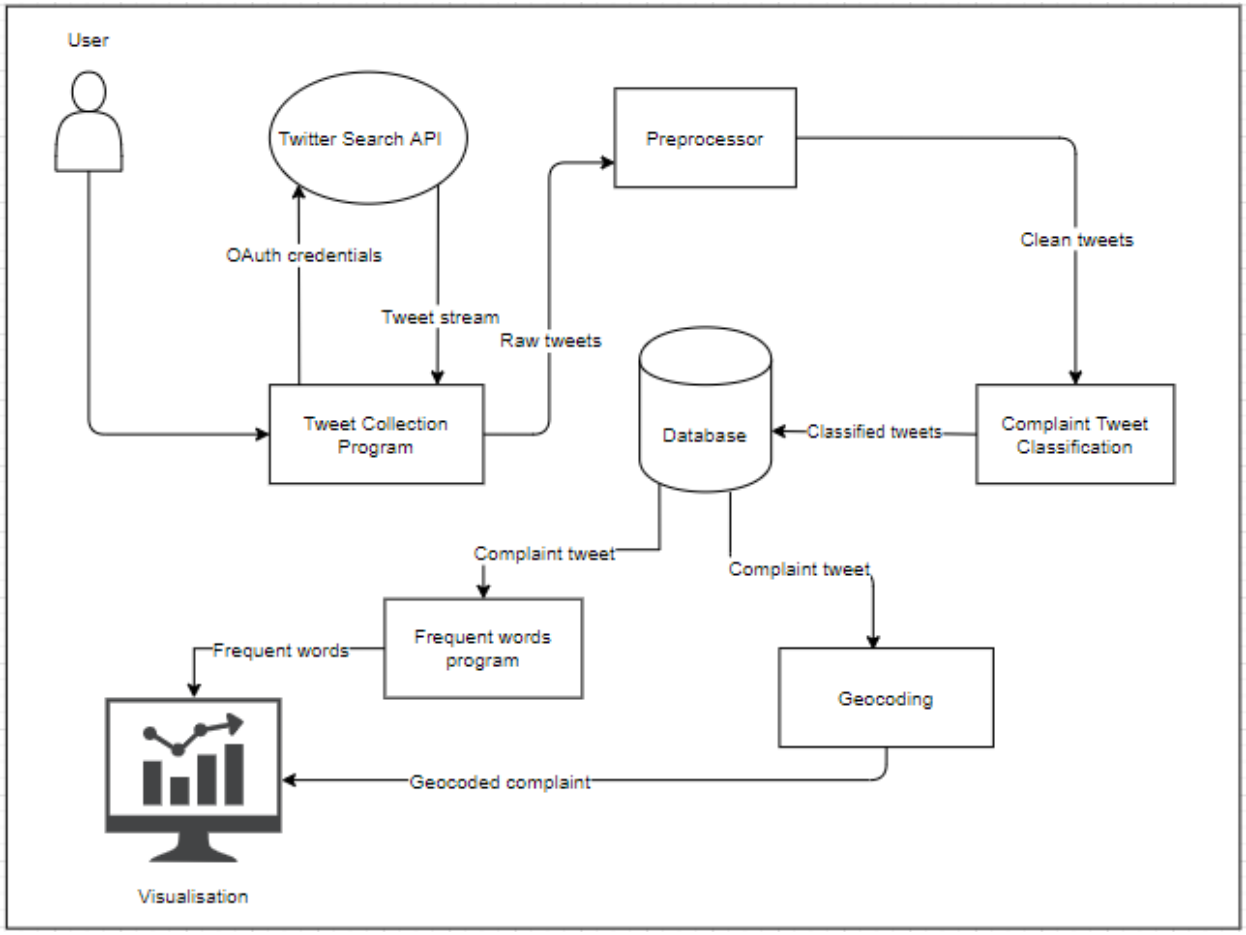


Figure 4.2 System Architecture

4.5 Context Diagram

The context diagram illustrates the boundary of the system, its entities and requirements. The user sends a request to search tweets. The system passes the request to the Twitter Search API. The Twitter Search API sends an API verification to the system to enable it use the API. The tool provides the confirmation details to establish the connection. The API retrieves tweets matching the request and returns them to the system which then returns them to the user. The user then issues

a complaint classification request to the system which classifies the tweets and passes complaint tweets to the Geocoding API. The geocoding API geocodes the complaint tweets and returns results back to the system. The user makes a visual result request to the system and receives visual feedback. The context diagram is illustrated in Figure 4.3 below.

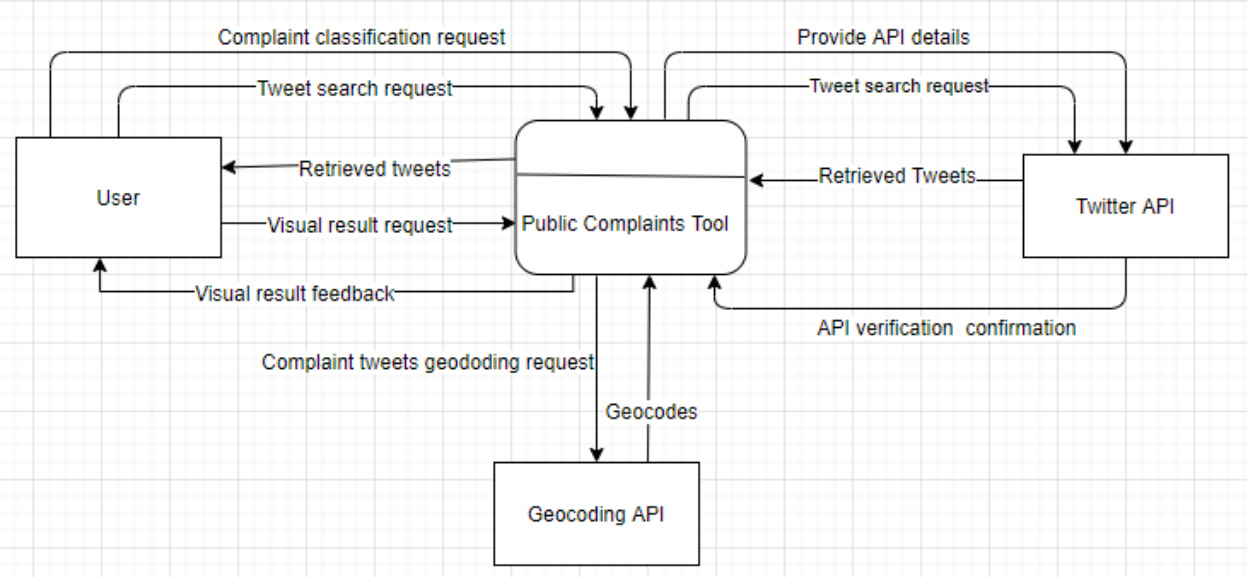


Figure 4.3 Context Diagram

4.6 Level 1 Data Flow Diagram

The level 1 data flow diagram in figure 4.4 below gives a more detailed view of the tool by indicating the entities contained in the system, the processes and the data stores. Arrows are used to illustrate data flow among the various components in the DFD. Process 1 (Collect Tweets) receives a Tweet search request from the user and forwards the request to the Twitter Search API entity. The Twitter API validates the API credentials and returns retrieved tweets matching the search request if the validation is successful. Retrieved tweets for process 1 are stored in data store D1(Retrieved Tweets). The user sends a complaint classification request to process 2(Preprocess Tweets). Process 2 reads tweets from data store D1 and cleans them. The clean tweets are then stored in data store D2 (clean tweets). Process 3 (Classify Tweets) reads clean Tweets from D2 classifies and labels them into complaint and non-complaint using machine learning. The labelled Tweets are then stored in D3(Labelled Tweets). Process 4(Identify Locations) reads labelled

complaint tweets from D3, identifies locations mentioned in the tweets and stores them in D4(Complaint Locations). Process 5(Geocode Locations) reads complaint locations from D4 and passes the to the Geocoding API which returns the geocodes. Process 5 then stores the geocodes into D5(Geocodes). Process 6(Identify Frequent Words) reads labelled complaint tweets from D3, identifies frequently mentioned words and stores them in D6(Frequent Words). The user makes a request to view visualised results via process 7(Visualize Results). Process 7 reads Geocodes from D5 and frequent words from D6 then returns the Visual result feedback to the user.

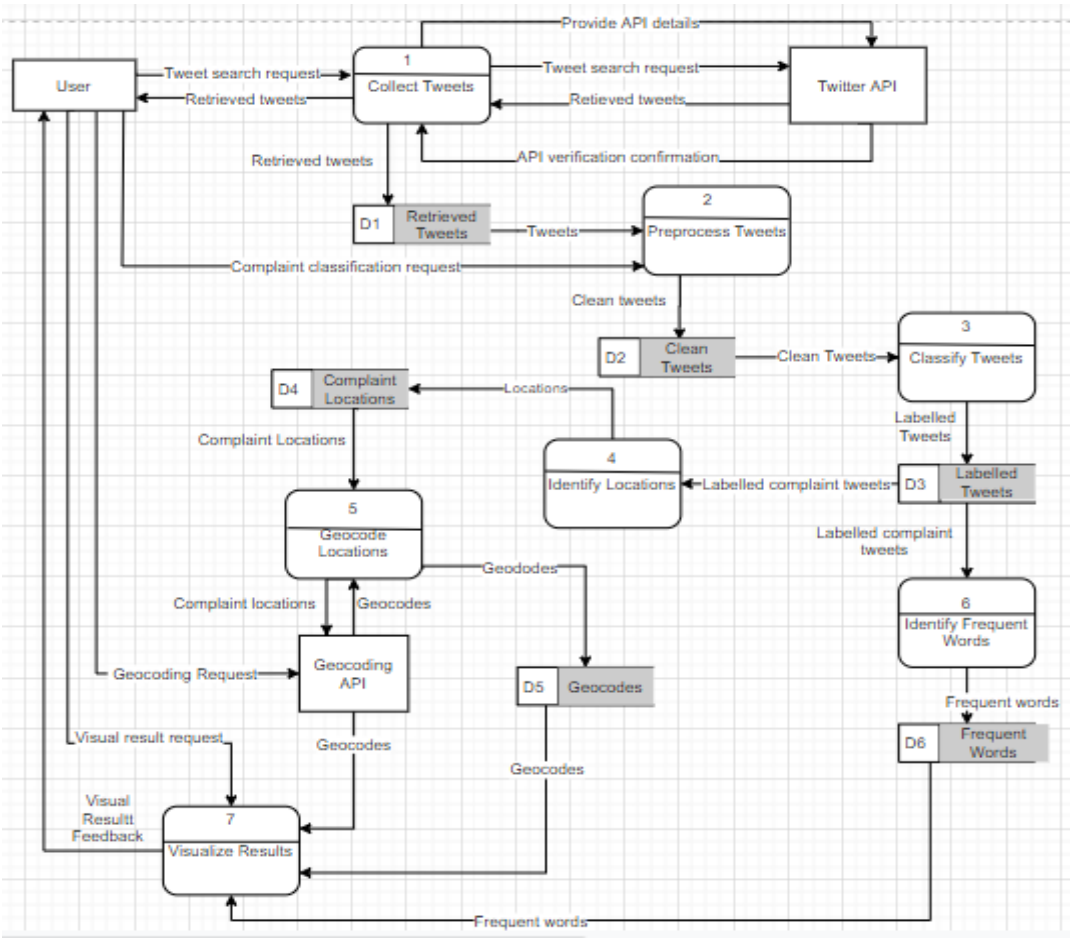


Figure 4.4 Level 1 Data Flow Diagram

4.7 Data Model

A data model refers to a conceptual representation of data objects, the associations between different data objects and the rules. Emphasis is placed on the data needed and how it should be organized rather than the operations that should be performed on the data.

4.7.1 Conceptual Data Model

A conceptual data model defines what the system is made up of. This is illustrated using entities, attributes and their relationships. A conceptual data model for the public complaints tool is defined in Figure 4.5.

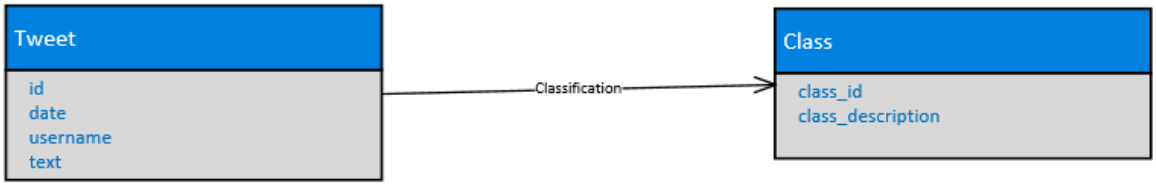


Figure 4.5 Conceptual Data Model

4.7.2 Logical Data Model

A logical data model defines how the system should be implemented regardless of a DBMS. This is done by defining the structure elements and the relationships between them. Figure 4.6 illustrates the logical data model for the tool.

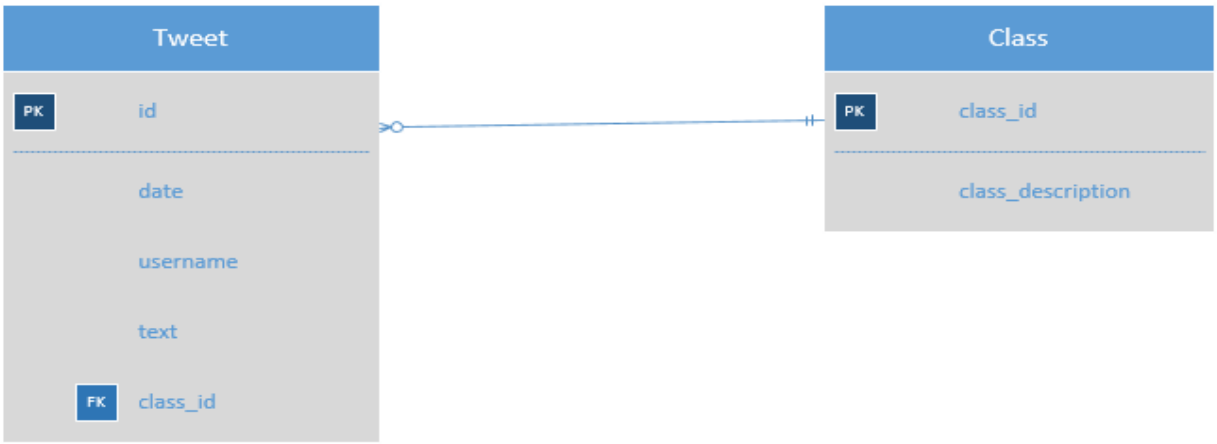


Figure 4.6 Logical Data Model

4.7.3 Physical Data Model

A physical model defines how data will be implemented using a specific DBMS. It visualizes the database structure taking into account column keys, constraints, indexes and triggers. The physical

data model for the tool is designed specific to a MySQL database. The tables and the relevant components are shown in Figure 4.7 below

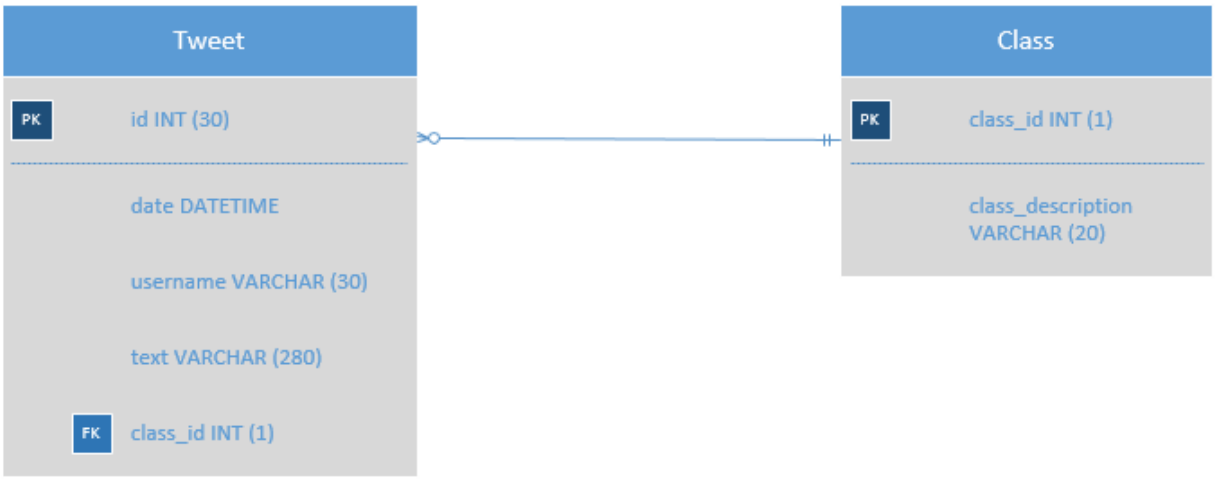


Figure 4.7 Physical Data Model

The tables below describe the attributes of the tables in Figure 4.8 above. PK is used to define the primary key while FK is used to define the foreign key

Tweet Table

Table 4.1 shows the tweet table which holds tweet details

Table 4.1 Tweet Table

Field	**Data type and Length**	**Details**	**Notes**
Id	Int (30)	PK	This is the tweet id obtained from Twitter. Each tweet has a unique id
Date	Datetime		This is the tweet creation date derived from Twitter
Username	Varchar (30)		This is the username of the tweet author
Text	Varchar (280)		This contains actual text from the tweet
class_id	Int (1)	FK	This is the tweet classification label.

Class Table

Table 4.2 shows the class table which holds classification label details.

Table 4.2 Class Table

Field	Data type and Length	Details	Notes
class_id	Int (1)	PK	This is the class id.
class_description	Varchar (20)		This is the class label description

4.8 System Wireframes

A wireframe is a visual representation of the system to be developed. The figures below represent the graphical user interface to the tool. The function of the wireframes is also elaborated by the researcher.

Figure 4.8 shows the wireframe for the homepage.

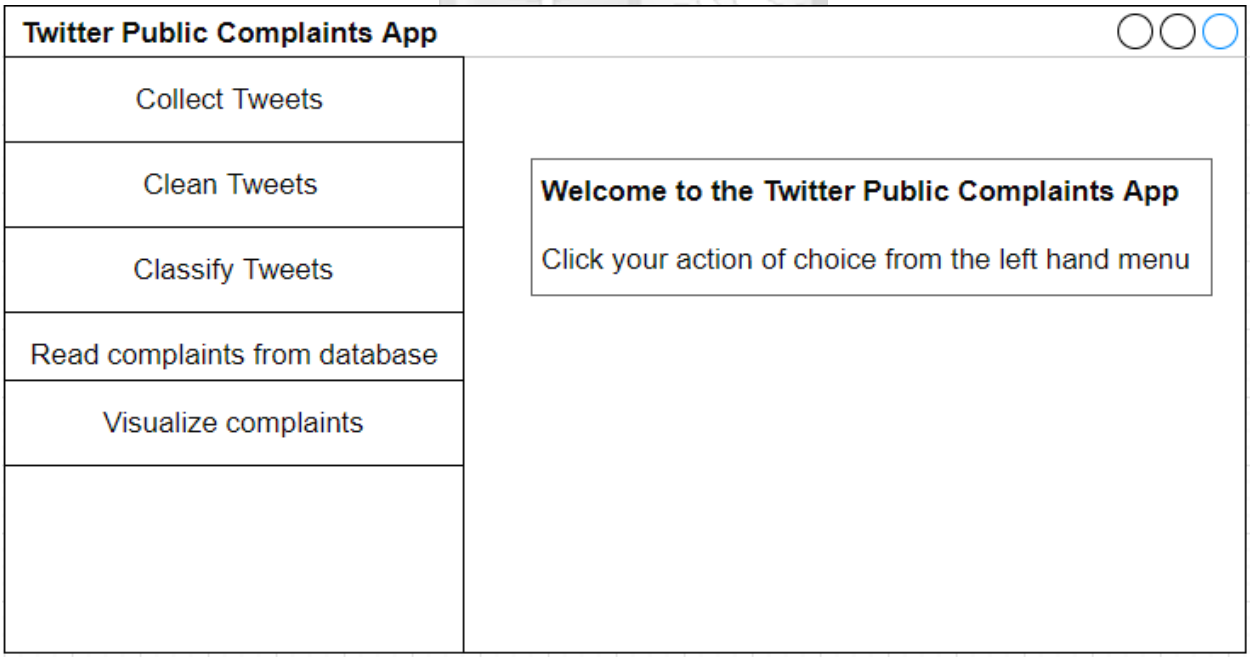


Figure 4.8 Homepage Wireframe

Figure 4.9 shows the wireframe for the Collect Tweets screen.

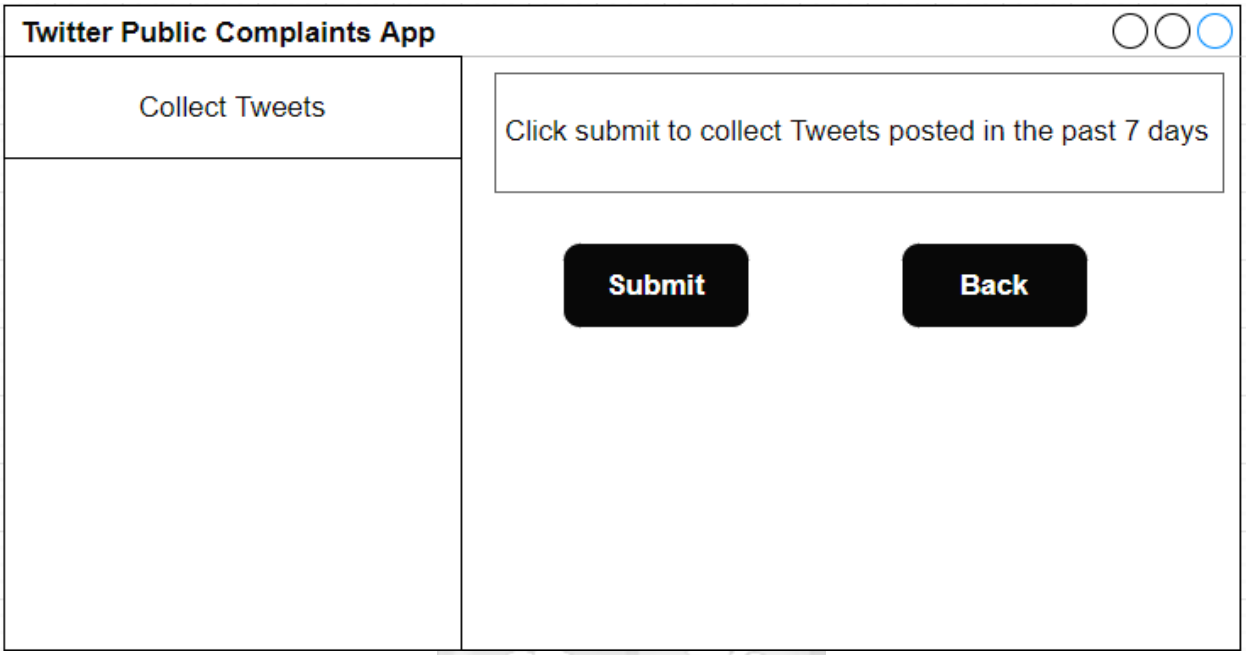


Figure 4.9 Collect Tweets Wireframe

Figure 4.10 shows the wireframe for the Clean Tweets screen.

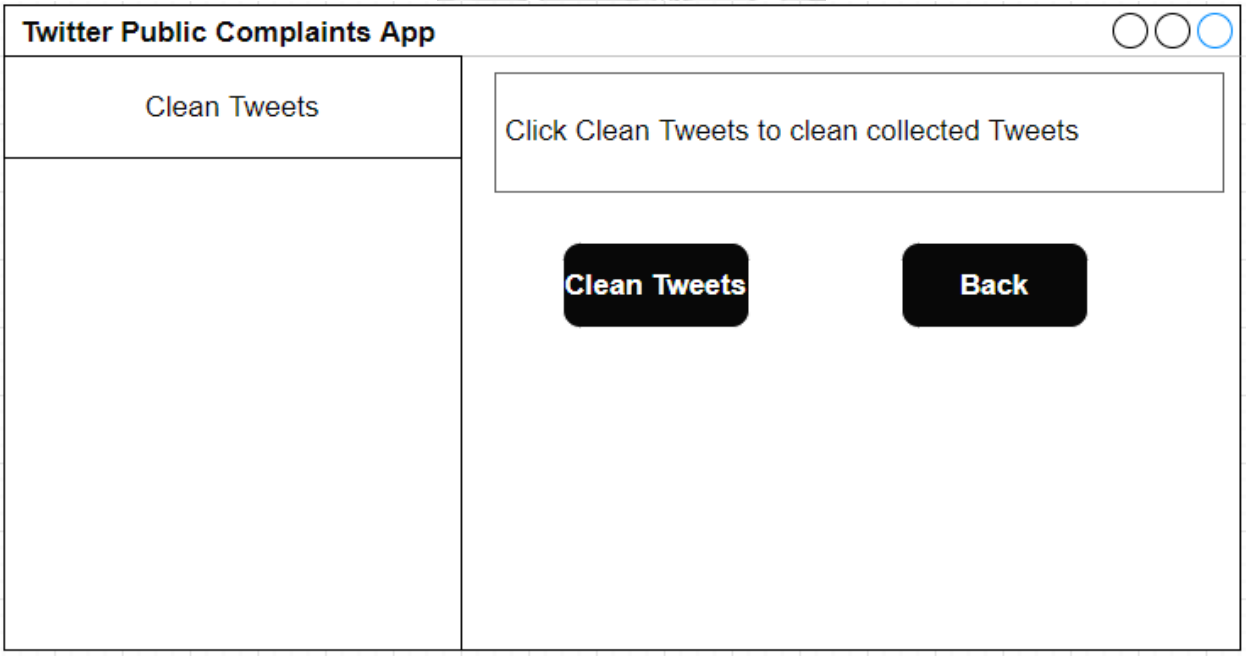


Figure 4.10 Clean Tweets Wireframe

Figure 4.11 shows the wireframe for the Read Complaint Tweets screen.

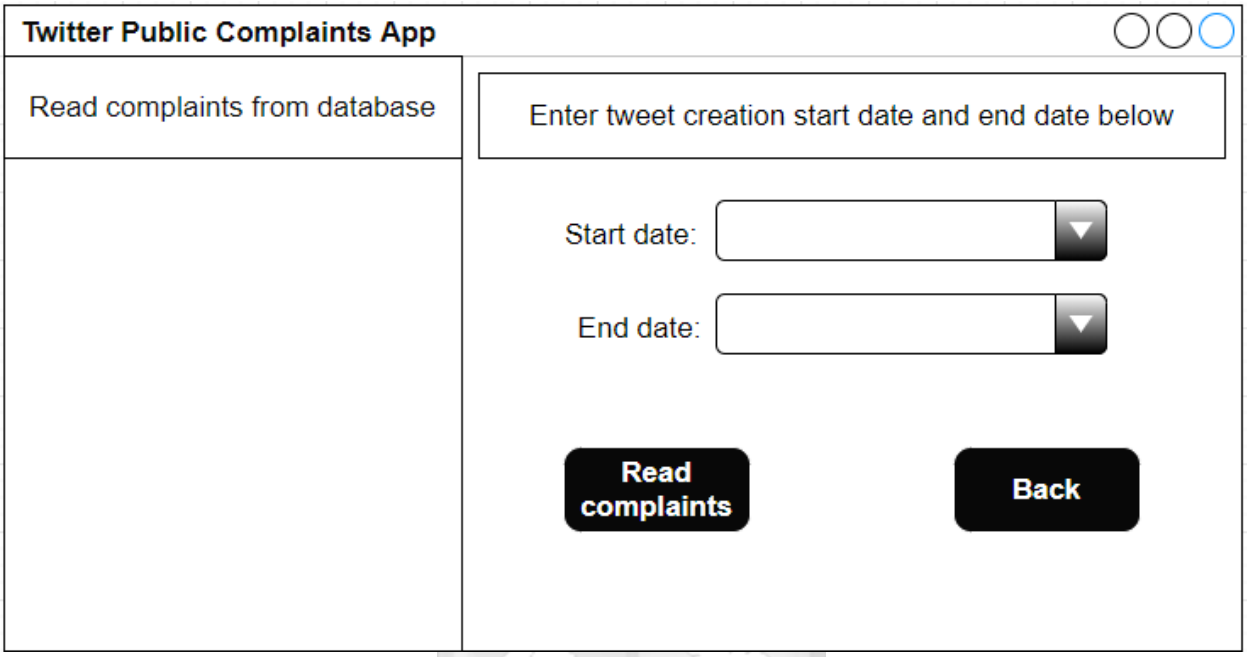


Figure 4.11 Read Complaints from Database Wireframe

Figure 4.12 shows the wireframe for the Visualize Tweets screen.

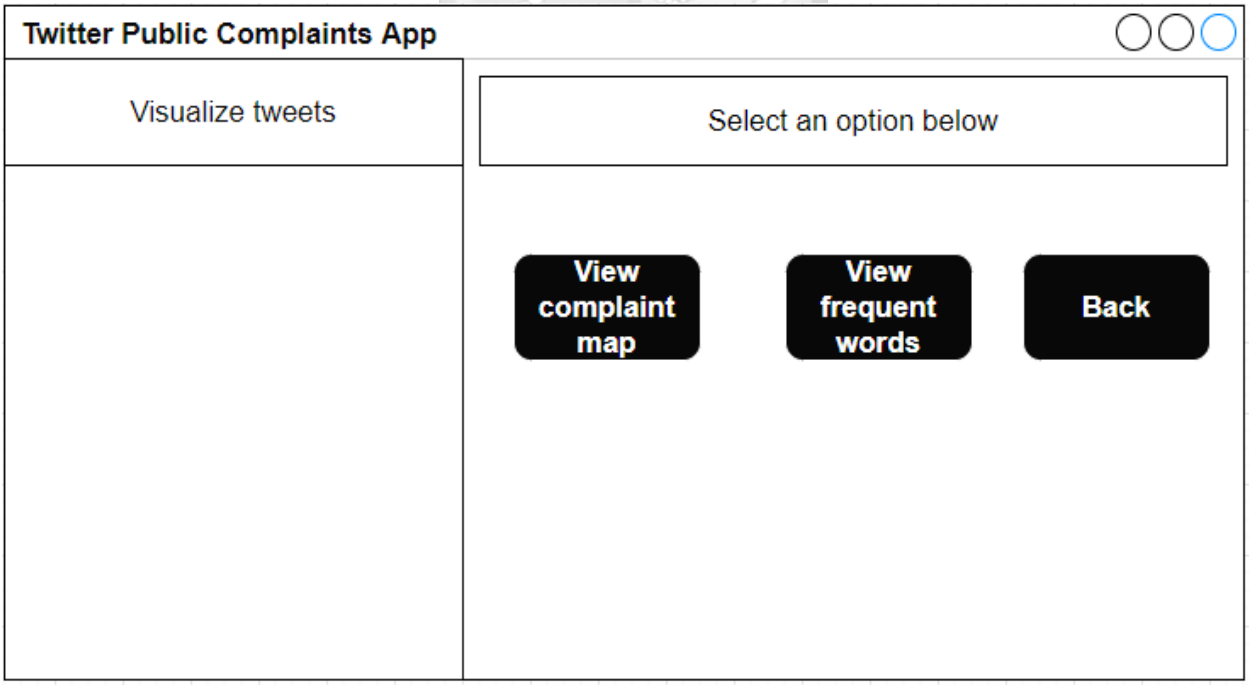


Figure 4.12 Visualize Complaints

Figure 4.13 shows the wireframe for the Complaint Map screen.

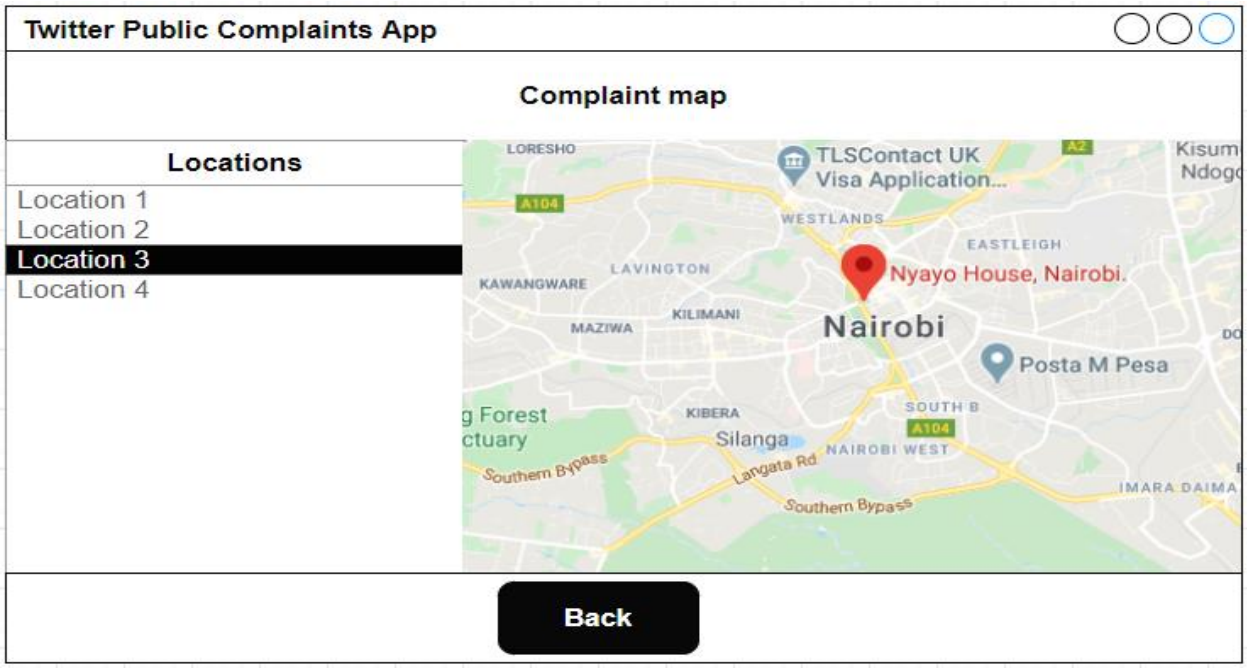


Figure 4.13 Complaint Map Wireframe

Figure 4.14 shows the wireframe for the Frequent Words screen.

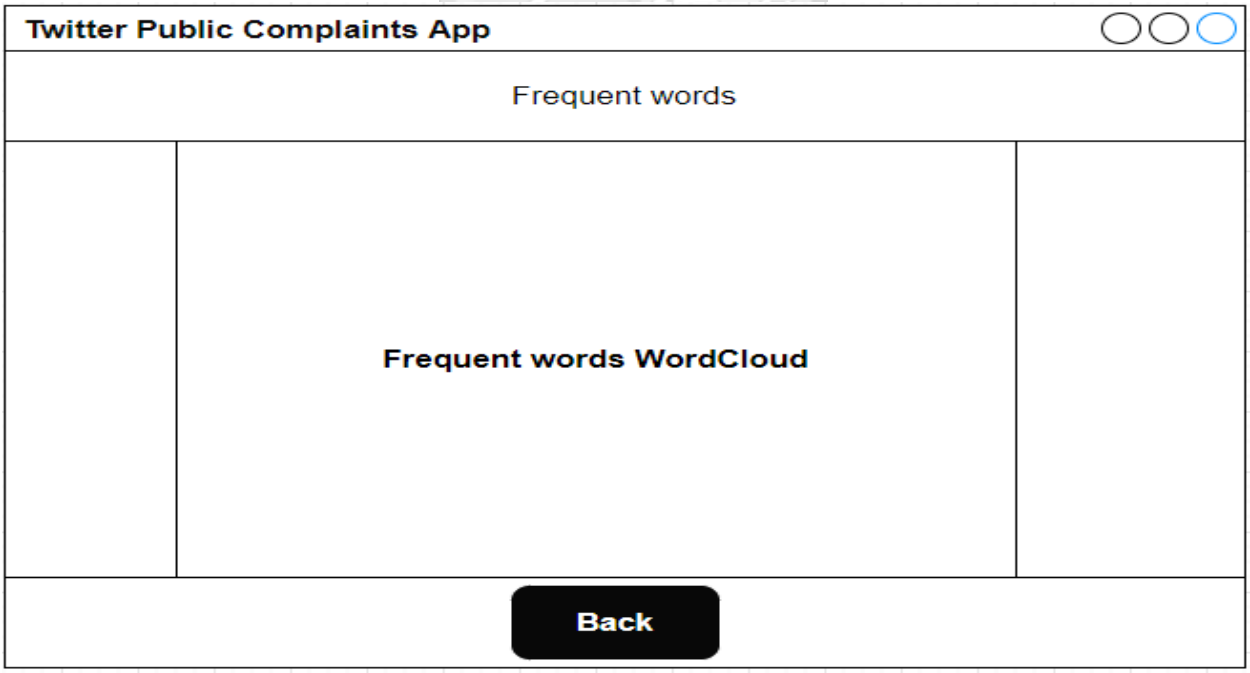


Figure 4.14 Frequent Words Wireframe

Chapter 5 System Implementation and Testing

5.1 Introduction

This chapter describes how the tool was implemented and tested. It describes the process of Tweet collection to build a corpus, Tweet pre-processing then classification of complaint and non-complaint tweets using machine learning. The testing section analyses the functionality and usability of the tool to ensure that it satisfies the requirements and objectives that were defined in chapter 1.

5.2 Building the Corpus

Tweets were collected using GetOldTweets3 which is an improvement of the of the original GetOldTweets Library by Jefferson Henrique. GetOldTweets3 is compatible with Python 3 which was used in the implementation of this tool. The improvement makes for a seamless running of the package on Windows OS. Tweets were collected using the keywords specified in section 3.4.

Figure 5.1 illustrates the tweet collection process using GetOldTweets3.

Figure 5.1 Collecting Past Tweets from Twitter

5.3 Preprocessing

The data collected consisted of Tweets which are unstructured in format. It was key that the data was cleaned to make it suitable for use in the machine learning classification model. A sample of the raw unstructured data is shown in figure 5.2 below.

	A	B	C	D	E	F	G	H
1	date	username	to	replies	retweets	favorites	text	geo
2	07/02/2020 14:04	Ma3Route		0	0	3	17:04 @TrishEmbenzy @NairobiCityGov Who do you think is to blame for the mess? The boda boda cyclists? Urban planners for not buil	
3	07/02/2020 14:04	OK_Onyatta	TrishEmbe	0	0	0	Who do you think is to blame for the mess? The boda boda cyclists? Urban planners for not building enough roads? Or motorists for fillin	
4	16/01/2020 16:49	Di_Mystro	SakajaJoh	1	1	2	That should mean fixing the #drainage and #garbage disposal systems. These are major reasons why our roads cannot withstand even th	
5	16/01/2020 15:45	Sir_Labz		0	0	5	Our mental health is affected by trafficked roads, especially for those who live close to noisy roads. Commuting by bike is not just fun a	
6	14/01/2020 18:57	Ma3Route		2	0	5	21:57 @NairobiCityGov @Paul_Frohmiller He should be arrested, no need to tolerate people who donâ€™t respect tax payers money. W	
7	14/01/2020 18:56	shafiee65	Ma3Route	0	0	0	He should be arrested, no need to tolerate people who donâ€™t respect tax payers money. We need to to follow the law to the later. Fr	
8	13/01/2020 16:15	thedunne		0	13	21	These rains have revealed the fake veneer of our roads. The potholes are too big and too deep for us to look the other way. Before long	
9	12/01/2020 06:33	KenyanTraffic		0	0	3	I guess potholes & sinkholes on our roads and streets are great equalisers like traffic jam has taken that role as well @NairobiJiji @Nairo	

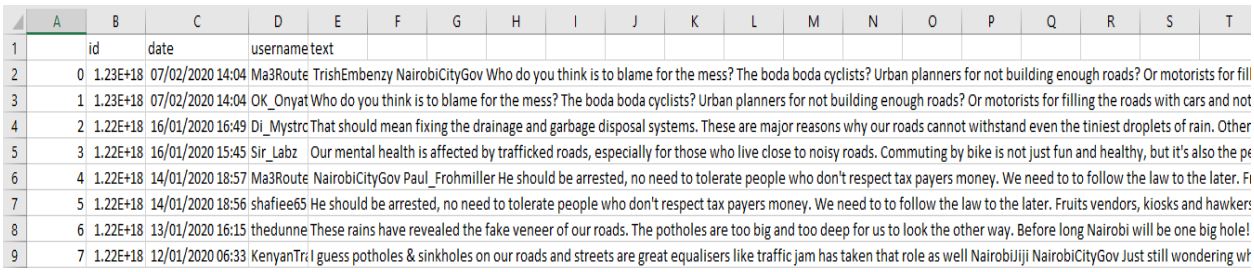
Figure 5.2 Raw Tweets

The data collected consisted of a number of attributes including date, username, to, replies, retweets, favorites, text, geo, mentions, hashtags, id and permalink. Only four attributes were considered important for this study and they are as follows id, date, username and text. The text was the only attribute required to build the corpora. Upon observation, the text contained special Twitter terms and characters which were not relevant to the research. For example, hashtags (#), mentions (@username), retweets (RT), URL links and non-utf8 characters. These were removed by using code written in regular expressions as shown in Figure 5.3 below.

```
import pandas as pd
import csv
import re
def clean_tweets():
    #reading csv file into panda dataframe
    df=pd.read_csv("test1.csv", error_bad_lines=False)
    print("CSV file read")

    print("Removing # in hashtags from tweets")
    df.text=df.text.str.replace("#" ,"")#removing hashtags

    print("Removing URLs from tweets")
    df.text=df.text.str.replace("https?:\/\/.*[\r\n]*" ,"")#removing URLs

    print("Removing @ in user from tweets")
    df.text=df.text.str.replace("@" ,"")#removing @

    print("Removing RTs from tweets")
    df.text=df.text.str.replace("RT" ,"" ,False)#removing RTs

    print("Replacing â€™ with ' in tweets")
    df.text=df.text.str.replace("â€™" ,"'")#replacing â€™

    print("Removing non UTF-8 characters from tweets")
    df.text=df.text.str.replace("[^\x00-\x7F]+" ,"")#remove non-UTF8 characters

    print("Removing time from tweets")
    df.text=df.text.str.replace("[0-9][0-9]:[0-9][0-9]" ,"")#removing time

    print("Removing remainder time from tweets")
    df.text=df.text.str.replace("[0-9]:[0-9][0-9]" ,"")#removing time

    print("Removing pic.twitter.com from tweets")
    df.text=df.text.str.replace("pic.twitter.com/[A-Za-z0-9]+" ,"")#removing pic URLs

    print("Removing via from tweets")
    df.text=df.text.str.replace("via [A-Za-z0-9][A-Za-z0-9\__]+[A-Za-z0-9_]" ,"")#removing vi

    # dropping ALL duplicte values
    print("Removing duplicate tweets")
    df.drop_duplicates(subset ="text",
                       keep = 'first', inplace = True)

    print("Writing clean tweets")
    df.to_csv("clean\clean_test1.csv" ,encoding="utf8")#print clean tweets

clean_tweets()
```

Figure 5.3 Removal of Special Terms and Characters Using Regex

The clean Tweets were stored as a CSV file. A sample of clean tweets are shown in Figure 5.4 below



	A	B	C	D	E
1		id	date	username	text
2	0	1.23E+18	07/02/2020 14:04	Ma3Route	TrishEmbenzy NairobiCityGov Who do you think is to blame for the mess? The boda boda cyclists? Urban planners for not building enough roads? Or motorists for fill
3	1	1.23E+18	07/02/2020 14:04	OK_Onyat	Who do you think is to blame for the mess? The boda boda cyclists? Urban planners for not building enough roads? Or motorists for filling the roads with cars and not
4	2	1.22E+18	16/01/2020 16:49	Di_Mystrc	That should mean fixing the drainage and garbage disposal systems. These are major reasons why our roads cannot withstand even the tiniest droplets of rain. Other
5	3	1.22E+18	16/01/2020 15:45	Sir_Labz	Our mental health is affected by trafficked roads, especially for those who live close to noisy roads. Commuting by bike is not just fun and healthy, but it's also the pe
6	4	1.22E+18	14/01/2020 18:57	Ma3Route	NairobiCityGov Paul_Frohmiller He should be arrested, no need to tolerate people who don't respect tax payers money. We need to to follow the law to the later. Fr
7	5	1.22E+18	14/01/2020 18:56	shafiee65	He should be arrested, no need to tolerate people who don't respect tax payers money. We need to to follow the law to the later. Fruits vendors, kiosks and hawkers
8	6	1.22E+18	13/01/2020 16:15	thedunne	These rains have revealed the fake veneer of our roads. The potholes are too big and too deep for us to look the other way. Before long Nairobi will be one big hole!
9	7	1.22E+18	12/01/2020 06:33	KenyanTra	I guess potholes & sinkholes on our roads and streets are great equalisers like traffic jam has taken that role as well NairobiJiji NairobiCityGov Just still wondering wh

Figure 5.4 Clean Tweets Sample

The cleaned tweets were then labelled as 0 for complaint, 1 for appreciation and 2 for neutral. The labelled Tweets were used to train the machine learning classification model. A sample of labelled Tweets is shown in Figure 5.5 below.

	A	B	C	D	E	F
1		id	date	username	text	label
18	16	1.21E+18	19/12/2019 16:08	Sir_Labz	Hey, city planners, you should consider the happiness and well-being that cycling engenders. Throw it in	3
19	17	1.21E+18	17/12/2019 03:59	Sir_Labz	How to design our roads better by prioritizing people. Protected bike lanes, protected intersections and	3
20	18	1.21E+18	13/12/2019 10:25	Kdmuya	FoodForThought suppose pres Kenyatta direct the army and NYS to do all delapilated roads in Nairobi be	0
21	19	1.20E+18	11/12/2019 15:04	vwamu_d	NairobiCityGov is decaying. And that's an ClimateEmergency . Drive along Uhuru Highway, it's beautiful.	0
22	21	1.20E+18	09/12/2019 22:47	MattoKaru	While SonkoInCourt this the condition of Kasarani_mwiki_road ..Its very annoying to see the road that cc	0
23	22	1.20E+18	09/12/2019 12:18	DouglasM	Some counties are crying of ran away corruption, nothing has been done. Case in mind is the governor of	0
24	23	1.20E+18	09/12/2019 07:08	JMutinda_	Good morning Nairobians, don't just tag KeRRA_Ke, KURAroads, KeNHAKenya & NairobiCityGov with rega	3
25	24	1.20E+18	05/12/2019 09:27	UoNSchoc	publicvoiceuhurupark2019 Prof Richard Mulwa caselapchss Clearly NairobiCityGov has failed in master pl	0
26	25	1.20E+18	27/11/2019 09:39	Mital_KE	This is where incompetence of MikeSonko, NairobiCityGov and KURAroads on roads have left us at risk w	0
27	26	1.20E+18	23/11/2019 17:40	KenyanTra	I love when it is rains, because you don't need further evidence that the repairs done on your roads were	1

Figure 5.5 Labelled Tweets Sample

5.4 Training the Model

The cleaned and labelled tweets were used to train the classification model. The CSV file containing cleaned and labelled tweets was loaded into a Pandas dataframe. A Data Frame is Pandas data structure which is two-dimensional and made up of columns of potentially different types. The text column was selected to create the corpus. The corpus was parsed to stem and group words with potentially similar meanings, for example buying and buy. The parsed data was then split into a training and test set using the sklearn.model_selection's train_test_split functionality. 30% of the corpus was used for test data while 70% was used for training. Data to test the performance of the model. The data was in text format and needed to be converted into a numerical format suitable for machine learning. TF-IDF matrix was used for this purpose. TF stands for Term frequency and is used to calculate how many times a term appears in a document. IDF stands for inverse document frequency and it calculates how important a word is. More weight is given to rare words as compared to the common ones. The TfidfVectorizer of sklearn was used to calculate

TF-IDF. The vectorizer discarded English stopwords and the number of maximum features was set to 1000. The training data was transformed into a TF-IDF matrix The vectorizer code snippet is shown in Figure 5.6 below

```
45 #train vectorizer with training set only
46 vectorizer = TfidfVectorizer(stop_words="english", max_features=1000, decode_error="ignore")
47 vectorizer.fit(X_train)
48 X_train_vectorized = vectorizer.transform(X_train)
49
```

Figure 5.6 TF-IDF Vectorizer

To select a suitable model, four models were trained using sklearn pipelines. Pipelines allow for the carrying out of steps in a sequence for a model. For example, convert training text into vectorised format and pass this to the model. Four pipelines were created using TF-IDF and CountVectorizer for vectorization and SGDClassifier and SVC (support vector classifier). The models were cross-validated using the cross_val_score feature of sklearn. The models were trained twice and their mean estimated accuracy recorded.

```
#Use pipeline to carry out steps in sequence with a single object
#Test several classifiera
# start with the classic
# with either pure counts or tfidf features
sgd = Pipeline([
        ("count vectorizer", CountVectorizer(stop_words="english", max_features=3000)),
        ("sgd", SGDClassifier(loss="modified_huber"))
    ])
sgd_tfidf = Pipeline([
        ("tfidf_vectorizer", TfidfVectorizer(stop_words="english", max_features=3000)),
        ("sgd", SGDClassifier(loss="modified_huber"))
    ])

svc = Pipeline([
        ("count_vectorizer", CountVectorizer(stop_words="english", max_features=3000)),
        ("linear svc", SVC(kernel="linear"))
    ])
svc_tfidf = Pipeline([
        ("tfidf_vectorizer", TfidfVectorizer(stop_words="english", max_features=3000)),
        ("linear svc", SVC(kernel="linear"))
    ])

all_models = [
    ("sgd", sgd),
    ("sgd_tfidf", sgd_tfidf),
    ("svc", svc),
    ("svc_tfidf", svc_tfidf),
    ]

unsorted_scores = [(name, cross_val_score(model, X_train, y_train, cv=2).mean()) for name, mo
scores = sorted(unsorted_scores, key=lambda x: -x[1])
print(scores)
```

Figure 5.7 Model Evaluation Using Cross-Validation

The results of model evaluation were as below:

[('svc_tfidf',0.7276632101308489), ('sgd_tfidf', 0.7248776948105147), ('sgd', 0.7235161474060376), ('svc', 0.7123038630493981)]

SVC with tf-idf features scored the highest estimated accuracy of 72.7% and was therefore chosen for creating the model. The model was then trained using the training data. The trained model was persisted using the joblib module to eliminate the need for training the model for future predictions. The model was then evaluated using the test data by calling the saved model. The code is shown in Figure 5.8 below

```
model = svc_tfidf
model.fit(X_train, y_train)

# save the model to disk
filename = 'saved_model.sav'
joblib.dump(model, filename)

# load the model from disk
loaded_model = joblib.load(filename)

y_pred = model.predict(X_test)

print("Confusion matrix")
print(confusion_matrix(y_test, y_pred))
print(accuracy_score(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

Figure 5.8 Model Training and Testing Using SVC_TFIDF

5.5 Testing the Model

The model was tested using the test data. The test data was passed to the saved model and the predicted results compared to the actual results. A confusion matrix was used to illustrate the performance of the model. The confusion matrix is illustrated in Table 5.1 below

Table 5.1: Confusion matrix

	Predicted 0	Predicted 1	Predicted 2
Actual 0	194	2	11
Actual 1	16	13	2
Actual 2	37	1	31

The accuracy score is calculated as

accuracy = total correct predictions / total predictions made * 100

The accuracy of the classification model was therefore calculated as below.

accuracy = 194+13+31 / 207+31+69* 100

= 238/307*100

=77.52%

The metrics: recall, precision and f-score were calculated using sklearn and are illustrated in Figure 5.9 below

	precision	recall	f1-score	support
0	0.79	0.94	0.85	207
1	0.81	0.42	0.55	31
2	0.70	0.45	0.55	69
accuracy			0.78	307
macro avg	0.77	0.60	0.65	307
weighted avg	0.77	0.78	0.76	307

Figure 5.9 Performance of the SVC_TFIDF Model

5.6 Predicting New Tweets

5.6.1 Collecting Tweets

To predict new Tweets, the user collects Tweets using the Twitter search API and passes them to the saved model for prediction. A Twitter application was created to obtain the necessary credentials needed by OAuth. An access key, access secret key, consumer key and consumer secret key are provided to enable the prediction program access the Twitter application. A user initiates Tweet collection using the keywords predefined in section 3.4. The program retrieves tweets from the API and stores them in a Pandas dataframe.

5.6.2 Preprocessing

The Tweets are cleaned using the preprocessing code in section 5.3. The text component is read, cleaned and stored in a Pandas dataframe. The cleaned tweets are used in the prediction of complaint, appreciation and neutral tweets.

5.6.3 Predicting Tweet Labels

The saved model is loaded. The model labels the cleaned tweets as complaint, appreciation or neutral and returns the result to the user in a dataframe. The dataframe is then written to a MySQL database using a preconfigured sqlalchemy database connection.

5.7 Visualize Complaint Tweets

5.7.1 Reading Complaint Tweets Stored in Database

The user initiates a database read on the MySQL database containing labelled tweets via a preconfigured sqlalchemy connection. The user inputs the start and end date for tweets that should be returned in the result. The complaint tweet texts are read and stored in a Pandas dataframe. A code snippet is illustrated in Figure 5.10 below.

```
# import the module

from sqlalchemy import create_engine
def read_db():
    # create sqlalchemy engine
    engine = create_engine("mysql+pymysql://{user}:{pw}@localhost:3308/{db}"
                           .format(user="root",
                                   pw="",
                                   #pt=3308,
                                   db="gov"))
    #first="2019-01-01 00:00:00"
    #last="2019-06-30 23:59:59

    tweet= pd.read_sql_query("select text from tweet WHERE class_id=0 AND date BETWEEN '2019-01-01 00
    return tweet.text
```

Figure 5.10 Database Read for Complaint Tweets

5.7.2 Location Identification Using NER

To identify locations within the result. Named entity recognition is carried out on the text using the Stanford NER tagger by Stanford Core NLP. The text is tagged using the 3 class classifier into person, organization or location. Upon observation, it was noted that the NER tagger tags location names composed of two words as separate locations. For example, Park Road is tagged as Park and Road. Such tags were merged into one using the python Groupby functionality. The result is stored in a dataframe and availed to the user. Sample tagged locations are shown in Figure 5.11 below.

```
[' Elgeyo Marakwet']
[' Muringa Road']
[' Taifa Road']
[' Nairobi']
[' Nairobi']
[' Kirinyaga']
[' River']
[' Nairobi County']
[' Nairobi']
[' Ngummo']
[' Nairobi']
[' Ngummo']
[' Nairobi']
```

Figure 5.11 Sample Tagged Locations

5.7.3 Grouping of Complaints by Location

The locations identified from the complaint tweets read from the database were clustered to remove repeats. The number of complaints per location was illustrated using a bar graph. This representation provides a quick overview of the locations with the highest number of complaints. A sample bar graph generated by the tool is shown in figure 5.12.

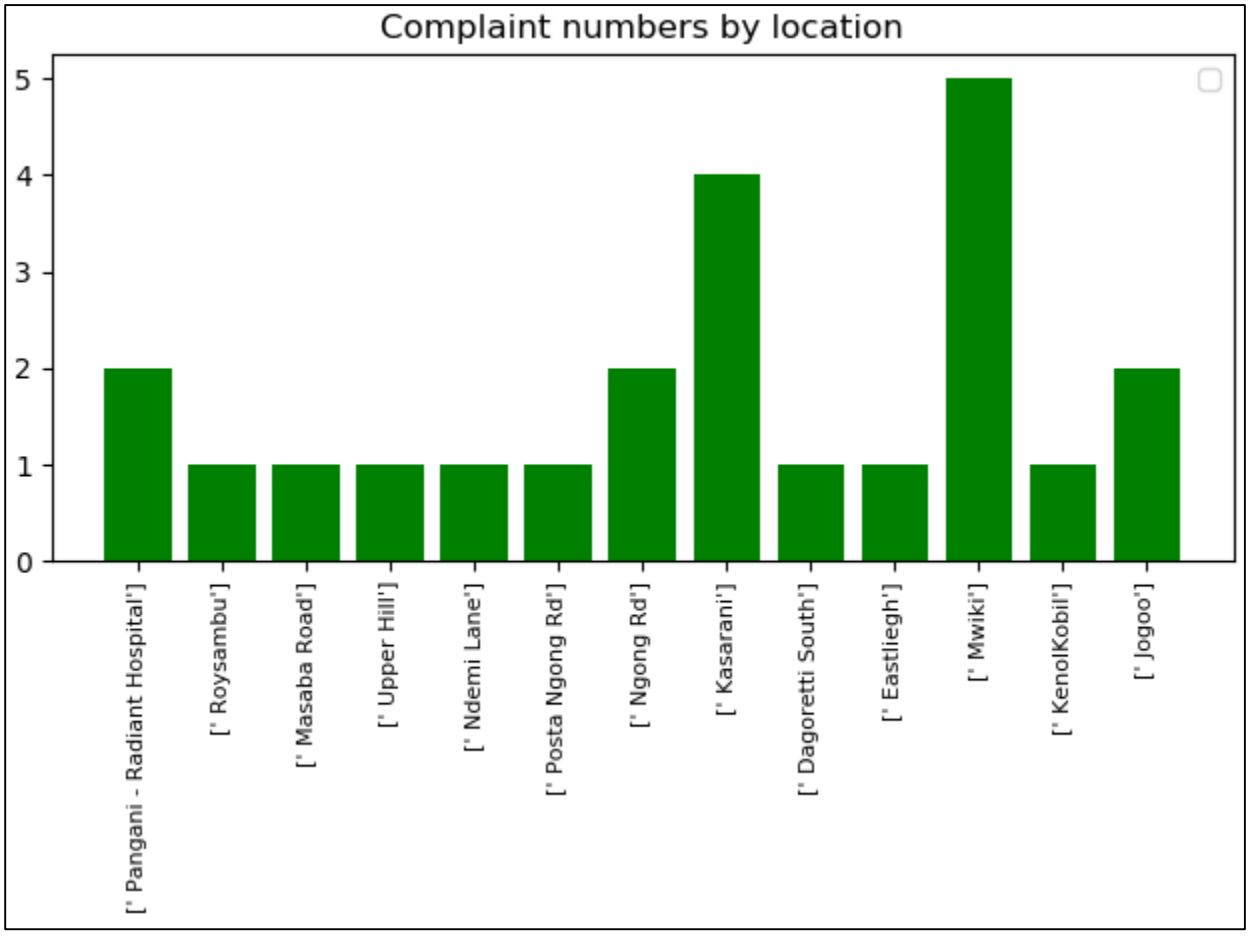


Figure 5.12 Complaints Numbers by Location Bar Graph

To facilitate further analysis, the tool was able to filter the complaints read as per the grouped locations. A sample filter of complaints by location is illustrated in figure 5.13 below.

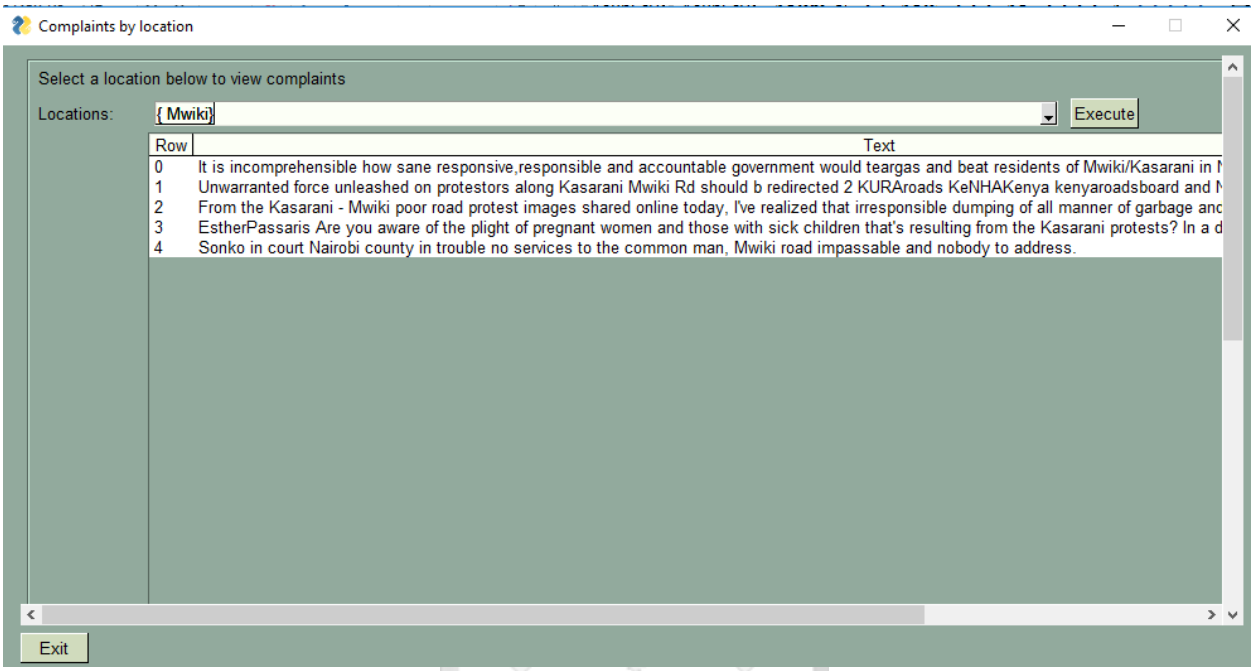


Figure 5.13 Complaints by Filtered Location

5.7.4 Geocoding and Mapping Locations

The locations identified in text format need to be geocoded and plotted on a map. The locations are in text format needed to be converted into geographical coordinates using the google geocoding API. A project was created on the google developer account as shown in figure 5.14 below.

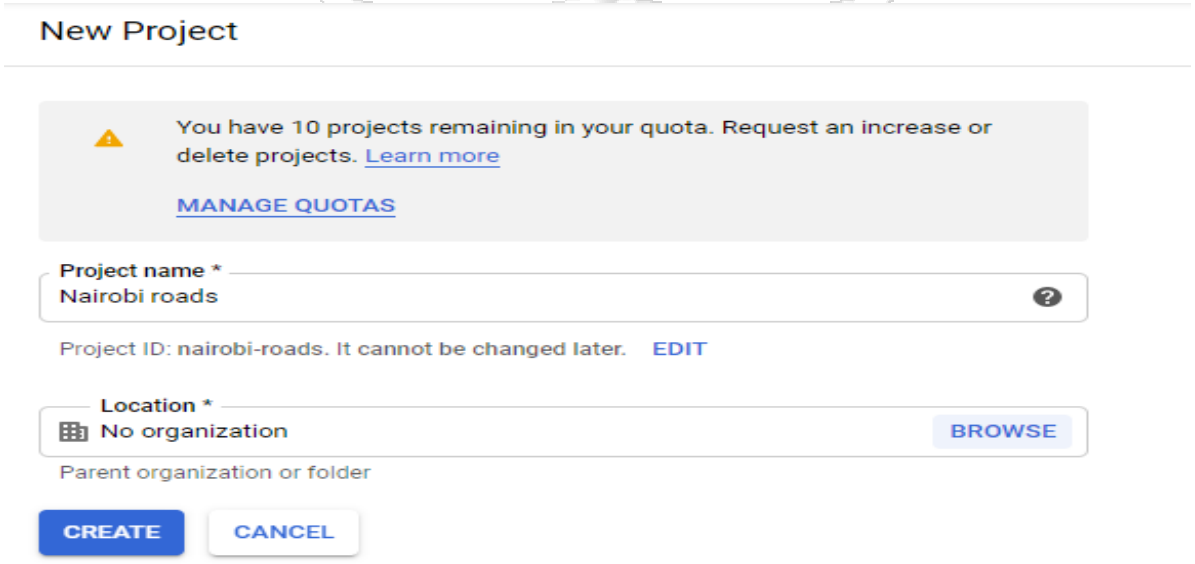


Figure 5.14 Google Developer Project Creation

The geocoding and Maps JavaScript APIs were enabled. The API key provided was used to establish a connection between the python program and the Google APIs. Reverse geocoding was performed on the list of locations and the latitude and longitude coordinates returned plotted on a static google map using the gmplot package and the Maps JavaScript API. A sample complaint map produced by the tool is shown in Figure 5.15.

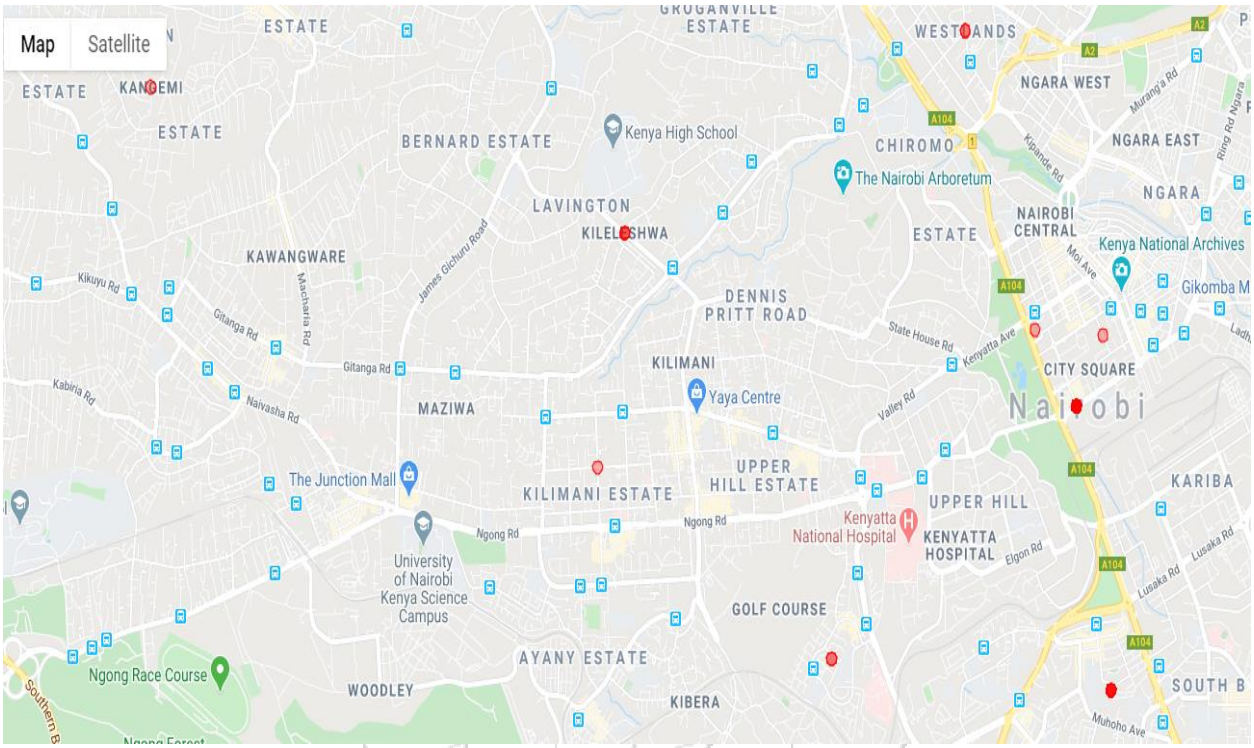


Figure 5.15 Complaint Map

5.7.5 Visualize Frequent Words

The complaint tweet text read in section 5.7.1 were visualised using the WordCloud package in Python. The text was first ridded of stopwords using the stopwords functionality in WordCloud. The resulting text was then split into tokens and converted to lowercase. A WordCloud image was then generated and displayed by the tool. Frequent words in the complaint tweets are displayed in bolder fonts. A quick analysis of the WordCloud indicates a mention of common problems related to roads in the complaint tweets, for example, potholes, congestion and traffic. A sample WordCloud image generated by the tool from complaint tweets is shown in figure 5.16 below.

Figure 5.16 Wordcloud for Complaint Tweets

Chapter 6 Discussion

6.1 Introduction

This chapter discusses the results of the research achieved in the previous chapters. The purpose of this study was to bring public complaints to the attention of government through the use of a tool that extracts and visualizes public complaint data sourced from citizens on social media using data mining techniques. The discussions are done describing how the objectives that were defined in the beginning were met by the results obtained.

6.2 Results Discussion

Existing frameworks relating to public sector complaint handling were evaluated. Two frameworks namely: 'Ladder of participation' and 'voice, choice and exit' were analysed. These frameworks help to understand the potential involvement of the citizen in public services by defining the hierarchy of citizen empowerment and how citizens exercise their influence by participating to have their concerns and complaints addressed. In general, the citizen should express their concerns to the government so that the government is aware of their wants, however, decision making should be left to government.

The study analysed current models of social media interactions in the public sector. With this regard, it was noted that the use of civic technologies translates into improved levels of interactions between citizens and their governments. This is because civic technologies facilitate citizen sourcing which enables governments to collect feedback from their large number of citizens. Social media is one of the forms of civic technology which governments are using to get feedback from citizens. Twitter was the chosen data source for the study because of the vast amount of data that can easily be obtained with fewer restrictions as compared to the other social media platforms.

The current techniques, approaches and architectures used for data mining aspects in social media were analysed. Previous studies have made use of data mining on social media using the official APIs provided. The most suitable method to mine data from Twitter is by using the official Twitter Search API or the Streaming API. However, the official APIs are only able to mine Tweets that go back a maximum of 7days. The GetOldTweets3 module which is an improvement of Jefferson Henrique's open source code was used to mine Tweets going back beyond 7 days. Natural

language processing has been widely used as a technique to extract meaningful information from social media data.

The study developed a domain specific tool for extraction and visualization of public complaints on social media. Twitter was the chosen data source for the study and roads the chosen infrastructure category. The tool developed was able to identify complaint tweets by classifying them into complaint (0), appreciation (1) and neutral (2). This was done using a SVC classifier with tf-idf features. The classifier was able to classify tweets into the predefined categories with an accuracy of 77.52%. The labelled tweets are stored in a structured format in a MySQL database which makes it easier to query the complaint tweets containing complaints.

The functionality of the developed tool was validated. The purpose of validation was to ensure that the tool was able to retrieve information that will be meaningful to the county government from the collected tweets. It was possible to query the database and view the resulting complaint tweets suited for the requested parameters. The tool was able to extract meaningful information from the complaint tweets, by using natural language processing. Named entity recognition was performed using the NLTK toolkit and the Stanford Core NLP NER tagger. Locations were identified in the complaint tweets in text format then reverse geocoded using the Google geocoding API. The results were visually presented to the user with the coordinates plotted on a google map. A visual representation of frequent words was also presented to the user in a WordCloud image.

The results of the study show that it is possible to use data mining to extract and visualize public complaints from social media. The data collected can be analysed using Natural language processing techniques to extract meaningful information from citizens that can be used by the County government for collecting public complaints from citizens.

Chapter 7 Conclusion and Recommendations

7.1 Conclusion

The purpose of this research was to develop a domain specific tool that extracts and visualizes public complaints from social media. This was intended to help Nairobi county gain feedback on service delivery from citizens by analysing the complaints collected. The tool was developed and is able to fulfil this requirement successfully.

This study evaluated existing frameworks relating to public sector complaint handling and how the governments are using these frameworks to gather public complaints from their citizens. Current models of social media interactions in the public sector were analysed and Twitter was identified as one of the best data sources for data mining on social media. The official Twitter search API being the most suitable for conducting such studies. Natural language processing approaches are key to extracting meaningful information from collected social media data. It is possible to develop a domain specific tool that extracts and visualizes public complaints based on social media data. The validation of the functionality of the tool illustrated how easy it is to obtain meaningful information from collected data.

7.2 Recommendations

Based on the findings of the research, the recommendation is to use a larger dataset for training the classification model. Better results could have been achieved in classification with the SVC_TFIDF model if a larger dataset was used for training. The dataset used was composed of 1023 labelled tweets. 70% of the labelled tweets were used for training and 30% were used for testing. The researcher acknowledges that better results would have been achieved with a larger dataset.

7.3 Future Work

Tweets are limited to a maximum of 280 characters. To satisfy this requirement, Twitter users make use of language that is made up of slang, abbreviations and non-English words. This makes the results of the classification model not as accurate as when grammatically correct English is used. Future researchers can focus on converting these terms into relevant English words to improve the accuracy of classification.

The focus social media platform on this research was Twitter only. Future research should be able to aggregate data from multiple social media platforms for use in the tool. The tool should be able to clean the various data formats into one uniform format that can be processed by the classifier during the tweet classification process. This will ensure that citizen complaints are captured across platforms that accommodate all members of the society regardless of age and their preferences.

References

Ahmad, J., Devarajan, S., Khemani, S., & Shah, S. (2005). *Decentralization and Service Delivery.* Retrieved 03 17, 2019, from The World Bank: http://www1.worldbank.org/publicsector/decentralization/decentralizationcorecourse2006/CoreReadings/Ahmad.pdf

Anggareska, D., & Purwarianti, A. (2014). Information extraction of public complaints on Twitter text for bandung government. *2014 International Conference on Data and Software Engineering (ICODSE)* (pp. 1-6). Bandung: IEEE. Retrieved from https://ieeexplore.ieee.org.ezproxy.library.strathmore.edu/document/7062658/authors#authors

Appelt, D. E. (1999). Introduction to Information Extraction. *AI Communications, 12*(3), 161-172. Retrieved from https://s3.amazonaws.com/academia.edu.documents/8361651/appisr99.pdf?response-content-disposition=inline%3B%20filename%3DIntroduction_to_information_extraction.pdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWOWYYGZ2Y53UL3A%2F20191106%2Fus-east

Arnstein, S. R. (1969). A Ladder Of Citizen Participation. *Journal of the American Planning Association, 35*(4), 216 — 224.

Aziz, R. B. (2015, July 01). The effectiveness of public service complaint management processes in contexts of autocratic governance: the case of Brunei Darussalam. Birmingham, West Midlands, England. Retrieved from https://etheses.bham.ac.uk/id/eprint/6213/5/Aziz15PhD_Final.pdf

Batrinca, B., & Treleaven, P. C. (2015). Social media analytics: a survey of techniques, tools and platforms. *AI & SOCIETY, 30*(1), 89–116. Retrieved from https://link.springer.com/article/10.1007/s00146-014-0549-4

Blase, J., & Blase, J. (2001). *Empowering Teachers: What Successful Principals Do?* (2 ed.). Thousand Oaks,Carlifornia: Corwin Press, Inc.

Born, C., Mainka, A., Meschede, C., & Siebenlist, T. (2019). Pushing Open Government Through Social Media. *Proceedings of the 52nd Hawaii International Conference on System Sciences* (pp. 3366-3375). Grand Wailea, Maui: Scholar space. Retrieved from https://scholarspace.manoa.hawaii.edu/bitstream/10125/59772/0332.pdf

Briscoe, T. (2010, 10 6). *Introduction to Linguistics for Natural Language.* Retrieved 08 01, 2019, from www.cl.cam.ac.uk: https://www.cl.cam.ac.uk/teaching/1011/L100/introling.pdf

Davis, F. D. (1986, 02 03). *A technology acceptance model for empirically testing new end-user information systems : theory and results.* Retrieved 02 15, 2020, from DSpace@MIT: https://dspace.mit.edu/handle/1721.1/15192

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease Of Use, And User Acceptance of Information Technology. *MIS Quarterly, 13*(3), 319-340.

Davis, F. D., Bagozzi, R., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35*(8), 982-1003.

Delima, R., Santosa, H. B., & Purwadi, J. (2017). Development of Dutatani Website Using Rapid Application Development. *International Journal of Information Technology and Eectrical Engineering, 1*(2), 36-44. Retrieved from https://jurnal.ugm.ac.id/ijitee/article/view/28362

Dennis, A., Wixom, B. H., & Roth, R. M. (2012). *System Analysis and Design* (5 ed.). New Jersey: John Wiley & Sons, Inc. Retrieved from http://www.saigontech.edu.vn/faculty/huynq/SAD/Systems_Analysis_Design_UML_5th%20ed.pdf

Farrell, C. M. (2010). Citizen and consumer involvement in UK public services. *International Journal of Consumer Studies, 34,* 503-507. doi:10.1111/j.1470-6431.2010.00915.x

Fatkhurrochman, Noviandha, F. D., & Setyanto, A. (2018). Twitter Classification of Public Service Complaints. *3rd International Conference on Information Technology, Information System and Electrical Engineering* (pp. 220-223). Yogyakarta, Indonesia: IEEE. doi:10.1109/ICITISEE.2018.8721006

Feldman, R., & Sanger,, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* New York: Cambridge University Press.

Garg, H., Bansal, C., Kaushal, R., & Thanaya, I. (2017). Identifying Actionable Information from Social Media for Better Government-Public Relationship. *10th International Conference on Developments in eSystems Engineering (DeSE* (pp. 206-211). Paris,France: IEEE. doi:10.1109/DeSE.2017.27

Government of Kenya. (2010, 01 01). *The Constitution of Kenya.* Retrieved 02 21, 2020, from Kenya Law: http://kenyalaw.org/kl/index.php?id=398

Government of Kenya. (2012). *County Governments Act.* Retrieved 03 16, 2019, from Kenya Law.org: http://kenyalaw.org/lex/rest//db/kenyalex/Kenya/Legislation/English/Acts%20and%20Regulations/C/County%20Governments%20Act%20Cap.%20265%20-%20No.%2017%20of%202012/docs/CountyGovernmentsAct17of2012.pdf

Government of Kenya. (2012). *Urban Areas and Cities Act.* Retrieved 03 16, 2019, from Parliament of Kenya: http://www.parliament.go.ke/sites/default/files/2017-05/UrbanAreasandCitiesAct_No13of2011.pdf

Hasby, M., & Khodra, M. L. (2013). Optimal path finding based on traffic information extraction from Twitter. *International Conference on ICT for Smart Society* (pp. 1-5). Jakarta: IEEE. Retrieved from https://ieeexplore.ieee.org.ezproxy.library.strathmore.edu/document/6588076

Henrique, J. (2017, 01 3). *GetOldTweets-python.* Retrieved 07 31, 2019, from www.github.com: https://github.com/Jefferson-Henrique/GetOldTweets-python

Hirschman, A. O. (1970). *Exit, Voice and Loyalty: Responses to Decline in Firms, Organizations and States.* . Cambridge, Massachussets: Harvard University Press.

Hu, Q. (2019). Twitter data in public administration: a review of recent scholarship. *International Journal of Organization Theory & Behavior*, 209-222. Retrieved from

https://www.emerald.com.ezproxy.library.strathmore.edu/insight/content/doi/10.1108/IJOTB-07-2018-0085/full/html#ref016

Iqbal, M., & Khan, S. (2018). Mining Facebook Page For Bipartisan Analysis. *Southern Association for Information Systems Conference SAIS 2018 Proceedings.* Atlanta, GA, USA: Association for Information Systems. Retrieved from https://aisel.aisnet.org/sais2018/42

Jianqiang, Z., & Xiaolin, G. (2017). Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis. *IEEE Access, 5*, 2870-2879. doi:10.1109/ACCESS.2017.2672677

Jumia Kenya. (2015). *Whitepaper: The Growth Of The Smartphone Market In Kenya.* Retrieved 04 11, 2019, from www.jumia.co.ke: https://www.jumia.co.ke/blog/whitepaper-the-growth-of-the-smartphone-market-in-kenya/

Khurana, D., Koli, A., Khatter , K., & Singh, S. (2017). Natural Language Processing: State of The Art, Current Trends and Challenges. *Arxiv*. Faridabad, India. Retrieved from https://arxiv.org/ftp/arxiv/papers/1708/1708.05148.pdf

Kibanya, M. (2015). Factors Influencing Customer Service Standards In County Governments In Kenya: A Case Study Of Nairobi County Government. *Strategic Journal of Business & Change Management, 2*(67), 615-629. Retrieved 03 19, 2019, from http://strategicjournals.com/index.php/journal/article/view/142/149

Kothari, C. R., & Garg, G. (2014). *Research Methodology Methods and Techniques* (3 ed.). New Delhi: New Age International (P) Limited Publishers.

Kumar, R. (2011). *Research Methodology: A Step by Step Guide for Beginners* (3rd ed.). London, Great Britain: SAGE Publications Ltd. Retrieved 04 15, 2019, from http://www.sociology.kpi.ua/wp-content/uploads/2014/06/Ranjit_Kumar-Research_Methodology_A_Step-by-Step_G.pdf

Leukis, E. (2018). Citizen-Sourcing for Public Policy Making: Theoretical Foundations, Methods and Evaluation. *Public Administration and Information Technology, 25*, 179-203. doi:10.1007/978-3-319-61762-6_8

Likert, R. (1967). *The human organization: Its management and value.* New York: McGraw-Hill.

Likert, R. (1981). System 4: A Resource for Improving Public Administration. *Public Administration Review, 41*(6), 674-678.

Lönn, C.-M., Uppström, E., & Nilsson, A. (2016, 6 15). Designing An M-Government Solution: Enabling Collaboration Through Citizen Sourcing. *EUROPEAN CONFERENCE ON INFORMATION SYSTEMS (ECIS).* Istanbul: Association for Information Systems. Retrieved from https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1071&context=ecis2016_rp

Majeed, M. J., Malik, M. M., Khan, M. T., & Khalid, S. (2016). Real-time government policy monitoring framework using expert guided topic modeling. *2016 Sixth International Conference on Innovative Computing Technology (INTECH)* (pp. 47-51). Dublin: IEEE. doi:10.1109/INTECH.2016.7845096

Maldonado, M., Alulema, D., Morocho, D., & Proaño, M. (2016). System for monitoring natural disasters using natural language processing in the social network Twitter. *2016 IEEE International Carnahan Conference on Security Technology (ICCST)* (pp. 1-6). Orlando, FL,: IEEE. doi:doi: 10.1109/CCST.2016.7815686

Mergel, I. (2013). A framework for interpreting social media interactions. *Government Information Quarterly, 30*(4), 327-334. Retrieved from https://doi.org/10.1016/j.giq.2013.05.015

Mitullah, W. V. (2016, 06 30). Retrieved 03 15, 2019, from afrobarometer.org: http://afrobarometer.org/sites/default/files/publications/Dispatches/ab_dispatchno105_kenya_devolution.pdf

Mottl, D. (2019, 11 27). *GetOldTweets3 0.0.11*. Retrieved 02 15, 2020, from PyPi.org: https://pypi.org/project/GetOldTweets3/

Müller-Seitz, G., Dautzenberg, K., Creusen, U., & Stromereder, C. (2009). Customer acceptance of RFID technology: Evidence from the German electronic retail sector. *Journal of Retailing and Consumer Services, 16,* 31-39.

Muriu, A. R. (2013). *Decentralization, citizen participation and local public service delivery.* Retrieved 03 20, 2019, from University of Potsdam: https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/deliver/index/docId/6337/file/master_muriu.pdf

Ng'ang'ira, J. N. (2018). A Prototype for mapping of tweets on state services. Nairobi, Nairobi, Kenya. Retrieved 07 15, 2019, from https://su-plus.strathmore.edu/handle/11071/6003

Nguyen, T. T., & Kravets, A. G. (2016). Analysis of the social network facebook comments. *2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA)* (pp. 1-5). Chalkidiki: IEEE. doi:doi: 10.1109/IISA.2016.7785412

Pade-Khene, C., Thinyane, H., & Mwazvita, M. (2017). Building Foundations before Technology: An Operation Model for. *European Conference on Digital Government* (pp. 118-126). Lisbon: Academic Conferences and Publishing International Limited. Retrieved 03 20, 2019, from http://collections.unu.edu/eserv/UNU:6206/ECDG17_118.pdf

Pavlou, P. A. (2003). Consumer acceptance of electronic commerce: integrating trust and risk with the technology acceptance model. *International Journal of Electronic Commerce,, 7*(3), 69-103.

Republic of Kenya,National Treasury. (2018). *2018 budget policy statement.* Retrieved 04 22, 2019, from treasury.go.ke: http://treasury.go.ke/component/jdownloads/send/195-budget-policy-statement/732-2018-budget-policy-statement.html

Ronoh, G., Mulongo, L. S., & Kurgat, A. (2018, January). Challenges of Integrating Public Participation In The Devolved System Of Governance For Sustainable Development In Kenya. *International Journal of Economics, Commerce and Management, VI*(1), 476-491. Retrieved from http://ijecm.co.uk/wp-content/uploads/2018/01/6132.pdf

Rosen, A. (2017, November 7). *Tweeting Made Easier*. Retrieved 7 23, 2019, from blog.twitter.com: https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html

SeeClickFix. (2020, 03 30). *SeeClickFix*. Retrieved 03 30, 2020, from SeeClickFix: https://seeclickfix.com/

Shafeek, S. A. (2011). E-learning technology acceptance model with cultural factors. *MSc. Thesis, Liverpool John Moores University, School of Computing and Mathematical Sciences.* . Liverpool. Retrieved from MSc. Thesis, Liverpool John Moores University, School of Computing and Mathematical Sciences.

Shagholi, R., Abdolmalki, R., & Moayedi, A. A. (2011). New Approach in Participatory Management, Concepts and Applications. *New Approach in Participatory Management, Concepts and Applications, 15*, 251-255. Retrieved 03 30, 2020, from https://www.sciencedirect.com/science/article/pii/S1877042811002618

Sommerville, I. (2011). *Software Engineering* (9 ed.). Boston, Massachusetts : Pearson Education, Inc. Retrieved 07 24, 2019, from https://edisciplinas.usp.br/pluginfile.php/2150022/mod_resource/content/1/1429431793.203Software%20Engineering%20by%20Somerville.pdf

Susilawati, E. (2016). Public services satisfaction based on sentiment analysis: Case study: Electrical services in Indonesia. *International Conference on Information Technology Systems and Innovation* (pp. 1-6). Bandung ,Bali: IEEE. doi:10.1109/ICITSI.2016.7858241

Theeyan, B. (2015). Arnstein's Ladder of Citizen Participation A Critical Discussion. *Asian Academic Research Journal of Multidisciplinary, 2*(7). Retrieved from https://www.researchgate.net/publication/322682350_ARNSTEIN'S_LADDER_OF_CITIZEN_PARTICIPATION_A_CRITICAL_DISCUSSION

Trappey, A. J., Lee, C.-H., Chen, W.-P., & Trappey, C. V. (2010). A Framework of Customer Complaint Handling System. *In Service Systems and Service Management (ICSSSM)* (pp. 1-6)). Tokyo: IEEE.

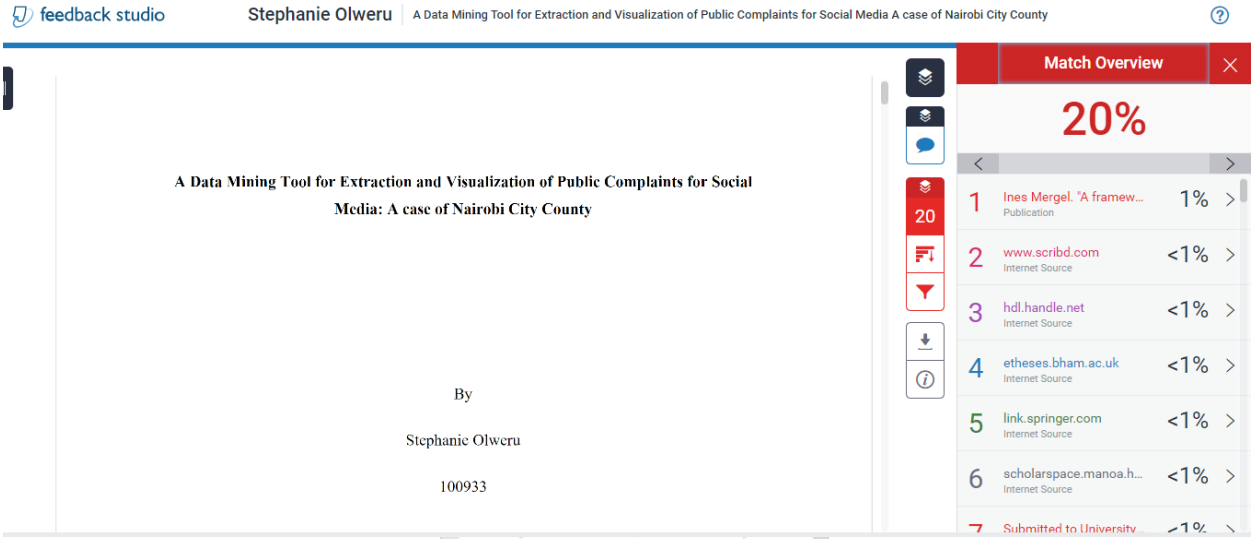
UK AID. (2015). *Urban infrastructure in Sub-Saharan Africa – harnessing land values, housing and transport.* Retrieved 03 17, 2019, from https://assets.publishing.service.gov.uk/media/57a0899c40f0b652dd0002fa/61319C_DfID-Harnessing-Land-Values-Report-1.8-Nairobi-Case-Study-20150731.pdf

Venkatesh, V., & Davis, F. D. (1996). A model of the antecedents of perceived ease of use: Development and test. *Decision Sciences, 27*(3), 451-481.

Zhou, L., Dai, L., & Zhang, D. (2007). Online Shopping Acceptance Model — A Critical Survey Of Consumer Factors In Online Shopping. *Journal of Electronic Commerce Research, 8*(1), 41-62.

Appendices

Appendix A: Originality Report



APPENDIX B: Ethical review clearance

Strathmore UNIVERSITY

28th January 2020

Ms Olweru, Stephanie
stephanie.olweru@strathmore.edu

Dear Ms Olweru,

RE: A Prototype for Public Participation on Urban Area Infrastructure Issues: a case of Nairobi County

This is to inform you that SU-IERC has reviewed and **approved** your above research proposal. Your application approval number is **SU-IERC0591/19.** The approval period is **28th January, 2020 to 27th January, 2021.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-IERC.
iii. Death and life threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-IERC within 72 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-IERC within 72 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to SU-IERC.

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology and Innovation (NACOSTI) https://oris.nacosti.go.ke and also obtain other clearances needed.

Yours sincerely,

for: Dr Virginia Gichuru,
Secretary; SU-IERC

STRATHMORE UNIVERSITY INSTITUTIONAL ETHICS REVIEW COMMITTEE (SU-IERC)
28 JAN 2020
TEL: +254 (0)703 034 000
P. O. Box 59857 - 00200
NAIROBI - KENYA

Cc: Prof Fred Were,
Chairperson; SU-IERC