UNIVERSITY OF NAIROBI
SCHOOL OF COMPUTING AND INFORMATICS

TOPIC:
A FRAMEWORK TO SUPPORT STUDY GROUP FORMATION USING ONLINE SOCIAL MEDIA -THE CASE OF TWITTER

BY:
OMUYA ODHIAMBO ERICK
P58/70565/2011

SUPERVISOR:
DR. ROBERT OBOKO

NOVEMBER, 2014

A research project submitted in partial fulfillment of the requirements of the Degree of Master of Science in Computer Science

DECLARATION

I hereby confirm that the research project that is presented on this report is my original work and that to the best of my knowledge it has not been presented anywhere else for any other University Award.

Signed: …….. Date…….

Student: OMUYA ODHIAMBO ERICK **Admission Number**: P58/70565/2011

This project has been submitted as part of fulfillment of the requirements for the award of Master of Science in Computer Science of the School of Computing and Informatics of the University of Nairobi, with my Approval as the University Supervisor.

Signed: …….. Date…….

Supervisor: Dr. Robert Oboko

DEDICATION

I dedicate this work to my beloved daddy and mummy, Charles Hezekiah Omuya and Jane Atieno Omuya, for their honest and continuous support during my entire academic life. May the Lord bless you so much.

ACKNOWLEDGEMENT

Let me begin by thanking the Almighty God for the gift of life and strength that He has given me to successfully accomplish this research. May Glory and Honor return to Him.

I am also very grateful to my supervisor Dr Robert Oboko, for his dedication in guiding me through the process of this study. I humbly appreciate the help, corrections and encouragement I received from him during the period of this research.

I also thank my wife, Spencer Atieno, for the peace of mind she has given me during my entire research period.

Finally I want to thank the Nairobi Institute of Business Studies for allowing me to use their resources to test the prototype. Special thanks to the students who commented on the system and helped me evaluate it. Thank you all and God bless you.

ABSTRACT

Online social media has positioned itself as one of the best media of communication and information sharing. People are able to write short messages on their walls using various social media like Twitter, Facebook, Bebo, hi5 and Ibibo. Through these messages they share and discuss about things like news, jokes and what they are going through. The short messages are generally called status updates and specifically tweet when you are using tweeter. Tweets have become so important in the world of information and communication because they have a great potential to pass information very fast. The knowledge generated from twitter has however not been adequately harnessed and utilized as it ought. The purpose of this study was to develop a way of searching, filtering, organizing and storing the information from social media so that it can be put to some good use. The social media in itself does not have the ability to facilitate harnessing and utilizing of the information that passes through it. This research therefore addressed this limitation by using the social media to cluster students and by so doing supported group learning.

After being developed, the prototype was evaluated at the Nairobi Institute of Business Studies with a group of twenty students being involved. The tools for data collection that were used included interviews and questionnaires the interviews especially for requirements gathering and system evaluation. The users interacted with the prototype for a period of two weeks and evaluated it based on usability and functionality.

All the students involved in the evaluation had twitter accounts. Most of them used twitter for social purposes while very few used it for academic reasons. The system was generally simple to use and so most of the users were comfortable with it. The users' response on functionality and usability of the prototype was generally positive. This study only covered development of a framework for forming groups so that current and new learners can easily get involved in academic discussion. The framework however did not capture how actually discussions can be done and facilitated. This is a component that would call for further discussion. The data in this study was descriptively analyzed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

API	Application Programming Interface
CAK	Communication Authority of Kenya
LMS	Learning Management System
ELE	Electronic Learning Environment
TV	Television
WWW	World Wide Web
SQL	Structured Query Language
NLTK	Natural Language Toolkit
NLP	Natural Language Processing

DEFINITION OF TERMS

Social Media- Social media refers to the means of interactions among people in which they create, share, and exchange information and ideas in virtual communities and networks.

Corpus- This is large and structured collection of texts or writings that are usually electronically stored and processed.

Twitter- Twitter is an online social networking service and blogging service that enables its users to send and read text-based messages of up to 140 characters, known as "tweets".

Profile- This contains someone's wall, photos and videos, a list of your friends, your favorite activities and interests, and anything else you choose to share.

Tweet – The message you post and send out to your followers is called a tweet.

Follower - A follower is a Twitter user who has subscribed to your account so he or she can see all your posts and updates on your own page.

Learning- Learning is the process of acquiring new, or modifying existing, knowledge, behaviors, skills, values, or preferences and may involve synthesizing different types of information.

E-learning – This refers to the use of various kinds of electronic media and information and communication technologies (ICT) in education.

Problem Solving Groups- This refers to classes of people that are able to handle a given task or tasks and come up with a solution.

CHAPTER 1.0 INTRODUCTION

This chapter gives an introduction to online social media and the need for clustering data obtained from social media sites especially twitter. It also looks at E-learning, group learning, problem statement, objectives of the study and the justification for the study.

1.1 Background of the Study

1.1.0 Online Social Media

In the recent past, online social media has proved to be one of the best media of communication and information sharing. This has been realized through methods like update of status, blogging, sharing of data online and social networking. People are able to write short messages on their walls using various online social media like Twitter, Facebook, Bebo, hi5 and Ibibo. Through these messages they share and discuss about things like news, jokes and what they are going through. The short messages are generally called status updates and specifically tweet when you are using tweeter. Tweets can consist of plain text, images, links or a combination of such and are normally about some event, topic of interest like music or someone's deep thoughts or personal opinion.

Tweets have become so important in the world of information and communication because they have a great potential to pass information very fast. As a result various researchers across the globe have launched serious analysis of the micro-blogging systems. Some research areas are discovering blog user characteristics (Dongwoo, Yohan, Chul, Oh, 2010) and detecting spam on social media (Yardi *et al.* 2010).

The information generated from twitter has however not been adequately harnessed and utilized as it ought. This can be done for instance by clustering the tweets and utilizing that information in various applications. Tweet clustering can assist in recommendation of users as well as viral marketing where online marketers can accurately post different advertisements to different user groups according to the common interests and classes. It can also assist in forming common groups of users that can involve in E-learning or even learning in groups.

1.1.1 E-Learning

E learning has been in place for some time now and has radically changed and positioned itself as the backbone of learning especially in post-secondary learning institutions. In these institutions, it has changed the way learning is managed right from registration to management of results and so it is actually a form of a learning management system (LMS). Learners are able to do virtually all the activities that pertain to their learning while online. This includes getting published content which they should work on and submit as well as reports on both the course and learner activities from the system.

E-learning generally facilitates teaching and learning online through network technologies and in fact is one of the most powerful responses to clamor for higher education. It employs various techniques and methods which learning institutions need to take keen interest on to ensure that e learning initiatives succeed. It can broadly be classified as either synchronous or asynchronous electronic learning.

Asynchronous E-learning is facilitated by media such as email and discussion boards and it supports discussions among learners and teachers even when both parties are not online at the same time. Participants can log on to the E-learning environment (ELE), download documents and send messages and queries to teachers or peers. This technique is flexible enough to allow participants to combine education with work, family and other commitments. Synchronous E-learning on the other hand is facilitated by media like videoconferencing and chats and supports e learners on developing learning groups and communities. This makes learning more social and it actually makes the learners feel like they are more of participants rather than isolates.

E-learning systems are built on web based technologies such as web 1.0 in which tutors author and publish content that would be utilized by the learners. The web technologies have improved from time to time due to the development of new generations of learners. Twitter is a classic example of modern web based social networking site which can be greatly used in the field of education. Even though it has not been tailor made for constructing and managing learning experiences, it has a great potential of being used in online education. Twitter as a platform gives students the opportunity to share thoughts and interact with their colleagues and teachers. The

current advanced generation of learners is able to use web based tools like twitter, Google and you tube to post ideas and questions on the online network and get immediate, relevant and up-to-date results.

The education policy makers need to relook at the concept of teaching and learning with respect to the growth in technology so as to cope with the new generation of learners. They can interrogate the current learning environment alongside available information communication technologies and the demands of the current networked society. Learning professionals can be engaged to critically analyze web based tools like twitter that learners use on a daily basis to see how best they can be used to efficiently facilitate learning. Social media has a humongous educational potential and so if analyzed properly with respect to the concepts of education, training and group learning, it can play a pivotal role in education not only as a medium of learning but also a learning platform.

1.1.2 Study Groups and Group Learning

Group working in a virtual environment is an ever increasing phenomenon both in the industry and tertiary education. In most institutions online modes of study have been used for external students. This can however be added to the class-room based traditional teaching methods (Light et al. 2000).

Group learning is a technique of learning that involves interaction between students, their colleagues and tutors who may meet several times either physically or virtually within some set period of time. The learning tends to be focused upon the discussion of predefined subject area and enables the participants to gain greater understanding of class material, share study tips and ideas, complete class projects quickly, establish new networks with friends and familiarize with business practices by learning how to work as a team.

Group learning can be easily be supported by information technology especially in networked environments. This is where designed systems can be used to mediate through various cognitive learning practices in a group setting. Such systems can actually enable participants to develop documents that express certain theory with a view to developing them more and more as time progresses. The members participating in group learning can develop a more advanced document

that contains more issues, concepts and technical arguments that would have been very difficult to produce if only one person were involved. This is because participation in the group process would enable the members to think through the theory being discussed in different perspectives and so come up with a more concrete result. The group process can better be achieved when it is done inline.

A model of teaching and learning has been put forward by Salmon in an online environment based on five distinct stages of online learning, characterized by differing needs of students, and differing nature of the interactions encountered. (Salmon, 2000)

- *Stage one*, Access and motivation, relates to ensuring students have access to the system and providing an overview of the process of computer mediated communication and reassurance to students of the availability of necessary support structures.

- *Stage two*, Socialization, involves the encouragement of the student to engage in online interactions and allows time for students to become familiar with the use of the technology within a communication process.

- *Stage three*, Information Exchange, sees the learners beginning to engage with information relating to learning outcomes.

- *Stage four*, Knowledge construction, involves the learner becoming more focused on the content matter, and taking more responsibility for their own learning, including openly collaborating with others.

- Stage five, Development, involves reflection on the learning process including identification of process skills developed as well as content knowledge.

1.2 Problem Statement

The theory underlying interaction between components of learning processes has seen the development of new frameworks that have been used to manage the frameworks. However, there are new and better ways that are needed to enhance understanding of both the processes and their interactions. Most of these frameworks are technological and are implemented through the social networking web sites. Since the implementation of social networking web sites, their use has gradually increased in different areas of the society. The main group that has quickly embraced the systems worldwide is that of the youthful age group among which the students fall. These students have managed to use the social network sites often and even ended up creating the virtual learning set ups that have enabled them to handle a lot of applications in relation to teaching and learning resources. This has been made possible by the fact that students generally prefer consulting their colleagues first in the event that they need some information or a set of solutions to certain problems.

The social media sites like facebook and twitter give a classic platform for students to engage in these kinds of interactions which can enhance knowledge creation and sharing. Even though a lot of students use these social networks to interact such that a lot of knowledge is created, this knowledge is generally wasted because there is no clear way of harnessing and applying it. This implies that as the knowledge being generated takes only a while in the network then the same disappears. There should therefore be a way of searching, filtering, organizing and storing the information so that it can be put to some good use. The social media in itself does not have the ability to facilitate harnessing and utilizing of the information that passes through it. This research therefore addressed this limitation by using the social media to cluster students and by so doing supported group learning.

1.3 Project Goal

This research aimed at developing a framework for harnessing information from social media and using it to classify students into study groups for group learning.

1.4 Objectives

i. To study group formation in relation to social media and whether the groups can translate to problem solving groups.

ii. Create a way of harnessing data from social media users.

iii. Establish relevant attributes of twitter users that can be useful for the clustering purposes.

iv. Develop a classifier that is able to use the above data to create clusters (study groups for students).

v. Develop a set-up that is able to classify a new twitter user (student) into one of the clusters.

1.5 Justification

In the current society, social media has turned into an inevitable means of interaction between people at different levels. People use the social media like facebook and twitter time and again to share digital information through posting information on the wall and chatting. This has been replicated even in the learning arena where the social media has provided a superb platform for quick information sharing and learning. This has been stimulated by the fact that students generally prefer consulting their colleagues first in the event that they need some information or a set of solutions to certain problems and the social media is the quickest way of doing that. Even though a lot of students use these social networks to interact such that a lot of knowledge is created, this knowledge is generally wasted because there is no clear way of harnessing and applying it. This implies that as the information is being generated it takes only a while in the network then the same disappears. There should therefore be a way of searching, filtering, organizing and storing the information so that it can be put to some good use. This research therefore addressed this limitation by using the social media to cluster students and by so doing supported group learning.

1.6 Summary of the Solution (Conceptual Model)

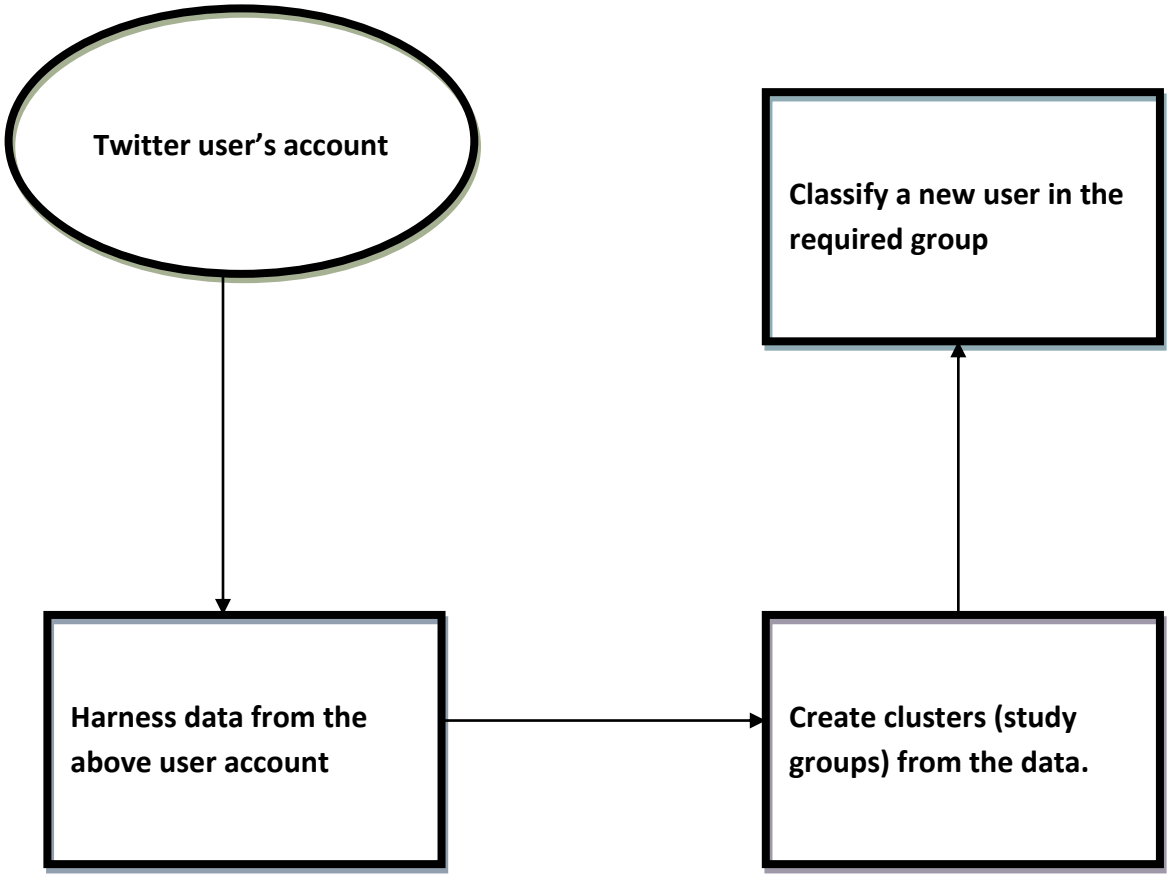


Figure 1.0 The Solution

This diagram (figure 1.0) above illustrates the solution. There had to be one or more users with twitter accounts from which data was extracted and used as input to the classifier. The data also had to be preprocessed. It was then used to create clusters which were equivalent to the study groups for students. When a new user or student was identified, he or she had also to be classified into one of the groups that were created.

CHAPTER 2.0 LITERATURE REVIEW

2.1 Introduction

This review explores a number of aspects which include study groups and group learning dynamics especially in a virtual environment, development of electronic learning, clustering tools, the internet, social media growth in Kenya and their application in learning management systems and group learning.

2.2 Group Learning Dynamics

Group learning is a technique of learning that involves interaction between students, their colleagues and tutors who may meet several times either physically or virtually within some set period of time. The learning tends to be focused upon the discussion of predefined subject area and enables the participants to gain greater understanding of class material, share study tips and ideas, complete class projects quickly, establish new networks with friends and familiarize with business practices by learning how to work as a team. "A learning group is characterized by a willingness of members to share resources, accept and encourage new membership, regular communication, systematic problem solving and preparedness to share success (Brook & Moore, 2000)".

Groups working especially in a virtual environment are an ever increasing phenomenon both in the industry and in tertiary education. Whilst an online mode of study has often been used for external students, it has more recently been a useful addition to classroom based, traditional teaching methods (Light, 2000). Group dynamics characterize online groups that operate for instance in a tertiary education undergraduate environment. The group dynamics literature has been so common in disciplines such as psychology, management science and now information communication technology which focus on virtual environments. According to researchers, it is the absence of proximal face-to-face interaction between members of virtual teams that makes them virtual and distinguishes them from traditional teams (Bell & Kozlowski, 2002). Group working only happen in a team which consists of members with complementary skills. All these people work towards achieving a common purpose and generally hold each member mutually

accountable. Teamwork is very important in the development of a group learning environment and in achieving the desired learning outcomes for a course (Yuen, 2003).

Group dynamics encompasses how various members of a group interact with each other and how this affects their interpersonal skills as well as their task performance. The processes of group formation and management dictate the manner in which the members of the groups will interact and the kind of results realized from the groups. There are models that can be used for group formation and interaction for instance the Tuckman model. Tuckman (2001) identifies five stages in group development, each possessing a particular group pattern of interpersonal relationships and the content of interaction relating to the task. The stages include group formation, storming, norming, performing and adjourning each of which has observable behaviors and actions. This model can be applied in analyzing interaction between students in electronic study groups in conjunction with various teaching and learning models.

2.3 Development of Electronic Learning

Most authors describe e-learning as the access to learning experiences through the use of technology. Electronic learning involves content strictly being accessible using technological tools that are either web-based, web-distributed, or web-capable (Nichols, 2003). However some scholars like Ellis slightly disagree. Ellis and group believe that e-Learning not only covers content and instructional methods delivered via CD-ROM, the Internet or an Intranet (Benson et al., 2002; Clark, 2002) but also includes audio- and videotape, satellite broadcast and interactive TV. E-learning as a learning platform provides accessibility, flexibility, connectivity and is able to provide varied interactions (Hiltz & Turoff, 2005). E-learning has evolved in different ways in business, education, the training sector, and the military and in fact it currently means different things in different sectors.

In the field of education, the development of e-learning was fueled by the emergence of the idea of building machines that could aid teaching and learning. Some developments were realized especially in the 1950s. In the early 1980s when the desktop computers were developed, education applications for individual users were designed. These applications could however not

be used in a network as they were basically stand-alone. All the same, this development gave a great foundation and led to a great milestone in realizing the dream of developing computer aided learning systems.

In the early 1990s, the concepts of networking become more developed with the advent of the client-server paradigm. With this development it was possible for files, data and applications to be stored in remote servers such that through the use of client workstations people could access the content of the servers. The introduction of the World Wide Web (www) gave a fresh breath to e-learning as it now made it possible to link up millions of files, data, servers and generally e-learning resources using uniform resource locator. These resources could be accessed through the internet. This therefore meant that one would be able to access any resource available online from wherever they were.

The kind of interface and nature of content that the online presentations provided however led to multiple challenges being realized in the use of the systems. The content was simply for reading, looking and taking an online exam. A good number of people had problems browsing the e-learning content because they were largely static and not interactive. They did not have an interactive set of instructing guidelines which would keep the learners engaged and have an interesting experience in learning. These challenges have however been sorted through the development of more flexible, demonstration based and interactive platforms aided with tools like you tube.

2.4 The Internet and Social Media Growth in Kenya

Internet access and the social media play a major role in the success of electronic learning in any country. The growth of internet access in Kenya has increased exponentially hence making it easy to implement the e-learning systems. Various surveys have been conducted that actually attest to this fact. According to the survey done on internet penetration in African countries, Kenya is ranked fourth among Africa's top internet countries as at December 31 2011 figures. Nigeria is ranked the number one country with 45 million users but this is attributed to its huge population of over 155 million people. The performance of Kenya is way above even South Africa which is much more developed. This is illustrated in the chart below.

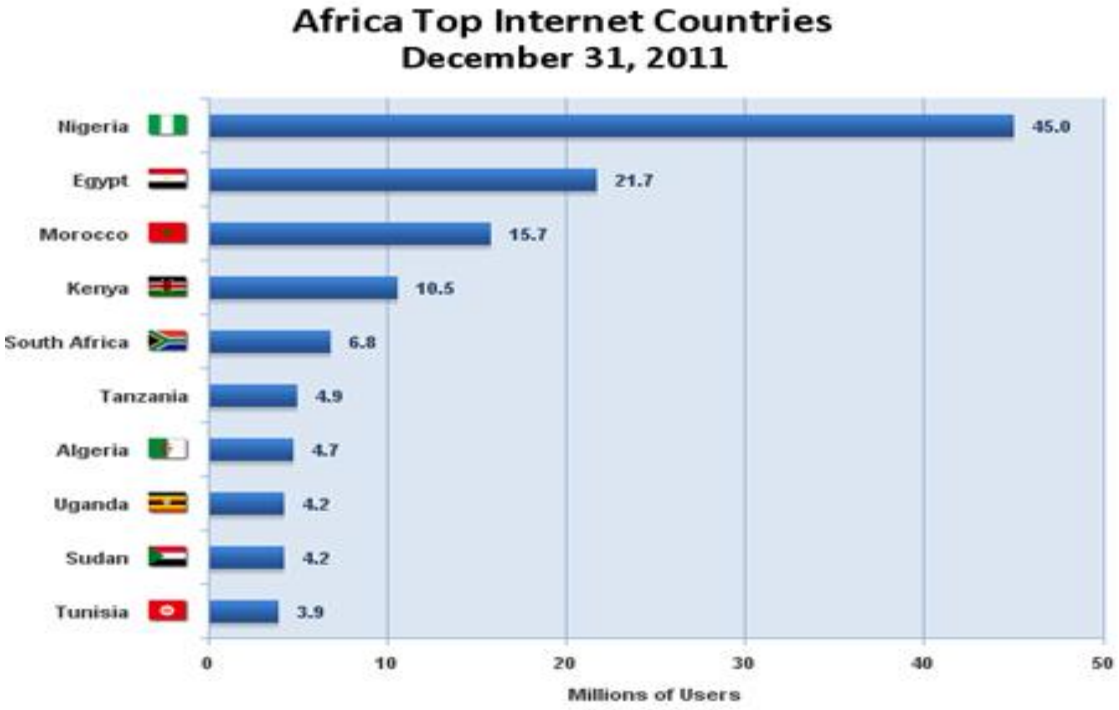
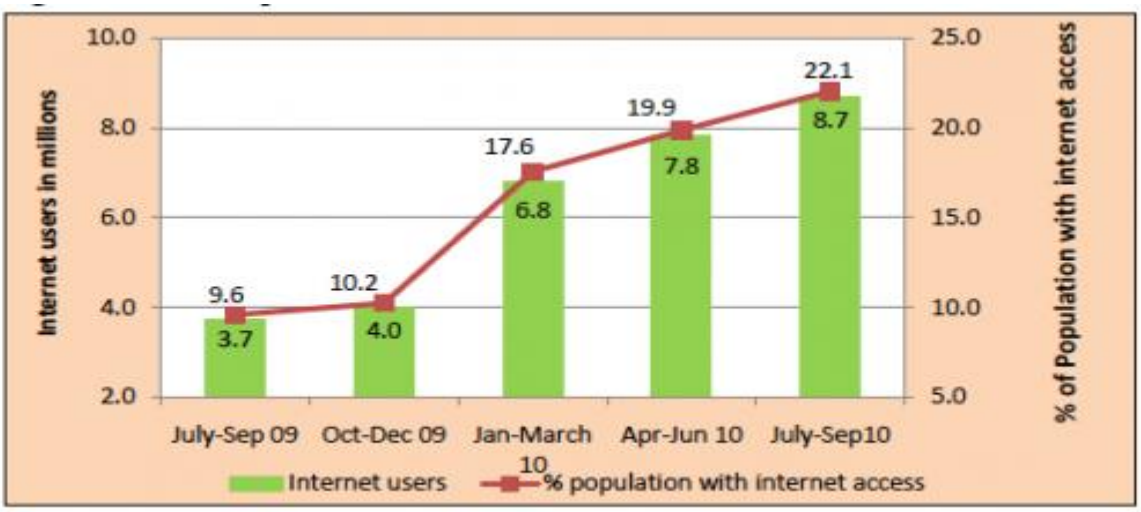


Figure 2.0: Internet penetration in Africa

The Communications Authority of Kenya also released a report that shows a tremendous increase in internet usage in Kenya due to increased use of internet enabled mobile phones. Part of the report says, "Kenyan Internet users increased by 95.63 % in the last one year showing a tremendous growth in the country's technology fuelled by high number of mobile phone usage, reveals the Communication Authority of Kenya, CAK, 2011/2012 sector report". The report indicated that the number of Internet users grew to 17.38 million as at December 2011 compared to 8.89 million users in the previous year. This, compared to the previous quarter, represents a growth of 21.55 percent. CAK statistics indicated 14.3 million Internet users in the previous quarter. CAK attributed the increase to intensified promotions on social media by mobile operators. The use of internet had been on a rising trend, with the figure showing that 44.12 percent of the populations have access to the Internet with majority accessing internet through mobile phones. This is illustrated in the chart below:



Source: CCK, Operators' returns

Figure 3.0: Internet growth in Kenya

If the above statistics are anything to go by, it goes without saying that internet penetration is on the rise in Kenya. The young people have greatly contributed to this because they are the ones that mostly use the social media either using computers or mobile phones. Most of the populations that are in schools and colleges are young and so the people in the education sector must wake up and look at how social media can be used to facilitate learning that targets students while they are in the environment that they cherish i.e. the social media.

2.5 Social Media and Learning

In the recent past social media has been used for basic communication and general interaction between people at different points. The social media like twitter and facebook has a great potential of being used as learning platforms. Twitter for instance is one of the most popular micro-blogging systems are being used often because of its elegance, robustness and simplicity. One only needs to create an account and immediately begin to twit. People generally use twitter to communicate with each other, seek for clarifications, post their points of view on certain issues, support and advise others. It is generally used to communicate in real time and the results can be received through web, SMS, and other instant messaging clients.

Twitter has a great potential of facilitating e-learning as it can create virtual learning communities where different groups of people can share a lot of information and do very

constructive academic discussions. Through twitter chartrooms, we can have platforms for engagement in which people can brainstorm and seriously ventilate on various issues. Twitter chats are a fantastic way of sharing informal knowledge and sparking conversation within a field or even inside an organization. This can enable generation of numerous amounts of information that can be very useful in facilitating innovations in various academic areas. The articulate twitter page is a good resource for articulate users since it provides news, tutorials, and links to helpful tips and tricks.

Jane Hart, a social media and learning consultant, classifies Twitter and other micro-blogs as tools for personal and informal learning. "The point of social media is to turn learning into a more participatory activity," she says. Learners use social media tools to ask and answer each other's questions, and as Hart maintains, "Micro-blogs can support collaboration and understanding." Many educators already use micro-blogs to create community around a class or an activity. Instructors who've used Twitter say it is a useful back channel during and after class. "As an instructor, you can have immediate feedback on the relevance of your class," Hart says. After class, instructors can encourage micro-blogging to support relationships among the people from the class and to further their learning. Teachers post tips of the day, questions, writing assignments, and other prompts to keep learning going. Some believe that Twitter is even more powerful as a social learning tool outside the context of the classroom.

Another popular use of Twitter and other micro-blogging sites is the building of professional networks. Michele Lentz, a technical writer and professional blogger, began using Twitter to get to know other learning professionals. Within months, she was posting regular updates about her work, getting help from experts, and attracting followers of her own. Currently, Lentz has 1,000 followers on Twitter and teaches courses on how to use micro-blogging as a learning tool. She recently polled her followers via a Twitter polling application, about why they like Twitter. The top reasons were: it accelerated their learning curve; it helped them with personal learning and also expanded their learning circle.

Twitter can also be used as a tool for exploring group learning and writing. It promotes writing as a fun activity, fosters editing skills and develops literacy skills. It can give students a chance

to record their cognitive trails and then use them to reflect on their work. Students can also use tweets to send out questions and observations to the group while engaged in classroom activities. These are just but a few areas that show the great potential that twitter has in education and learning.

The use of twitter as a learning platform currently however has a lot of challenges that have been met. As it is used now, there is no true system for filtering, searching and organizing information. The speed at which information is generated is the same speed at which it disappears into older posts. There is not really a provision for the knowledge that is being generated to be classified, saved, and easily and quickly retrieved. This results in a great loss as the information that has been generated cannot be easily reserved. This can however be sorted through the use of current technologies and the findings of this research.

2.6 Clustering Tools

2.6.1 Introduction

Clustering can be considered the most important unsupervised learning problem. It deals with finding a structure in a collection of unlabeled data. It is the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering. It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection).

Clustering users has been done based on user interest. This involves computing user similarity leveraging both textual contents and social structure, according with Twitter's role, not only a news media but also a social network. These features include tweet text, URLs, hashtags,

following relationship and retweeting relationship, all of them are closely correlated with user's interests. Then similarity is used as a measure to cluster users (Boyd et. al, 2008)

Clustering can be done through topic models. This representation is convenient to compute document similarity and perform clustering. Topic models do not make any assumptions about the ordering of words (Steyver & Griffiths, 2007). This is known as bag-of-words model2. It disregards grammar as well. This is particularly suitable to handle language and grammar irregularities in Twitter messages. Each document is represented as a numerical vector that describes its distribution over the topics.

Banerjee, Kang and Rangrej managed to cluster rich site summary (RSS) feed items. They achieve improvement over a baseline expanding the vectors to include key concepts returned by querying Wikipedia with the content of the feed (Banerjee et al, 2007). Kang used affinity propagation algorithm to cluster similar tweets and Rangrej conducted a comparative study, comparing three clustering algorithms: kMeans, singular value decomposition and affinity propagation. Experimenting on a small set of tweets they conclude that affinity propagation is best suited for short, though not so sparse, texts1. Scalability is not addressed in their comparison.

Pak and Paroubek developed a Naive Bayes algorithm with discretionary feature selection to analyze the emotion embedded in each individual tweet. Though topic classification and sentiment polarization analysis provide some useful information on the public behavior and mood, they fail to answer questions like why people are happy?" and which aspect people like the subject of interest?" (Pak and Paroubek, 2010).

Clustering tools can be applied in a number of areas ranging from data mining to computational lexicography. This research is interested in document clustering which may include recall in information retrieval, browsing a collection of documents for purposes of information retrieval and organizing results provided by a search engine. Various algorithms exist that can be used to do clustering. Some of these include:

2.6.2 Hierarchical Agglomerative Methods

The hierarchical agglomerative clustering methods are most commonly used. The construction of a hierarchical agglomerative classification can be achieved by the following general algorithm.

- Find the 2 closest objects and merge them into a cluster
- Find and merge the next two closest points, where a point is either an individual object or a cluster of objects.
- If more than one cluster remains , return to step 2

Individual methods are characterized by the definition used for identification of the closest pair of points, and by the means used to describe the new cluster when two clusters are merged. There are some general approaches to implementation of this algorithm, these being stored matrix and stored data, are discussed below:

In the second matrix approach , an N*N matrix containing all pair wise distance values is first created, and updated as new clusters are formed. This approach has at least an O(n*n) time requirement, rising to $O(n^3)$ if a simple serial scan of dissimilarity matrix is used to identify the points which need to be fused in each agglomeration, a serious limitation for large N.

The stored data approach required the recalculation of pairwise dissimilarity values for each of the N-1 agglomerations, and the O(N) space requirement is therefore achieved at the expense of an $O(N^3)$ time requirement.

2.6.3 The Single Link Method

The single link method is probably the best known of the hierarchical methods and operates by joining, at each step, the two most similar objects, which are not yet in the same cluster. The name single link thus refers to the joining of pairs of clusters by the single shortest link between them.

2.6.4 The Complete Link Method

The complete link method is similar to the single link method except that it uses the least similar pair between two clusters to determine the inter-cluster similarity (so that every cluster member is more like the furthest member of its own cluster than the furthest item in any other cluster). This method is characterized by small, tightly bound clusters.

2.6.5 The Group Average Method

The group average method relies on the average value of the pair wise within a cluster, rather than the maximum or minimum similarity as with the single link or the complete link methods. Since all objects in a cluster contribute to the inter –cluster similarity, each object is , on average more like every other member of its own cluster then the objects in any other cluster.

2.6.6 Text Based Documents

In the text based documents, the clusters may be made by considering the similarity as some of the key words that are found for a minimum number of times in a document. Now when a query comes regarding a typical word then instead of checking the entire database, only that cluster is scanned which has that word in the list of its key words and the result is given. The order of the documents received in the result is dependent on the number of times that key word appears in the document.

2.7 Conclusion

Going by the discussion above, it goes without saying that social media has a great of potential in the area of education especially e-learning. It can create very good virtual platform for group learning in which peers can meet and discuss a lot of issues hence contribute to knowledge creation in various fields. Currently, people who participate in the discussions have to sign in then just join the groups that they are working with. Through this project were able to generate the groups by clustering using twitter based information before the discussions were initiated.

CHAPTER 3.0 RESEARCH METHODOLOGY

3.1 Introduction

In this chapter, we look at how the prototype for creating discussion groups was developed as well as a detailed explanation of the research method that was used to realize the objectives of the study.

3.2 Methodology for Developing the Prototype

The research methods used in achieving the objective of this project are discussed here. The system design methodology was incremental prototyping. In incremental prototyping, the whole requirements are broken down into building blocks which are incremented each time a new component is integrated based on an overall design solution. Typically development starts with the external features and user interface, and then adds features as prototypes are developed. Requirements and Architectural Design can be done up front and then each prototype developed as the project progresses. The solution is complete when all the components are in place.

The general model of the prototype was created and then the other features were added incrementally step by step until the objectives of the system were met. This was part of an implementation of the top down approach in systems development. In a top-down approach an overview of the system is formulated, specifying but not detailing any first-level subsystems. Each subsystem is then refined in yet greater detail, sometimes in many additional subsystem levels, until the entire specification is reduced to base elements. Once these base elements are recognized then we can build these as computer modules. Once they are built we can put them together, making the entire system from these individual components.

This methodology was selected because its approach generally reduced the development cost and time. For each level of development, there was an expected output which was a part of the overall expected solution. This was evaluated against the individual level deliverables as well as the overall objectives of the proposed prototype. The whole process involved establishing the requirements specification and determining the structural design of the prototype.

3.3 Requirements Specification

This research majorly targeted a group of people that were quite conversant with the use of computers and by extension the social media. This is the new generation of learners that are characterized by a lot of curiosity of learning new things very fast. The nature of this group of learners is the main thing that will push the stakeholders in education to change the manner in which learning is actually done.

The process of requirements gathering for developing the prototype for discussion group formation involved analyzing how e-learning and group formation was done. This enabled understanding the pros and cons of the process currently and seeing how this can be done easily using the proposed methodology. The case study was done at Nairobi Institute of Business Studies in Kiambu County. There is an option of e-learning or distance learning in which students who are not able to attend classes register and get materials online. These materials include notes, assignments, exercises and even examinations.

In order to establish and understand the dynamics of this kind learning of learning, a group of 10 e-learning students at the institution were interviewed. This majorly captured how discussions groups were formed, how group discussions were done and the challenges that were met when conducting these processes. The students first confirmed that indeed, social media has a great potential for supporting teaching and e-learning. In order to confirm that the students actually understood the processes of group formation and learning, they were expected to describe how they conducted the process. Through this we could identify the major challenges that they met.

There were common shortcomings of the current way of doing things which were likely to come up. Firstly, it would take too long for the e-learning students to link up and interact between themselves so that they could agree on the nature of groups to form. This is because there was no automated system that is used to facilitate this process hence it was largely done manually. Secondly, it was generally difficult for the students to agree on how to form the groups in terms of the nature of the groups. This was because at a glance, they were not able to understand who to pair with in relation to their seriousness and the areas in which they were interested. The other challenge that was encountered was accessibility of the systems. Some of the learners were in

remote areas where they were not able to access computers with internet connectivity easily hence they were not in constant touch with their colleagues. These students however had to find a way of being connected because the course itself was online and so were the activities of our prototype. Analysis of these challenges played a pivotal role in development of the prototype.

3.4 Structural Design

We used a twitter application programming interface majorly in the development of our prototype. There are a number of activities that were performed to come up with the system.

The first task was to retrieve details of each of the students from their twitter accounts using an extension script which is part of the twitter API.

The second task involved identifying the right kind of data to use for training the expected prototype as well as testing it. This generally dealt with preliminary processing of the data collected from the users to do away with any inconsistencies and outliers. These unwanted features are not very good because they can easily cause the system to perform irregularly.

The third step involved using the data already preprocessed above to train the prototype. The kind of training we used was unsupervised learning in which the system was given the data so that it automatically analyzed and created clusters of it. The relationship between the data items was established using the k-nearest neighbor technique. From this, we identified the groups that students fell which were then turned into discussion groups.

The fourth step was testing the prototype. The end result of the learning process was the model which was able to do classification with very minimal margins of error. The prototype was then subjected to testing using the test data. This is a collection of data whose class labels are already known. They are part of the data that was used to train the system but its results are already known. They were used to confirm that the system indeed accurately did the classification given some data items.

Finally, we used the model to classify a new user into a group. This involved picking the details of a new student from twitter and trying to predict the class hence group that he could join. The illustration of the process for prototype development is given below:

.

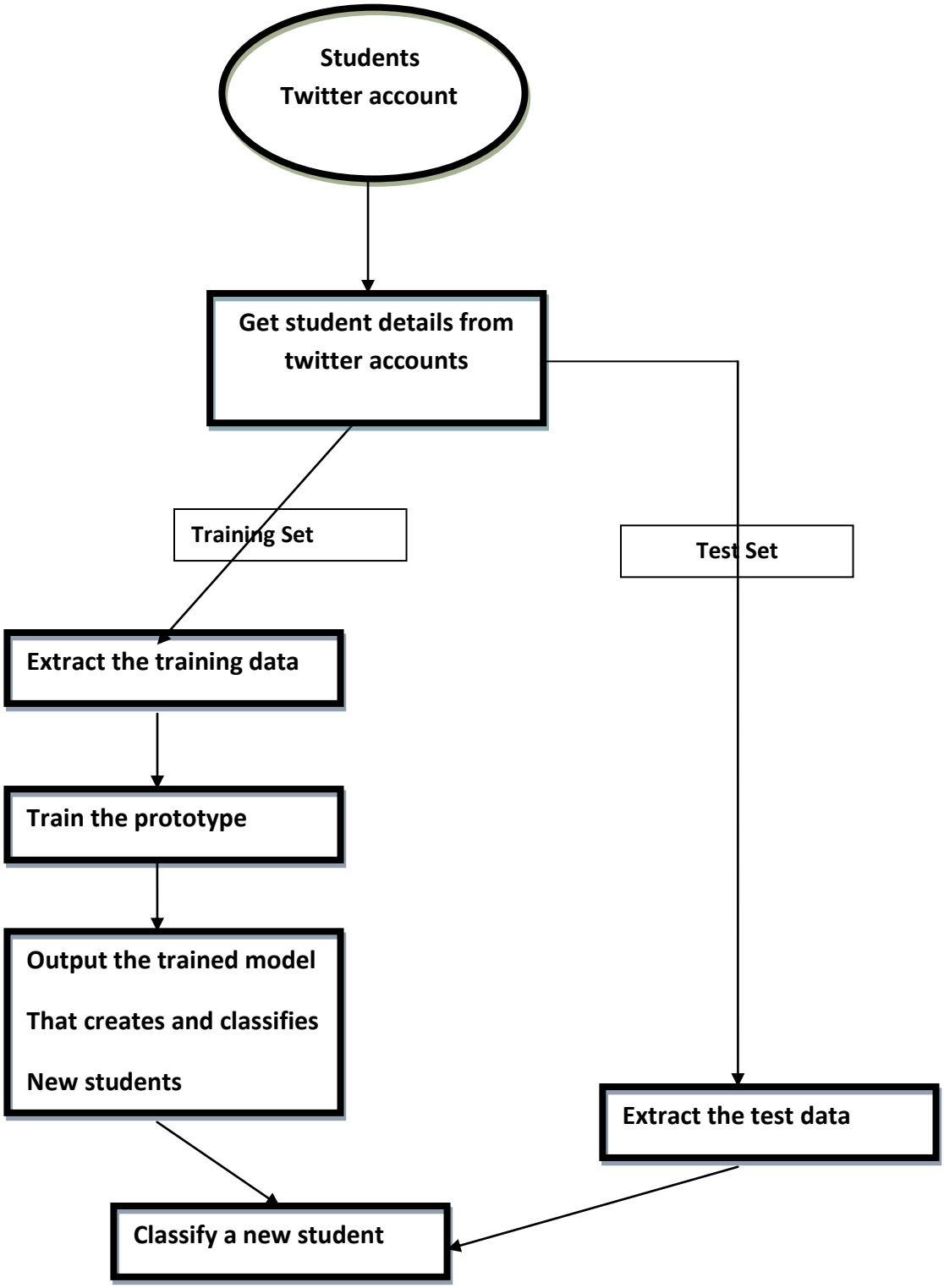


Figure 4.0: Detailed illustration of proposed solution

3.5 Prototype Evaluation

In order to evaluate the system performance, a group of twenty business students at Nairobi Institute of Business Studies were used. The exercise took a period of two weeks which involved some students being used for training while others for testing. These students were supposed to be quite conversant with the use of computers and so by extension the social media. They were required to have twitter accounts that have been used for a period of time. Those who do not have, however, were required to open their accounts immediately and use for some time. This is because it is the cumulative data in the twitter accounts that were collected and used in the prototype. The prototype was then used to extract data for fifteen students and using various attributes clusters them so as to form discussion groups. These fifteen students were used to generate the training data. The other five students were then classified into the existing groups depending on the attributes. The students were able to confirm the groups that they belong to and possible link up online and do collaborative learning. At the end of the two weeks the students were required to evaluate the system based on usability and functionality. In order to check the usability and functionality of the system, a questionnaire was used.

CHAPTER 4.0 PROTOTYPE DEVELOPMENT AND IMPLEMENTATION

4.1 Prototype Design

The methodology used in developing this prototype is incremental prototyping. This is one of the methods of Rapid Application development. The process started by developing a general model of the prototype that captures all the main features that would be used to achieve the overall objectives of the system. This also captures the overall objectives of this study. The general model was then broken down into various components depending on the objectives of the study. The summary of the design process is captured in the illustration below.

	Objective	**System Features**
1	Design how to extract data from twitter	Twitter Extractor (API)
2	Design a classifier of the above data into clusters	-Data Preprocessor -Naïve bayes classifier
3	Design a platform for storing the clustered data and classifying a new user	A database (SQL) for storing the clusters and enabling easy retrieval for comparison.

Figure 5.0: Summary of Design Process

4.2 Tweeter Application Programming Interface (API)

The tweeter Application Programming Interface played a pivotal role in extracting the tweets that would be used as the input to the prototype. For the purpose of this study, tweeter users that were identified posted comments to a tweeter account guided by a hash tag e.g. programming. The tweets extracted by the extractor script would then be fed into the classifier. According to the tweets posted, the users were classified into various categories based on the programming task.

4.3 Data Preprocessing

The nature of data that is used in machine learning is normally very important. It is one of the main factors that affect the success of machine learning. It is generally difficult, for instance, to retrieve knowledge accurately from data during training if the data that is being used is irrelevant, redundant and noisy. Such kind of data is actually unreliable if used in learning. It is therefore important to preprocess data, inasmuch as the process may take quite long. The main activities that data preprocessing involves include: data cleaning, data integration, extraction of attributes and selection (Han J, Kamber M, 2006).

The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: −100), impossible data combinations (e.g., Gender: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis (Lu H et al, 1996). The data extracted from twitter is not suitable to use in the classifier as raw as it is. It therefore needs to undergo preprocessing so as to give better results.

The process of data pre-processing can take quite long if a big chunk of data is noisy and unreliable. In this study a script was written to run through the data so that if it encountered any outliers, they were removed. Most of these would be elements that were not really relevant to the group classifier.

Some of the areas of concern included:

Tokenization: This involves splitting a sentence into its constituent tokens. For segmented languages like English, the existence of whitespace makes tokenization relatively easier.

Stemmer: In order to reduce the size of the initial feature set, we remove misspelled or words with the same stem. A stemmer (an algorithm which performs stemming), removes words with the same stem and keeps the stem or the most common of them as feature. For example, the words “train”, “training”, “trainer” and “trains” can be replaced with “train”.

Removal of Stop words like a, is, the and with. The full list of stop words can be found at Stop Word List. These words do not necessarily add value to the classifier hence they were removed.

Repeating letters – This captures letters in the tweets that sometimes people repeat so as to stress the emotion, for instance, gooooosh, gaaaaaoosh for 'gosh'. We can look for two or more repetitive letters in words and replace them by 2 of the same.

Punctuation – this can involve removing punctuation marks like a comma, single/double quote, question marks at the start and end of each word. E.g. beautiful!!!!!! Replaced with beautiful

4.4 Classification

Classification is the process that involves predicting fixed categories or groups of objects depending on a given set of attributes. A model or a classifier is built to do the prediction of labels. There are three methods of classification namely supervised learning, unsupervised learning and reinforcement learning. Supervised learning involves a case where the categories that data is assigned to are known before the actual computation is done. So they are being used in order to 'learn' the parameters that are really significant for those groups. Unsupervised learning on the other hand is where various data sets are assigned to segments without the groups being known beforehand. Reinforcement learning involves learning various actions with respect to the payoff. Actions that maximize payoff are normally selected. There are various algorithms that can be used to do classification namely artificial neural networks, decision trees, Naïve Bayes classifier and K-nearest neighbor. This study used the Naïve Bayes Classification algorithm to predict the label for a given input sentence. Naïve Bayes classifier is a supervised learning classification method. The Naïve Bayes classifier was built using Python and run together with the natural language toolkit (NLTK) for natural language processing (NLP).

4.5 Naïve Bayes Classifier

Classification using Naïve Bayes is normally based on the Bayes theorem. A simple Bayes classification namely the Naïve classifier is comparable in performance with decision tree and neural network classifiers. Naïve Bayes classifiers have also exhibited high accuracy and speed when applied to large database. Naïve Bayes classifier assumes that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called *class conditional independence*. It is made to simplify the computations involved and, in this sense, is considered "naïve". While applying Naïve Bayes classifier to classify text, each word position in a document is defined as an attribute and the value of that attribute to be the word found in that position.

Naïve Bayes is formalized as the product of the prior probability which is based on previous experience and the likelihood of a given attribute being in a given class, this forms the posterior probability.

To classify an unlabeled example it is just a matter of using the prior probabilities of a given category and multiplying them together. The category which produced the highest probability would be the label/classification for the unlabeled example. Only the words found in the unlabeled example would be looked up in the feature vector. 30

The equation below can be used to classify an unlabeled example. Given a document d and a class c. If the goal is to predict the probability that the document d belongs to class c, the following formula can be used.

P (c/d) = Argmax (P (d/c). P (c))

4.6 Python

Python is the main programming language that was used to build the classifier. The interface was however done using Hypertext Mark-up Language and JavaScript. Python programming language is dynamically-typed and an object oriented interpreted language.

The main advantage of python especially in this case is that it allowed the programmer to easily and rapidly code the prototype. It is also powerful and holds a mature set of standard libraries that makes it easily support large-scale production-level software engineering projects as well. Python has a very shallow learning curve and is an excellent online learning resource.

4.7 Natural Language Toolkit (NLTK)

Python programming language comes with most of the features that are needed to perform simple tasks related to natural language processing. These features are however quite simple and so do not support standard or more advanced natural language processing tasks. There is therefore need to install the Natural Language Toolkit so as to handle such tasks. NLTK is a group of modules and corpora, released under an open source license that allows users to learn and conduct research in NLP.

A corpus is a large collection of texts. It is a body of written or spoken material upon which a linguistic analysis is based. It can either be monolingual, bilingual, open or closed. Monolingual corpora represent only one language while bilingual corpora represent two languages. An *open corpus* is one which does not claim to contain all data from a specific area while a *closed corpus* does claim to contain all or nearly all data from a particular field. Corpora are used in the development of NLP tools. Applications include spell-checking, grammar-checking, speech recognition, text-to-speech and speech-to-text synthesis, automatic abstraction and indexing, information retrieval and machine translation. Corpora also used for creation of new dictionaries and grammars for learners.

The most important advantage of using NLTK is that it is entirely self-contained. Not only does it provide convenient functions and wrappers that can be used as building blocks for common NLP tasks, but it also provides raw and pre-processed versions of standard corpora used in NLP literature and courses.

4.8 Feature Selection, Extraction and Feature Vector

Feature selection involves a series of activities for choosing a specific sub group of terms that occur in a given training set and using this subset that has been obtained as the set of features for text classification. Feature selection is very vital because it simplifies the training and classification hence making it very efficient by reducing the vocabulary size. This directly impacts on the time needed for training. The process also effectively reduces the noise in the data to be input into the classifier hence increasing the accuracy of classification. Existence of a lot of noise in the data greatly reduces the accuracy of classification of new data.

Learning can be performed by extracting clues from the text which may lead to correct classification (Yessenov and Misailovic 2009). Clues about the original data are usually stored in the form of a feature vector (Fn=fi,f2...fn). Each coordinate of a feature vector represents one clue also called a feature Fi of the original text the value of the coordinate may be a binary value indicating the presence or absence of the feature. Proper selection of features strongly influences the subsequent learning. The main goal of selecting good features is to capture the desired properties of the training data in numerical form. This research involves selecting the properties of the training data that would facilitate correct classification. There are algorithms that can be used in feature selection.

The most important concept that is used in concept in implementing a classifier is the feature vector. Feature vector directly determines how successful the text classifier will be. It is used to build a model that the classifier uses to learn from the training data. This model cal also be further used to classify previously unseen data or new data.

The key words that appear in the training data were used as features in this research. These key words are related to the classes and sub classes that the classification will be based. The training data consists of a set of label documents. Each document is split into a set of individual words called unigrams. These are used to define significant words to be added to the feature vector that finally assist in determining the class label of a given tweet or tweeter user. The words that are deemed to not having a say in the class label of a tweet is filtered out.

4.9 Implementing the Classifier using a Web Based Application

A web based environment was used to implement the classifier, especially the interface, so that end users use it to do the actual grouping. This came after successfully building, training and testing the classifier. Python, Html and JavaScript are the main languages that were used to build the web based interface. A web development framework called Django was also used. Structured Query Language (SQL) was used to build the database for holding the tweets and classes and for general database manipulation. For the extraction of tweets, twitter API was used.

The system runs on Ubuntu operating system or Linux environment. End users can use the system to directly extract tweets using a group or hash tag e.g. programming. The system displays the last 100 tweets on the browser. These tweets can then be used to train the classifier which would then be able to classify new tweets in the correct groups. This accomplished the purpose of this study.

CHAPTER 5.0 EVALUATION OF RESULTS AND THE ACHIEVEMENTS

In this section, we discuss the results from evaluation of the system that was done at the Nairobi Institute of Business Studies, Thika Road Campus. Twenty ICT students were sampled and used in the system evaluation.

5.1 Preliminary Study

Preliminary study was done so as to get some background information on the rate of usability of social media especially by current or prospective students. This would also capture the use of social media for educational purposes, other methods probably used and the challenges that have been faced.

This was achieved through the use of interviews and questionnaires. The questionnaire required participants to provide background data on whether they had social media accounts especially twitter, and if they did whether they used them for academic related issues. All the twenty students had social media accounts, and fifteen of them were frequent users who visited their accounts quite often.

The other aspect that was being tested is whether the students had participated in social media groups like chat rooms especially for educational purposes. Three-quarters (15) of the students had been involved in online social groups but only one-third (5) of them had used the groups for academic purposes. The feedback from those who had used social media for academic purposes was quite positive. It was generally easy to connect with the fellows that would participate in learning or that would be consulted. The platforms were also generally comfortable to use and most of the users did not need a lot of training for them to use them. The students who had not used social media for educational purposes stated that this was due to most people preferring to use the media for social activities and networking. The pie chart below summarizes the information extracted from the respondents.

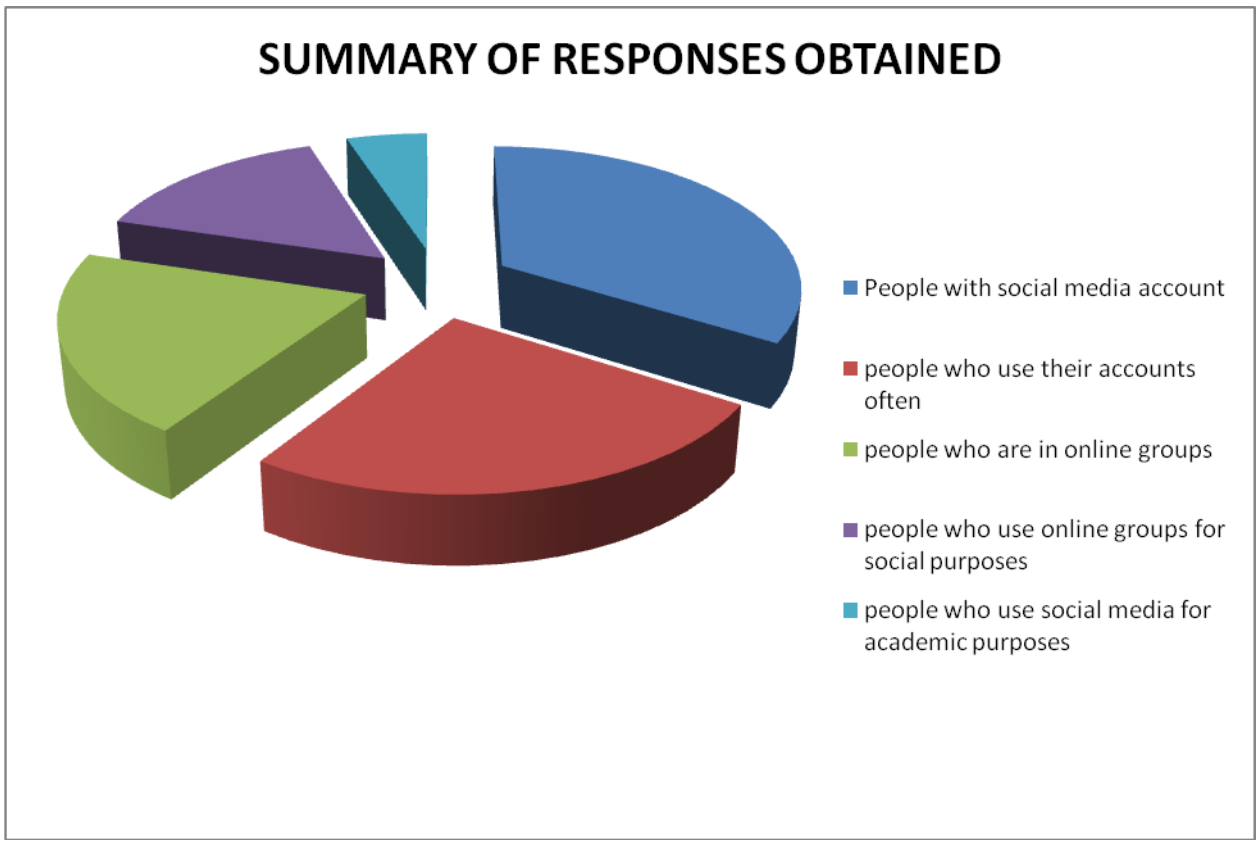


Figure 6.0: Summarized Responses from Questionnaires

5.2 Evaluating the Prototype on Usability

In this section the students were asked to use the prototype for grouping after tweeting on the given hash tag and then rate they system on its ease of use. Most of them responded positively and actually confirmed that the prototype is easy to use. This is because the interface of the system was quite clear and well organized such that they could easily find the links that they needed to accomplish various tasks. The interface was generally similar to some of the social media platforms that they have used before and so it was generally easy to navigate. Most of the learners also commented that the procedure, manner of retrieval, and organization of tweets to be used for learning and classification was generally well organized and so easy to use. They therefore gave the system a clean bill of health in terms of achieving its overall objective and so indicated that they would recommend it for use by both students and teachers.

5.3 Evaluation of the Prototype on its Functionality

This is the section that captured the users view on the functioning of the prototype. A questionnaire was used to assist in determining if the prototype achieved its overall goal which is grouping students through social media for discussion. On this question touching on the overall goal, 90% of the students emphatically agreed that the system actually enabled them to be classified into groups and they were therefore able to know their group members and comfortably interact with them on a given task that they were assigned. They also confirmed that the system simplified the process of group formation and made inclusivity of distant students in the groups possible. Fifteen students which is an equivalent of 85% agreed that the system is able to classify a new user or student in a group. This was an encouragement because it meant that any students who would want to join the groups later would be catered for. Thirteen learners indicated that they would continually use the system for the purposes of group formation and discussion. This is summarized in the chart below.

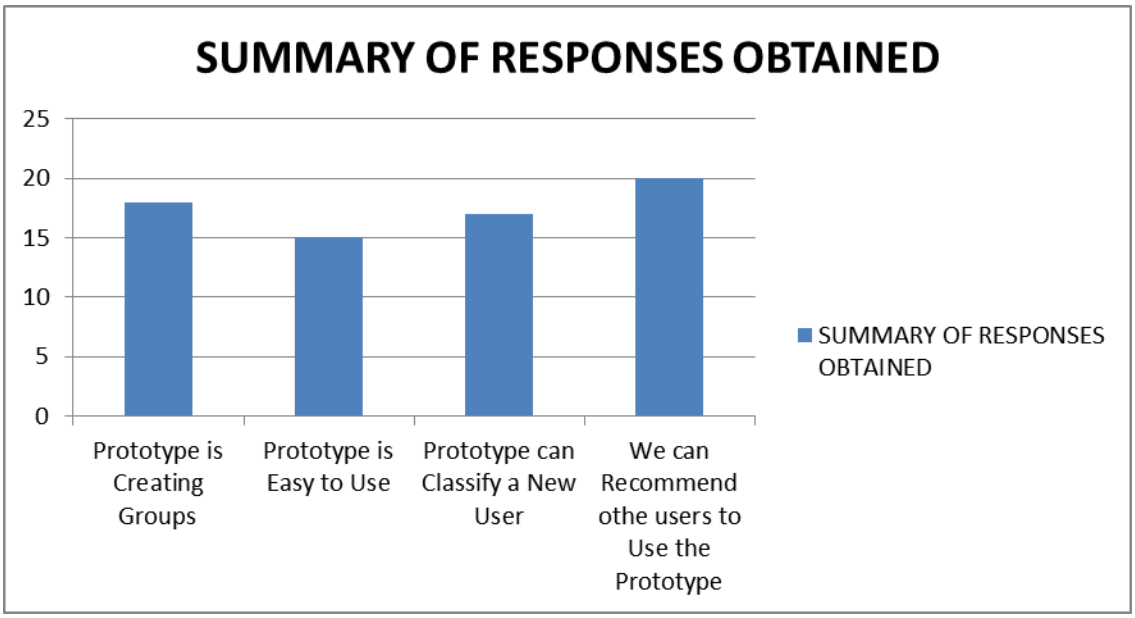


Figure 7.0: Summary of Responses Obtained

5.4 Evaluation of the Naïve Bayes Classifier

The Naïve Bayes Classifier was also tested to evaluate its accuracy, precision and recall. In experimenting with the Naïve Bayes Classifier, we relied on the NLTK module which provides functions for calculating these measures for the classifier. A total of 200 tweets were extracted and used for this test which was summarized in a confusion matrix. This matrix consists of the following parameters: TP, TN, FP and FN, which are defined below.

True Positives (TP): number of positive examples, labeled as such.
False Positives (FP): number of negative examples, labeled as positive.
True Negatives (TN): number of negative examples, labeled as such.
False Negatives (FN): number of positive examples, labeled as negative.

Classifier Accuracy, Precision and Recall

Accuracy: This is the proportion of correct results that a classifier achieved. If, from a data set, a classifier could correctly guess the label of half of the examples, then we say its accuracy was 50%.

Accuracy = (TP + TN)/ (TP + TN + FP + FN)

	Classified positive	Classified negative
Positive class	10	5
Negative class	15	100

Figure 8.0: Results from Sample Classifier Analysis

From these dummy results the accuracy can be calculated as:
Accuracy = (10 + 100)/(10 + 5 + 15 + 100) = **84.6%**

Precision: This measure determines what fraction is correct out of all the examples the classifier labeled as positive. **Precision = TP/(TP + FP)**

Recall – This measure determines what fraction the classifier picked up out of all the positive examples that were there. **Recall = TP/ (TP + FN)**

The results below illustrate a summary of what was obtained when 200 tweets were used to test the Naïve Bayes Classifier.

Feature	**Accuracy**	**POS Precision**	**POS Recall**	**NEG Precision**	**NEG Recall**
Unigrams	***0.714***	***0.502***	***0.950***	***0.937***	***0.426***

Figure 9.0: Results from Actual Classifier Analysis

This classifier was doing the classification using the unigrams. This is where the tweets were being divided into single words which were analyzed before being classified. From this analysis the classifier performed above average with an accuracy of 71.4%. Precision and recall were however average. These measures can be improved if large amounts of data are used to train the classifier before being used to do actual classification.

5.5 Discussion of Objectives' Achievement

In this section, we discuss the findings of the research and how it is related to objectives of the study.

5.5.1 Objective One

The first objective was to study group formation in relation to social media and whether the groups can translate to problem solving groups.

This was achieved through research and extensive reading that is captured in the literature review. It is possible to create groups that can solve problems through social media.

5.5.2 Objective Two

The second objective was to create a way of harnessing data from social media users (students).

This objective was achieved through the use of twitter API and a script. The script was able to access and extract tweets posted by participants on some hash tag in relation to a task that was given. The script was then able to pass these tweets to the classifier as input for learning and classification.

5.5.3 Objective Three

The third objective was to establish relevant attributes of twitter users that can be useful for the clustering purpose.

This was done through desk research. It involved identifying specific keywords for the expected categories. These key words were related to the classes and sub classes that the classification was based. The training data consists of a set of label documents. Each document is split into a set of individual words called unigrams.

5.5.4 Objective Four

The fourth objective was to develop a classifier that is able to use the above data to create clusters (study groups).

This was achieved by developing a Naïve Bayes classifier and training it using the first batch of tweets that was extracted. The tweets extracted from the task hash tag e.g. programming were then fed as input to the classifier. The classifier than gave an output of classified tweets according to the groups that had been established.

5.5.5 Objective Five

The last objective was to develop a set-up that is able to classify a new twitter user (student) into one of the clusters.

This was achieved by creating a way of automatically assigning a new user a group once the system has learnt and done the initial classification. Generally, the platform created was able to achieve the overall goal of the study which was to create a framework for group formation for purposes of group study.

From the discussion of the objectives above in terms of their achievement, it is clear that the prototype was able to achieve the major set goal for the study. Through the study, we were able to address the limitation of the social media of not being properly utilized as a platform for supporting learning activities like group formation. Most of the information that passes through social media was being used majorly for social interaction. The study has proved that it can actually be used constructively in learning in various institutions.

CHAPTER 6.0 CONCLUSION AND RECOMMENDATIONS

6.1 Introduction

This chapter captures the conclusion of this study as well as the areas that need further investigation in machine learning as applied in group formation.

6.2 Conclusion

The main aim of this study was to develop a framework to support group formation using social media specifically twitter. In the beginning of the research, social media was highlighted as the one of the current platforms that has revolutionized communication and general interaction among people including students around the world. It therefore has a great potential in the area of teaching and electronic learning (J. Neuman, 2011).

Through the study, it was underscored that inasmuch as the social media has a great potential in education, this has not been exploited to a greater percentage. The techniques that are currently used in group formation and learning are mostly manual and so not efficient. They therefore come with a lot of challenges including time wastage. Through social media a better and more efficient way can be used to enable online learning generally and group formation specifically.

A prototype was then developed by the researcher so as demonstrate the learning capability of the social media by coming up with a way of creating study groups from the information shared across the social media. The prototype was able to extract tweets from various social media accounts based on a given hash tag (task) and then pass them to a Naïve bayes classifier as input. The classifier then grouped the users into different categories based on various tweets that they posted on the task. The classifier was also able to assign other or new users groups also according to their tweets and the learning that the system had undergone.

The prototype was able to address the limitation of the social media of not being properly utilized as a platform for supporting learning activities like group formation. Most of the information that passes through social media was being used majorly for social interaction. The study proved that it can actually be used constructively in learning in various institutions.

6.3 Challenges

In the course of the research, the researcher met the following challenges:

- Availability of only a small number of people that use twitter compared to other social media.
- Maintenance of the twitter system in the course of the study.

6.4 Suggestions for Further Study

In this study, classification of users was done using the actual tweets that users posted. This is the method that is widely used. There is therefore need to establish how other elements of classification can be used for instance the actual entry details of each user.

This study also stopped at the formation of groups using the tweets of various users. It did not capture anything more that happens in the groups after formation for instance coordination of learning. Further study therefore needs to be done on how the same platform for group formation can also be used to facilitate group learning especially while the students are online.

REFERENCES

Bell, BS & Kozlowski, SWJ .(2002). *A typology of virutal teams: Implications for effective leadership', Group & Organization Management*, vol. 27, no. 1, pp. 14-50.

Benson, A. (2002). *Using online learning to meet workforce demand: A case study of stakeholder influence. Quarterly Review of Distance Education*, 3(4), 443−452

Brook, C. and Oliver, R.(2003). *Online learning communities: Investigating a design framework. Australian Journal of Educational Technology*, 19(2), 139-160.

Clark, R. (2002). *Six principles of effective e-Learning: What works and why. The E-Learning Developer's Journal*, 1−10.

Dongwoo, K., Yohan, J., Il-Chul, M., and Oh, A. (2010). *Analysis of twitter lists as a potential source for discovering latent characteristics of users. Workshop on Microblogging at the ACM Conference on Human Factors in Computer Systems*

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). *Topics in semantic representation. Psychological Review, 114,*211-244. (pdf) (topic modeling toolbox)

Han J, Kamber M (2006). *Data mining: Concepts and Techniques, 2nd edn.* Morgan Kaufman Publishers

Hiltz, S. R., & Turoff, M. (2005). *Education goes digital: The evolution of online learning and the revolution in higher education.* Communications of the ACM, 48(10), 59−64, doi:10.1145/1089107.1089139

Light, V, Nesbitt, E, Light, P & Burns, JR .(2000). *Let's You and Me Have a Little Discussion: computer mediated communication in support of campus-based university courses, Studies in Higher Education*, vol. 25, no. 1

Lu H, Sun S, Lu Y (1996). *Preprocessing data for effective classification.* ACM SIGMOD'96 workshop on research issues on data mining and knowledge discovery, Montreal, QC
Nichols, M. (2003). *A theory of eLearning. Educational Technology & Society,* 6(2), 1−10.

Oblinger, D. G., & Oblinger, J. L. (2005). *Educating the net generation. EDUCAUSE.* Retrieved from. http://net.educause.edu/ir/library/pdf/pub7101.pdf

Pak, A., & Paroubek, P. (2010). *Twitter based system: Using Twitter for disambiguating sentiment ambiguous adjectives.* In Proceedings of the 5th International Workshop on Semantic Evaluation (pp. 436- 439). Association for Computational Linguistics.

Salmon, G. (2000). *E-Moderating: The key to teaching and learning online*, Kogan Page Ltd, London

Tuckman, GW. (2001), *Developmental sequence in small groups, Group Facilitation*, vol. 3, no. Spring, pp. 66-81.

Yardi, S.; Romero, D.; Schoenebeck, G.; and Boyd, D. (2010). *Detecting spam in a twitter network. First Monday* 15:1–4.

Yessenov, K. and Misailovic, S. (2009). *Sentiment Analysis of Movie Review Comments.* Available: http://people.csail.mit.edu/kuat/courses/6.863/report.pdf . Last accessed 13th Jul 2011.

Yuen, A.H. (2003). *Fostering learning communities in classrooms: A survey research of Hong Kong schools.* Education Media International, 40, 153-162.

APPENDIX

APPENDIX 1: Requirements for Prototype Set-up

The requirements for setting up the prototype are as follows:

Django==1.6.5

MySQL-python==1.2.5

twitter==1.14.3

TextBlob ==0.8.4

nltk == 2.0.4

Unidecode==0.04.16

They should be typed in a text editor and saved as twitter_app_requirements.txt on the desktop and installed through the terminal using the command below:

./pip install -r ~/Desktop/twitter_app_requirements.txt

APPENDIX 2: Questionnaire

This is part of the questionnaire that was used in the prototype evaluation.

Please indicate your level of agreement with the following statements regarding the platform you have been using:

1-Strongly Agree, 2 -Agree, 3-No Opinion, 4-Disagree, 5-Strongly Disagree

I have a twitter account					
I use my social media account for learning purposes					
The prototype was easy to use and I managed to easily interact with the interface of the platform.					
I will continue using the application					
I will recommend the application to others					

APPENDIX 3: Sample Source Code

manage.sh file

This file is used to access the server and display the interface of the system.

```
#Run server via py in our venv
#/home/omuya/apps/twitter/venv/bin/python manage.py  syncdb
#/home/omuya/apps/twitter/venv/bin/python  manage.py collectstatic
/home/omuya/apps/twitter/venv/bin/python manage.py  runserver  localhost:3000
```

views.py file

This section of code is used to view the process of data extraction and classification.

```
from django.shortcuts import render
from app import models
from app import appforms as forms
# Create your views here.
#@author Omuya O. Erick
#from django.views.generic import DetailView
from django.views.generic import ListView, CreateView, FormView
from django.shortcuts import render_to_response
from django.template import RequestContext
from django.http import HttpResponse
import twitter
import datetime
import re
from unidecode import unidecode

#WE Need this for NLTK Classification
import os
from nltk.corpus.reader.plaintext import PlaintextCorpusReader
from nltk import NaiveBayesClassifier
from nltk.tokenize import word_tokenize
```

```
import random
class ExtractTweets(ListView):
    model = models.Tweet
    template_name = "tweets/tweets.html"
    context_object_name = 'tweets'

    CONSUMER_KEY = 'gKEHQ9FZn02uisF3a6IrbCcCh'
    CONSUMER_SECRET ='LAvCkiTOQAZtbzrtRz8dgNYlnGumwjlADgjOnl8PelNGYLPVlk'
    ACCESS_TOKEN ='123246186-wEhC5hb2XpoLMLbRdYnNm4qpeMlucmDXpKyBSrhF'
    ACCESS_SECRET = 'iJLINv34JnCFQg5qiaL3gOzKvt4ZvurWBHwHDb3MGZqlV'

auth = twitter.OAuth(
        consumer_key=CONSUMER_KEY,
        consumer_secret=CONSUMER_SECRET,
        token=ACCESS_TOKEN,
        token_secret=ACCESS_SECRET
    )
    def get_context_data(self, **kwargs):
        # Call the base implementation first to get a context
        context = super(ExtractTweets, self).get_context_data(**kwargs)
        context.update(**kwargs)
        context['categories'] = models.TweetCategory.objects.all()
        return context

    def get_queryset(self):
        print self.real, "Counting on real data"
        if self.real:
                return self.model.objects.filter(test=False)
        else:
            return self.model.objects.filter(test=True)
```

```
def get(self, request,  *args, **kwargs):
    self.real = True if request.GET.get('r', None) else False
    return super(ExtractTweets, self).get(request, *args, **kwargs)

def post(self, request, *args, **kwargs):
    self.real = True if request.GET.get('r', None) else False
    #Setting up Twitter API
    api = twitter.Twitter(
        auth=self.auth
    )
    topic = '#programming'
    topic = request.POST.get('tweet-topic', '#programming')

    #search = api.statuses.filter(track = topic)
    try:
        search = api.search.tweets(q=topic, lang='en', count=100, result_type='recent')
    except URLError,e:
        print "Connection Failed, Please check internet", e

        search = {"statuses":[]}

    for s in search["statuses"]:
        raw_text = s["text"].split()
        print "RAW TEXT", raw_text
        #remove emoticond
        text = ""
        for t in raw_text:
            try:
              text += unidecode(t.decode('utf8'))
```

```
                except:
                    pass

            print "New TEXT", text

            created_at = datetime.datetime.strptime(s["created_at"], "%a %b %d %H:%M:%S +0000
%Y")
            created_at = created_at.strftime('%Y-%m-%d %H:%M:%S+0000')
            owner = s["user"]["screen_name"]
            name= s["user"]["name"]
            tweet = models.Tweet()
            tweet.content = text
            tweet.owner = owner
            tweet.tweet_date = created_at
            tweet.test = not self.real
            tweet.save()
            print "Search Result ", (owner, created_at, text)
            print "
        # print "SHOWING DATA SEARCH ", search
        return super(ExtractTweets, self).get(request, *args, **kwargs)

class Tweets(ListView):
    model = models.Tweet
    template_name = "tweets/tweets.html"
    context_object_name = 'tweets'

    def get_context_data(self, **kwargs):
        # Call the base implementation first to get a context
        context = super(Tweets, self).get_context_data(**kwargs)
        context.update(**kwargs)
```

```
        context['categories'] = models.TweetCategory.objects.all()

    def post(self, request, *args, **kwargs):
        self.user_category_form = forms.TweetGroupForm(self.request.POST)
        if self.user_category_form.is_valid():
            user_group = models.TweetGroups()
            user_group.category = self.user_category_form.cleaned_data['category']
            user_group.group_name = self.user_category_form.cleaned_data['group_name']
            user_group.status = True
            user_group.save()
            return super(UserGroups, self).get(request, *args, **kwargs)
        else:
            return render_to_response(self.template_name, { \
                'user_category_form' : self.user_category_form,
                'user_categories' :self.model.objects.all(),
                'categories':models.TweetCategory.objects.all()},
                        context_instance=RequestContext(request))

class Categorize(ListView):
    model = models.UserCategory
    template_name = 'tweets/categorize.html'
    context_object_name = 'user_categories'

    def get_queryset(self):
        print "Looking for real", self.real
        if self.real:
            return self.model.objects.filter(test=False)
        else:
            return self.model.objects.filter(test=True)
def get_context_data(self, **kwargs):
```