ENHANCING CLUSTERING OF USERS IN SOCIAL MEDIA NETWORKS FOR IMPROVED DIGITAL MARKETING

EVELYNE CHANYA SHUMA

MASTER OF SCIENCE

(Computer Systems)

JOMO KENYATTA UNIVERSITY OF

AGRICULTURE AND TECHNOLOGY

2016

Enhancing Clustering of Users in Social Media Networks for Improved Digital Marketing

Evelyne Chanya Shuma

A thesis submitted in partial fulfilment for the degree of Master of Science in Computer Systems in the Jomo Kenya University of Agriculture and Technology

2016

DECLARATION

This thesis is my original work and has not been presented for a degree in any other university.

Signature: …… Date: ……

Evelyne Chanya Shuma

This thesis has been submitted for examination with our approval as the university supervisors

Signature: …… Date: ……

Prof. Waweru Mwangi

JKUAT, Kenya

Signature: …… Date: ……

Dr. Michael Kimwele

JKUAT, Kenya

DEDICATION

This work is dedicated to my dad for his love, support and encouragement and to Nicole my daughter and best friend to her I say “Always reach for the Stars”.

ACKNOWLEDGEMENT

I would like to express my very great appreciation to my supervisors, Prof.Waweru Mwangi and Dr. Michael Kimwele for their valuable and constructive suggestions during the planning and development of this research work. Their willingness to give their time so generously has been very much appreciated. I do thank them for their assistance in keeping my progress on schedule. I would also like to sincerely thank my family, Mark, Nicole, Tina, Loice, Mercy and dad for their never ending support and encouragement along the way.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS/ ACRONYMS

API Application Programming Interface

AB A and B

ASUS Ad space units

BT Behavioral targeting

CPM Cost per mil/cost per thousand impressions

CPC Cost per click

CTR Clickthrough rate

DM Data Mining

E-CRM Electronic Customer Relationship Management

GoK Government of Kenya

ICT Information and Communication Technologies

IT Information technology

IDF Inverse Document Frequency

LISREL Linear Structural Relations

OSN Online Social Networks

ORM Online reputation management

PPC Pay Per Click

PYV Promote your video

PR	Public Relations
ROI	Return on Investments
RFM	Recency, frequency, monetary
SERPs	Search Engine results page
SEM	Search Engine Marketing
SEM	Structured Equation Modelling
SEO	Search Engine Optimization
SN	Social Network
SMM	Social Media Marketing
SPSS	Statistical Package for Social Sciences
TF	Term Frequency
TF-IDF	Term Frequency – Inverse Document Frequency

ABSTRACT

The Internet has changed the world in which we sell. It reaches beyond being a new channel for marketing and offers a new paradigm for the way consumers connect with brands and with each other. Online Social Networks (OSN) which began in the form of generalized online communities that focused on bringing people together to interact with each other have now become an avenue for marketers to look for customers. The rapidly expanding social network audiences in the emerging markets will be huge drivers of social user growth. Changes affecting traditional marketing have been seen where people now spend less time watching TV and reading print newspapers each day and instead communicate through the use of mobile phones, watch videos on YouTube, read the newspaper online, look at photos in Flickr and exchange information through social networks. This transition has forced businesses to find alternative affordable ways to reach customers. The big question for marketers is how to convert users of OSN to customers. The purpose of this research work was Enhancing Clustering of Users in Social Media Networks for Improved Digital Marketing. This is in recognition of the fact that many customers have social media network accounts such as twitter and facebook that can be effectively utilized to convey advertisement information to them. Using these platforms, the digital marketers have access to behavioural and demographic data that can be mined to extract actionable patterns. However, social media data can contain a large portion of noisy data. As such, the definition of noise becomes complicated and relative because it is dependent on the task at hand. Therefore, this study sought to address the challenges of noise removal from social media data and clustering users in social media networks for improved digital marketing. The researcher formulated five objectives which were to: Identify noise in data depending on task at hand for enhanced clustering; Use weight allocation in Term Frequency-Inverse Document Frequency (TF-IDF) environment in order to enhance detection of noise in data; Determine optimal model for reducing noise in data; Make deductions that will lead to design of appropriate framework for social media networks marketing; and to test the clustering framework for social media network adverts. To achieve these objectives,

a quantitative research design was adopted. The data was collected from Facebook social media network open groups, namely Soko Kuu and Soko Nyeusi Official Group and coded into numerical format to allow for mathematical analysis. This study delimited itself to user demographics data that may influence customer buying trends. A target sample of 155 was employed and TF-IDF values were employed as a basis for noise elimination.

The study reveals social media data is indeed noisy and identification of the noise from the data depended on task at hand. Uniquely identifiable Keywords were formulated from the data and tf-idf weighting computed on these keywords as a basis for further eliminating of noise. Keywords that appeared in nearly all documents were given a low weight. Noise identification was achieved by observing the tf-idf values for the different measurable constructs and words with very low tf-idf values compared to their counterparts were regarded as noise and eliminated from further analysis. For the dataset used in this study Nationality, Youth, Group membership and Market were noisy and eliminated from further analysis. It was noted for latent behavioral variables elimination of variables with low tf-idf values as noise is not always obvious as constructs with lower tf-idf values may be indications for some interesting insights. To determine the optimal model for reducing noise in data, variables were adopted or dropped depending on their path costs. Results obtained from the model revealed that all the measurable variables were significantly above the threshold value of 0.05 implying the developed model was the attuned one for the given dataset. Indicators were hypothesized consisting of both behavioral and demographics variables to make deductions leading to design of appropriate framework for social media marketing. In this study demographic variables were used to design appropriate framework for social media marketing. Where tf-idf values for a comparable variable were higher than the other, advertisements were skewed to fit the variable with higher tf-idf value. In this study for automobile goods and services category, tf-idf values for adults were much higher than that of the youth; the consequence is that advertisements should be more skewed to fit adults than youth. Testing the clustering framework for social media marketing revealed among all the

variables buyer attitudes were the most important while religious views were the least important when arranged in their order of importance in the clusters formed. Hypothesis was formulated and were accepted or dropped from the results obtained. Possible combinations between demographics and behavioral constructs were carried out to test the hypothesis. When the test was run hypothesis whose significance levels were above the threshold of 0.05 had they null hypothesis accepted meaning the variables influenced each other. In the case that had null hypothesis rejected, the alternate hypothesis stated that the two constructs never influenced each other. In this study from the dataset used, the deductions made were physical address influenced buyer readiness, buyer attitude and benefit sought. Gender did not influence buyer readiness, buyer attitude or benefit sought. Age influenced buyer readiness and buyer attitude but did not influence benefit sought. Education did not influence buyer readiness, buyer attitude or benefit sought. All these contextual issues informed the study on enhancing clustering of users in social media networks for improved digital marketing. A set of recommendations and guidelines is presented, which could act as a refence point for improving digital marketing in social media networks not only in Kenya, but also in the wider world.

CHAPTER ONE

INTRODUCTION

1.1 Background to the Study

"The recent technology boom has created a digital age. The explosive growth in computer, communications, information, and other digital technologies has had a major impact on the ways companies bring value to their customers" (Kotler & Armstrong, 2012). "Digital technology has also brought a new wave of communication, advertising, and relationship building tools—ranging from online advertising, video-sharing tools, and cell phones to Web apps and online social networks. The digital shift means that marketers can no longer expect consumers to always seek them out. Nor can they always control conversations about their brands" (Kotler & Armstrong, 2012).

"According to Saas Marketers Guide to Analytics (2012) the World of Marketing is changing and there's been a continuous power shift to consumers. For example, technology empowers them to easily find the lowest-cost vendors for goods, and by using e-mail spam filters, they can avoid marketing communications from businesses that they don't want to receive. Today's customers can also influence tens of millions of people to buy from you – or not – by writing online reviews, tweeting and blogging. Another manifestation of this power shift to consumers is the fact that they expect product and service information that's personally relevant, timely and delivered via their preferred channels. Failing to do this will ultimately frustrate customers and turn them away, as your business will be perceived as out of touch – for instance, by blindly pushing products on customers rather than giving them timely access to helpful information at a time when they are open to offers and making purchasing decisions".

"With the rise of social media, the web has become a vibrant and lively Social Media realm in which billions of individuals all around the globe interact, share, post, and conduct numerous daily activities". Social media enables us to be connected and interact

with each other anywhere and anytime – allowing us to observe human behavior in an unprecedented scale with a new lens. This social media lens provides us with golden opportunities to understand individuals at scale and to mine human behavioral patterns otherwise impossible. As a byproduct, by understanding individuals better, we can design better computing systems tailored to individuals' needs that will serve them and society better" (Zafarani, Abbasi, & Liu, 2014). "Recognizing the potentials of this transformation, companies have responded by embracing the newly established patterns of many-to-many communication on platforms such as Facebook, Twitter, etc" (Berthon, 2007), thus establishing an additional marketing channel generally known as social media marketing (SMM)".

"Social media mining is an emerging field where there are more problems than ready solutions. For effective social media mining, we collect information about individuals and entities, measure their interactions, and discover patterns to understand human behavior. Here, we see opportunities as well as problems. The opportunities have been presented in the form of large groups of individuals with complex social relations found on social media who can be targeted with marketing offers. The problem is these individuals are also interested in connecting and interacting with content of interest. They are empowered through technology to filter and block content displayed to them that they deem not suitable at a particular time" (Zafarani, Abbasi, & Liu, 2014) .

Information collected about individuals and entities in social media networks can be analyzed to discover patterns and insights then used to segment them into groups for target marketing. This information also known as big data " big dat is growing exponentially in both size and strategic value to marketers who need to engage in a 1-to-1 manner with these consumers, as this data can be turned into a gold mine of unique customer insight" (Han & Kamber, 2006). "Unfortunately, social media data is significantly different from the traditional data that we are familiar with in data mining. Apart from enormous size, the mainly user-generated data is noisy and unstructured,

with abundant social relations such as friendships and followers-followees" (Zafarani, Abbasi, & Liu, 2014).

"A common complaint: is that "99% Twitter data is useless". – for example "Had eggs, sunny-side-up, this morning"– Can we remove the noise as we usually do in Data Mining? What is left after noise removal? Twitter data can be rendered useless after conventional noise removal. As we are certain there is noise in data, how can we remove it?" (Liu, 2013).

"Clustering is the Key to Big Data Problem as it is not feasible to "label" large collection of objects, no prior knowledge of the number and nature of groups (clusters) in data, clusters may evolve over time and clustering provides efficient browsing, search, recommendation and organization of data" (Jain, Chitta, & Jin, 2012).

In Kenya "Information and Communication Technologies (ICT) have assumed a highly strategic role in the development of the Kenyan economy in the current millennium. Between the years 2000 and 2012, the country's wider transport and communications sector, of which ICT is a part of, grew by a Compounded Annual Growth Rate (CAGR) of 7.7 percent, outperforming all other sectors of the national economy. IDC estimates that ICT spending in Kenya – covering the domains of hardware, packaged software, and IT and telecommunication services – has surged considerably over the past five years, growing from 8.9% of gross domestic product (GDP) in 2006 to an estimated 12.1% of GDP in 2013 (IDC Government Insights, 2014)".

The growth of ICT sector in Kenya will directly spur growth of individuals accessing the internet and social media networking sites. Marketers in Kenya will not be left behind in seeking to market their products to individuals online. Social media networks users generate content which can be mined into data. This data is noisy in that it contains information which can be irrelevant in the situation at hand. On removing the irrelevant information or noise the data can be used to cluster users into appropriate segments for target marketing. Having some prior information on market segments in the social media

networking sites will go along way in ensuring success of marketing campaings in social media networks in Kenya as Marketers can target these segments with relevant advertisements for goods and services.

1.2 Statement of the Problem

"According to The State of Always-On Marketing Study (Taylor & Colwell, 2014) 76% of marketers have failed to use behavioural data in segmentation analysis and targeting execution. Digital marketers have access to behavioural data and there is considerable talk and investment these last few years on big data, however, many businesses are struggling to translate this data with the right technology and skills into better data-led customer facing experiences".

"Mining social media data is the task of mining user-generated content with social relations. This data presents novel challenges encountered in social media mining. New challenge for mining social media data has been identified as noise removal fallacy" (Zafarani, Abbasi, & Liu, 2014).

"Noise removal fallacy is that by its nature, social media data can contain a large portion of noisy data and the definition of noise becomes complicated and relative because it is dependent on our task at hand. In classic data mining literature, a successful data Noise Removal Fallacy mining exercise entails extensive data pre-processing and noise removal as "garbage in and garbage out" (Zafarani, Abbasi, & Liu, 2014).

This study seeks to address the challenge of noise removal fallacy and clustering users in social media networks for marketing. Research will be done to determine how social media data can be mined, pre-processed, noise effectively removed depending on the task at hand, cluster users and get some hidden patterns within these clusters that will enable marketers target each user cluster with relevant advertising messages leading to successful social media marketing campaigns.

Research on how to mine social media data and effectively removing noise will enable people marketing in social media networks identify hidden patterns in the data and structure marketing campaings that will be relevant to individuals on social media networks translating to successfull digital marketing campaigns.

1.3 Justification

"According to a study conducted by Razorfish and Adobe consumers are now more connected and expect more meaningful brand experiences - in real time — than ever before. Companies recognize this and want to respond with real-time experiences and solutions. However, the research found that they struggle with even the most basic technology and marketing programs. Surprisingly, only 13 percent of businesses can target a recognized segment and measure results — indicating that most companies lack the ability to tie together the various elements of their business required to take action on their data and use technology to execute effective targeted experiences. The study also reveals that majority of the marketers lack the ability to tackle behavioral data". (Taylor, & Colwell, 2014).

In Kenya today most businesses have turned to online marketing to promote their businesses and generate sales leads. Majority of the businesses both small and large are now looking to engage customers on social media networks. It is important that these businesses can be able to use appropriate tools to reach and target the right audiences with their products and services. These tools will need to mine data on social media networks, clean the data and analyse it for insights.The insights obtained will be used to conduct successful social media network marketing campaings.

This study is important as improving the clustering process in social media network advertising will assist businesses struggling in their online social media network marketing efforts to effectively utilize technology. It will benefit small and medium size businesses in Kenya who currently do not use any form of analytics in their online social media marketing efforts when determining potential customers to target with marketing

messages. Implementing the improved clustering process is envisioned to improve marketing efforts on social media networks leading to revenue growth and profitability for small and medium size businesses in Kenya.

1.4 General Objective

The study will find out how noise can be identified in social media networks data and effectively removed with the application of tf-idf weighting then cluster users for improved digital marketing.

1.5 Specific Objectives

1. Identify noise in data depending on task at hand for enhanced clustering.
2. Use weight allocation in TF-IDF environment in order to enhance detection of noise in data.
3. Determine optimal model for reducing noise in data.
4. Make deductions that will lead to design of appropriate framework for social media networks marketing.
5. Test the clustering framework for social media network adverts.

1.6 Research questions

1. How can we identify noise in data depending on task at hand for enhanced clustering?
2. How can weight allocation be used in TF-IDF environment in order to enhance detection of noise in data?
3. What approach model is optimal for reducing noise in data?

4. How can deductions be made that will lead to design of appropriate framework for social media networks marketing?

5. How can testing be done on the clustering framework for social media network adverts?

1.7 Scope

The study will focus on improving the clustering process used to segment users based on their behavior and demographics in social media networks through noise removal. Relevant advertising messages will be sent to users based on the clusters they belong. The results based evidence of brand enhancement and competitive advantage gained by firms through behavioral marketing on social networks will be a major area of interest in this study.

1.8 Limitations

1. Privacy settings configured by social media network users on their profiles makes it difficult to mine all the data of interest.

2. It is difficult to validate if profile information put by users on their social media network profiles is a true picture of the real users, for example, users often lie about their age.

3. Due to privacy of user's data in Social Media Networks use of application programming interfaces (APIs) to mine data of interest was a limiting factor hence data mining had to be done manually from user's profiles using tally sheets.

CHAPTER TWO

LITERATURE REVIEW

2.1 The Introduction

'Digital marketing can simply be defined as: Achieving marketing objectives through applying digital technologies' (Chaffrey, Ellis-Chadwick, Mayer & Johnstone, 2013). It involves the promotion of products and services using digital distribution channels that reach consumers in a timely, relevant, personal, and cost-effective manner. At a high level, digital channels can have several categories, such as the internet, mobile, digital outdoors, and any form of interactive digital media. Each category has multiple digital tools/ sub-channels that can support digital marketing'. 'These include: • Internet – Email, banner ads, dedicated websites, pop-up ads, sponsored content, paid keyword search, podcasts, etc. Newer channels comprise social networks, blogs, wikis, widgets, virtual worlds, online gaming and RSS. • Mobile – SMS, MMS, mobile web, mobile applications, and mobile video. • Digital outdoors – Still/ video digital display, interactive kiosks. • Interactive digital medium – interactive television channels' (Jain, 2010). 'If marketing creates demand, digital marketing drives the creation of demand using the power of the Internet. The Internet is an interactive medium. It allows for the exchange of currency, but more than that, it allows for the exchange of value' (Stokes, 2011).

2.1.1 History of digital marketing

'The Internet has changed the world in which we sell. It reaches beyond being a new channel for marketing and offers a new paradigm for the way consumers connect with brands and with each other. The online medium provides consumers with more choice, more influence and more power. Brands have new ways of selling, new products and services to sell as well as new markets in which to sell. The roles played by marketing agencies are shifting too. Traditional agencies are getting better at digital marketing,

while agencies that started out as digital shops are starting to play in the above-theline space. More than ever, integrated strategies that speak to an overall brand identity are vital to achieving an organisation's goals' (Stokes, 2011).

'With computers and the Internet came digital advertising. The history of digital ads goes as far back as 1987, when Apple introduced the Macintosh which featured HyperCard, widely considered the first multimedia tool. Over time, digital ads have evolved from static images to interactive designs with the help of universally used tools as Quicktime, Adobe Photoshop and Illustrator, and Flash. The birth of social media and gadgets like the iPhone and the iPad created new opportunities for multimedia, location-based ads that span across platforms, making digital advertising as rich a field as it's ever been. Check out the infographic figure 1 below to see the evolution of digital advertising and figure 2 continuation of figure 2.1 below on the evolution of digital advertising (Franceschi-Bicchierai, 2012)

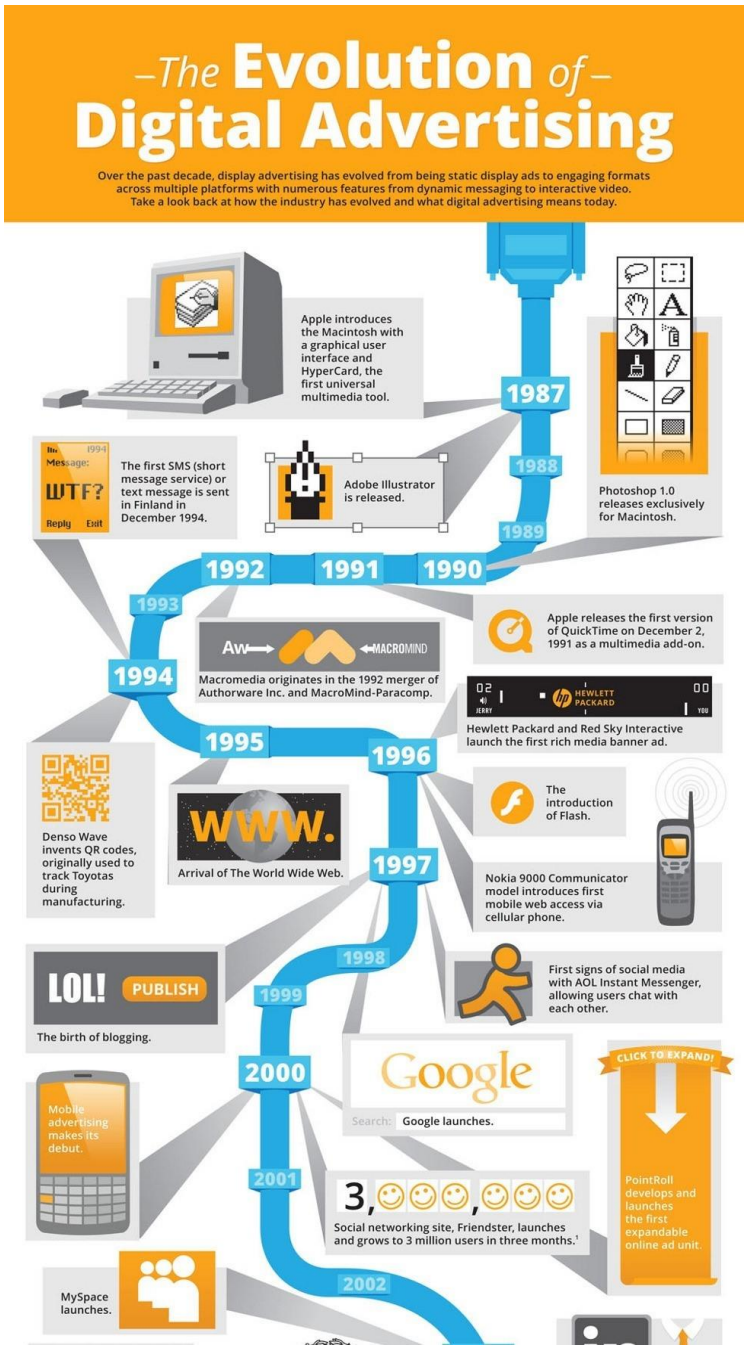


Figure 2.1: Infographic showing evolution of digital advertising (Franceschi-Bicchierai, 2012)

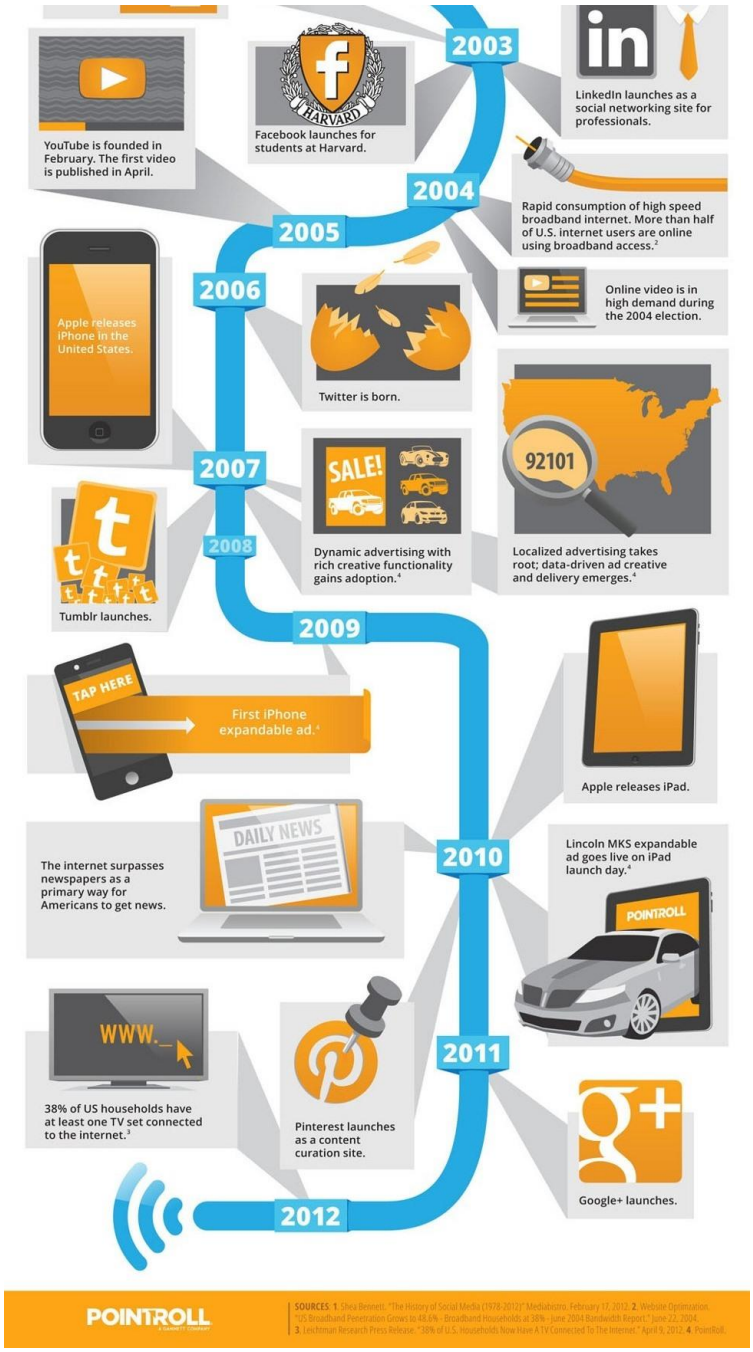


Figure 2.1: Infographic showing evolution of digital advertising (Franceschi-Bicchierai, 2012)

2.1.2 Digital Marketing Tactics

The following digital marketing tactics have been identified in the study.

i. Web development and design

"Web development can be seen as the thread that holds digital marketing together. After all, websites are the first thing we think of when we think of all things "Internet"! Whatever campaign is being run, there is no doubt that it will involve a website. With the crucial role that search engines play in the way that people access the Internet and visit websites, web development and design goes hand in hand with search engine optimization (SEO). And of course, campaigns such as pay per click (PPC), email marketing campaigns and even affiliate campaigns often require custom landing pages. Almost all digital marketing is designed to get users to a website where they convert into customers, so web development really is at the centre of all your online marketing activities" (Stokes, 2011).

ii. Email Marketing

"Email marketing is a form of direct marketing which uses electronic means to deliver commercial messages to an audience. It is one of the oldest yet most powerful of all digital marketing tactics. The power comes from the fact that it is: extremely cost effective due to a low cost per contact, highly targeted, customisable on a mass scale and completely measurable. Email marketing is a tool for building relationships with both existing and potential customers" (Stokes, 2011).

iii. Online advertising

"Online advertising is advertising on the Internet. Online advertising encompasses adverts on search engine results pages (Pay Per Click Advertising), adverts placed in emails and on social networks, and other ways in which advertisers use the Internet". (Stokes, 2011).

The following have been identified as disadvantages of Online Advertising. "Technical obstacles in that the nature of a lot of display advertising is intrusive, so pop up blockers can often prevent adverts from being served as they were intended by the advertisers. Connection speed where bandwidth can also be an issue, although this is a shrinking problem. Advertising fatigue where consumers are reported to be suffering from advertising fatigue, so while new technologies can provide great results, as soon as the market moves mainstream it can get saturated. Consumers are increasingly ignoring adverts. Ad blockers - as well as most browsers now blocking pop-ups, there are also extensions available for the Mozilla Firefox browser, such as AdBlock Plus, that will block advertising on web pages. Technologically savvy consumers are increasingly using these methods to limit the advertising that they see." (Stokes, 2011).

iv. Affliate Marketing

This is also reffered to as 'word of mouth' marketing. "If you recommend a restaurant to a friend, and that friend visits the restaurant because of your recommendation, the restaurant's revenue will have increased because of your referral. Affiliate marketing is used widely to promote websites, and affiliates are rewarded for every visitor, subscriber or customer provided through their efforts. Because of this, affiliates are sometimes viewed as an extended sales force for a website. Affiliates are paid for performance, so affiliate marketing is also referred to as performance marketing" (Stokes, 2011).

v. PPC Advertising

"Pay per click (PPC) advertising is an advertising system where the advertiser only pays for each click on their advert. Hence, pay per click. PPC advertising on search engines is keyword based – this means that it is based on the search term that a user enters into a search engine. A search term can have one word, or be made up of many words. PPC advertising on Facebook is interest and demographics based – this means that it is based on the interests that a user enters onto their profiles, or are listed on pages that contain

content that is related to interests. Advertisers target those interests for which they want their adverts to appear" (Stokes, 2011).

vi. **Social network advertising**

This is an "important growth area for PPC advertising. Adverts are shown to users of social networks. Social networks include Facebook, YouTube (via Google's AdWords) and LinkedIn. It often complements other advertising solutions offered by the social network. Targeting is often behavioural, based on user interests, or demographics, but can be keyword or content based as well (especially for YouTube advertising)" (Stokes, 2011).

"PPC adverts are targeted in a variety of ways, from keyword targeting on search engines, to behavioural and demographic targeting on social networks. PPC advertising can also be targeted based on personal behaviour. With Facebook advertising, adverts can be targeted based on people's profile settings and Facebook behaviour. You could target your advert based on some or all of the following: Gender, Location, Relationship status, Age group. You could also target your behaviour based on: Likes and interests, Brand interactions. Using the Google Display Network and AdWords, you can re-target visitors who came to your site via an AdWords advert based on actions that they took. This means that if someone came to your site, but did not complete a purchase, you can target adverts to them via the Google Display Network" (Stokes, 2011).

"Facebook offers two paid for advertising solutions: Ad Space Units (ASUs) which can be bought on either a cost per thousand impressions (CPM) or CPC cost per click (CPC) basis, or Facebook Engagement Ads which are bought on a CPM basis with a minimum spend threshold. ASUs have no minimum spend, and are the small adverts which appear on Facebook pages that are not the home page. These adverts are served based on interests and demographic information. For example, an advertiser can request to have their advert shown to all women in London who are interested in men, who are themselves single and between 25 and 35, and who like dogs or puppies" (Stokes, 2011).

vii. Google Adwords

John Wanamaker famously said, “I fully believe that half the money I spend on marketing is wasted. The trouble is, I don’t know which half.” And therein lies the fundamental problem of traditional marketing, which we’ve struggled to overcome since before Mr. Wanamaker uttered his famous words. I’ve met countless advertisers who gauge the effectiveness of their online campaigns based on gut feelings. I’m generally a proponent of trusting one’s gut, but not in the case of online advertising. In this world, it’s all about the data’ (Holdren, 2012).

‘AdWords is the name of Google’s auction-based advertising platform. Keywords are the basis of the AdWords auction. The auction is a competition between all advertisers who want ads to appear when a searcher types in the keyword or similar words. AdWords advertisers specify the maximum amount they are willing to pay for an ad click (designated in the account as Max CPC, which stands for maximum cost-per click). With AdWords, winners are not always the highest bidders, because quality is an important factor. The keyword bids and quality combined determine the winners of the auction, as well as the actual cost of individual clicks’ (Holdren, 2012).

‘To build a keyword list that competes well in the auction and results you need to keep in mind that you cannot force people to search in a particular way. If searchers don’t find a relevant result, they try different keyword combinations until they find the results they want. The limitless combinations of search queries makes keyword building a challenge if you want to include every possible way to connect with likely prospects. Keywords are categorized into two groups: keywords that trigger ads and keywords that prevent ads from showing. New advertisers often neglect the latter, called negative keywords. Negatives are a critical component of a successful AdWords account. In some cases, the number of negative keywords exceeds the number of positive keywords in a particular ad group. A thoughtful approach to building a keyword list helps ads showing

higher positions, at lower prices, to the most likely prospects. It also prevents ads from displaying on irrelevant searches' (Holdren, 2012).

viii. Retargeting

'Retargeting is the practice of serving ads based on prior engagement. While there is more than one form of this technology, the most frequently used is site-based retargeting. Other forms include search retargeting, email retargeting, facebook retargeting, and CRM retargeting" (ReTargeter, n.d.). Retargeting lets you selectively advertise to consumers who have already visited your site with a message that's tailored to the type of interest they've shown, keeping your brand top of mind and bringing the right customers back to your site for critical repeat visits that add up to a purchase' (Crosby & Talley, 2014).

2.2 Noise in Data

'Noise in data can be described as any unknown error source. There are various techniques that can be used to identify and remove noise in data. Data pre-processing is one of the techniques used to handle noise in data. Data is pre-processed because it is often incomplete, that is lacking attribute values or certain attributes of interest, noisy, that is containing errors or outlier values that deviate from the expected and often inconsistent, that is containing discrepancies. Data preprocessing when applied before mining can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining' (Han & Kamber, 2006)

There are various forms of data preprocessing as shown in Figure 2.2 below:

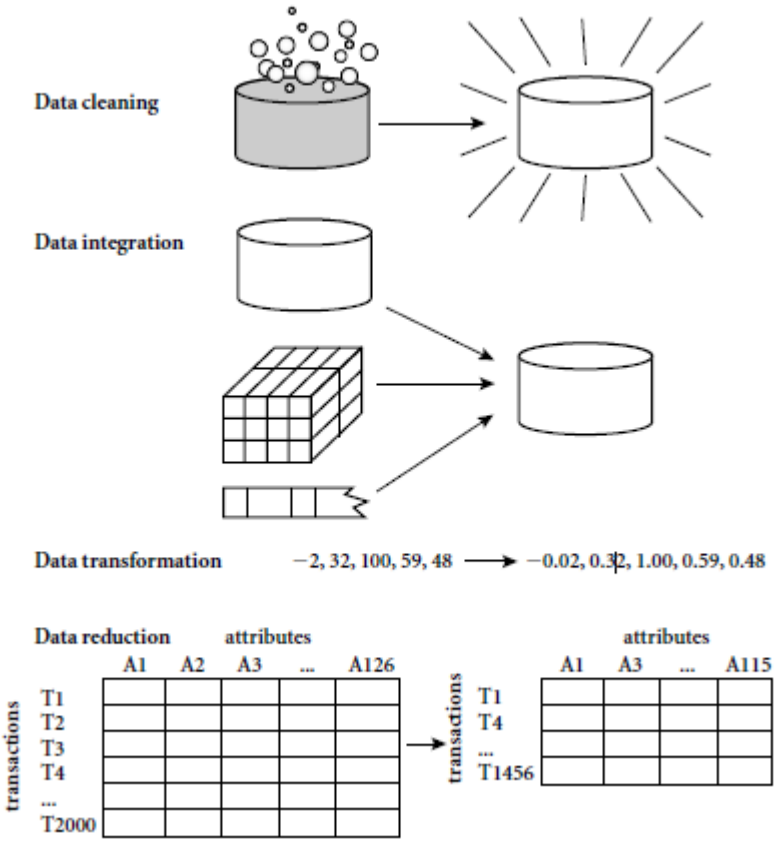


Figure 2.2: Forms of Data preprocessing (Han & Kamber, 2006)

The forms of data pre-processing in figure 2.2 above are explained below:

2.2.1 Data Cleaning

"Data cleaning is applied to remove noise and correct inconsistencies in the data. Real-world data tend to be incomplete, noisy, and inconsistent. Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data" (Han & Kamber, 2006).

The methods below are used to handle missingvalues in data:

i. Ignore the tuple:

   "This is usually done when the class label is missing (assuming the mining task involves classification). This method is not very effective, unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably". (Han & Kamber, 2006).

ii. Fill in the missing value manually

   'This approach is time-consuming and may not be feasible given a large data set with many missing values'. (Han & Kamber, 2006).

iii. Use a global constant to fill in the missing value

   'This is done by replacing all missing attribute values by the same constant'. (Han & Kamber, 2006).

iv. Use the attribute mean to fill in the missing value

   'This is done by using the attribute mean for all samples belonging to the same class as the given tuple' (Han & Kamber, 2006).

2.2.2 Data integration

"Data mining often requires data integration, that is, the merging of data from multiple data stores into a coherent data store, such as a data warehouse" (Han & Kamber, 2006).

2.2.3 Data transformation

"In data transformation, the data are transformed or consolidated into forms appropriate for mining". Data transformation methods can involve the following:

i. Smoothing

"Smoothing works to remove noise from the data. Such techniques include binning, regression, and clustering".

a) "Binning methods smooth a sorted data value by consulting its "neighborhood," that is, the values around it. The sorted values are distributed into a number of "buckets," or bins. Because binning methods consult the neighborhood of values, they perform local smoothing"(Han & Kamber, 2006).

b) "In regression data is smoothed by fitting it to a function". (Han & Kamber, 2006).

c) "Clustering consist of objects that are similar between themselves and dissimilar to objects of other groups. Representing data by fewer clusters necessarily loses certain fine details, but achieves simplification. The quality of a clustering result also depends on both the similarity measure used by the method and its implementation also the hidden patterns. Traditionally clustering are broadly divided into; Partitioning methods, Hierarchical methods and Density based methods" (Jagadeeswaran, 2013). "Unlike classification, clustering and unsupervised learning do not rely on predefined classes and class-labelled training examples" (Han & Kamber 2006).

ii. Aggregation

“Summary or aggregation operations are applied to the data For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts”. (Han & Kamber, 2006).

iii. Generalization of the data

“This is where low-level or “primitive” (raw) data are replaced by higher-level concepts through the use of concept hierarchies. For example, categorical attributes, like street, can be generalized to higher-level concepts, like city or country. Similarly, values for numerical attributes, like age, may be mapped to higher-level concepts, like youth, middle-aged, and senior” (Han & Kamber, 2006).

iv. Normalization

“In this the attribute data are scaled so as to fall within a small specified range, Attribute construction (or feature construction) where new attributes are constructed and added from the given set of attributes to help the mining process. An attribute is normalized by scaling its values so that they fall within a small specified range. Normalization is particularly useful for classification algorithms involving neural networks, or distance measurements such as nearest-neighbor classification and clustering” (Han & Kamber, 2006).

2.2.4 Data reduction

“Complex data analysis and mining on huge amounts of data can take a long time, making such analysis impractical or infeasible. Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results” (Han & Kamber, 2006). The following are strategies for data reduction:

i. “Data cube aggregation, where aggregation operations are applied to the data in the construction of a data cube”. (Han & Kamber, 2006).

ii. “Attribute subset selection, where irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed. Data sets for analysis may contain hundreds of attributes, many of which may be irrelevant to the mining task or redundant. The goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. Mining on a reduced set of attributes has an additional benefit. It reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand. Leaving out relevant attributes or keeping irrelevant attributes may be detrimental, causing confusion for the mining algorithm employed. This can result in discovered patterns of poor quality. In addition, the added volume of irrelevant or redundant attributes can slow down the mining process” (Han & Kamber, 2006).

iii. “Dimensionality reduction, where encoding mechanisms are used to reduce the data set size”. (Han & Kamber, 2006).

iv. “Numerosity reduction is where the data are replaced or estimated by alternative, smaller data representations such as parametric models (which need store only the model parameters instead of the actual data) or nonparametric methods such as clustering, sampling, and the use of histograms” (Han & Kamber, 2006).

v. “Discretization and concept hierarchy generation, where raw data values for attributes are replaced by ranges or higher conceptual levels. Data discretization is a form of numerosity reduction that is very useful for the automatic generation of concept hierarchies “(Han & Kamber, 2006).

2.3 Feature Selection and transformation

"The quality of any data mining method such as classification and clustering is highly dependent on the noisiness of the features that are used for the clustering process. For example, commonly used words such as "the", may not be very useful in improving the clustering quality. Therefore, it is critical to select the features effectively, so that the noisy words in the corpus are removed before the clustering. Feature selection is more common and easy to apply in the problem of text categorization in which supervision is available for the feature selection process. However, a number of simple unsupervised methods can also be used for feature selection in text clustering". Some examples of such methods are discussed below (Aggarwal & Zhai, n.d.).

2.3.1 Document Frequency-based Selection

"The simplest possible method for feature selection in document clustering is that of the use of document frequency to filter out irrelevant features. While the use of inverse document frequencies reduces the importance of such words, this may not alone be sufficient to reduce the noise effects of very frequent words. In other words, words which are too frequent in the corpus can be removed because they are typically common words such as "a", "an", "the", or "of" which are not discriminative from a clustering perspective. Such words are also referred to as stop words. Noisy text collections which are derived from the web, blogs or social networks are more likely to contain such terms. We note that some lines of research define document frequency based selection purely on the basis of very infrequent terms, because these terms contribute the least to the similarity calculations. However, it should be emphasized that very frequent words should also be removed, especially if they are not discriminative between clusters. Note that the TF-IDF weighting method can also naturally filter out very common words in a "soft" way" (Aggarwal & Zhai, n.d.). Below are ways of quantifying the importance of a term directly to the clustering process.

i. **Term Strength**

'A much more aggressive technique for stop-word removal is proposed. The core idea of this approach is to extend techniques which are used in supervised learning to the unsupervised case. The term strength is essentially used to measure how informative a word is for identifying two related documents' (Aggarwal & Zhai, n.d.).

ii. Term Contribution

"The concept of term contribution is based on the fact that the results of text clustering are highly dependent on document similarity. Therefore, the contribution of a term can be viewed as its contribution to document similarit. While this strategy makes these methods unsupervised, there is a concern that the termselection might be biased due to the potential bias of the assumed similarity function. That is, if a different similarity function is assumed, we may end up having different results for term selection. Thus the choice of an appropriate similarity function may be important for these methods" (Aggarwal & Zhai, n.d.).

2.3.2 TF-IDF

"The formal measure of how concentrated into relatively few documents are the occurrences of a given word is called TF-IDF (Term Frequency times In-verse Document Frequency). The measure called TF-IDF lets us identify words in a collection of documents that are useful for determining the topic of each document. A word has high TF-IDF score in a document if it appears in relatively few documents, but appears in this one, and when it appears in a document it tends to appear many times. In several applications of data mining, we shall be faced with the problem of categorizing documents (sequences of words) by their topic. Typically, topics are identified by finding the special words that characterize documents about that topic" (Leskovec, Rajaraman, & Ullman, 2014).

"In order to enable an effective clustering process, the word frequencies need to be normalized in terms of their relative frequency of presence in the document and over the entire collection. In general, a common representation used for text processing is the vector-space based TF-IDF representatio. In the TF-IDF representation, the term frequency for each word is normalized by the inverse document frequency, or IDF. The inverse document frequency normalization reduces the weight of terms which occur more frequently in the collection. This reduces the importance of common terms in the collection, ensuring that the matching of documents be more influenced by that of more discriminative words which have relatively low frequencies in the collection. In addition, a sub-linear transformation function is often applied to the term frequencies in order to avoid the undesirable dominating effect of any single term that might be very frequent in a document" (Aggarwal & Zhai, n.d.).

2.4 Critiques of the Existing Literature Relevant to the Study

The general objective of the study is to enhance clustering of users in social media networks for improved digital marketing by finding out how noise can be identified in social media networks data and effectively removed with the application of tf-idf weighting then cluster users for improved digital marketing.

Existing Literature and been reviewed on the subject of digital marketing and noise as critiqued below :

Evolution of digital marketing has been highlighted in depth together with the various digital marketing techniques in use. "John Wanamaker famously said, "I fully believe that half the money I spend on marketing is wasted. The trouble is, I don't know which half" (Holdren, 2012). This goes to show that most people marketing are not able to determine the right audience to market to, hence the statement half the money spent on marketing is wasted. Various techniques have been developed for online marketing to ensure the right audience is targeted in digital marketing execution. Techniques such as retargeting which is the practice of serving ads based on prior engagement have made

huge strides in ensuring the right target audience is reached in online marketing, it has however, not been discussed how retargeting can be used in social media networks marketing.

PPC advertising on social media networks has been discussed as also being interest and demographics based. It has however, not been discussed if advertisers are given some insights prior to selection of target audience in when using PPC marketing in these medias.

Google adwords, the name of Google's auction-based advertising platform where Ads are displayed at the moment someone is looking for something and presented as potential solutions to their search is another solution for digital marketers. As seen from review of literature, keywords are the basis of the AdWords auction which is a competition between all advertisers who want ads to appear when a searcher types in the keyword or similar words. An advertiser marketing online using Adwords has to identify keywords carefully as he does not have any prior information on which keywords users are most likely to use when conducting their search. In this study, keywords are carefully selected then used in computation of tf-idf weighting values to enhance detection of noise in the data mined from the social media sites.

Marketing on social media networks such as Facebook which target users based on behavior and demographics on their profiles is even more complicated when it comes to identifying customers to target with advertising messages. Social media network sites such as Facebook enable advertisers to target the right audience based on their behavior and demographics as long as the advertiser is able to specify the target audience. This is seen in a case say when an advertiser chooses to target Females aged 30 to 35 years of age within Nairobi City when advertising for a new clothing line. In this case the marketer uses gut feelings to select the target audience that will be interested in the cloths offered for sale as the marketer has no advance information to inform them this target group will be interested in the cloths being sold. This is to say some prior

information on the target audience is still needed when marketing in online social media sites such as Facebook. The results obtained from this study will help marketers address the marketer's gut feeling challenge in target market selection when working with similar datasets as those used in the study. This is possible as the results will give some direction on segments of users who will be most likely interested in certain goods/services leading to digital marketing success.

Stokes (2011) on his analysis on the future of online advertising states that "as technologies evolve the way we interact with content changes, so advertising follows and a little online research will reveal plenty of commentary declaring the decline of display advertising. Increasingly, consumers are becoming both weary and wary of advertising. Clickthrough rates on banners ads are dropping, so the effectiveness of display advertising is being questioned by some". This is a wakeup call on all digital marketers and especially those marketing on social media sites to rethink their online marketing strategies and ensure they are targeting the right audiences with relevant advertisements based on their behavior online at a given time.

A review of the literature has given the description of noise and the various ways it can be identified and removed. Various data preprocessing techniques have been identified such as data cleaning, data integration, data transformation and data reduction. This review gives great insights which are used in the methodology to identify and remove noise in the mined social media data. Techniques such as data transformation and data reduction wil play a key role in this study. Noise removal methods to be adopted in the study will depend on the tasks at hand.

TF-IDF a technique that assigns a high weight to a term that occurs frequently in the document but hardly ever in the whole document collection has been reviewed and will be used in this study to help achieve the objectives of enhancing noise removal and determining appropriate framework for social media marketing.

Structural equation modelling (SEM) will be used in this study in determining the optimum model for reducing noise in data. According to Ayodele and Olushina (2013) in their paper, "structural equation modeling (SEM) permits researchers in the social sciences, management sciences, behavioral sciences, biological sciences, educational sciences and other fields to empirically assess their theories. These theories are generally formulated as theoretical models for observed (manifest) and unobservable (latent) variables. When data are collected for the observed variables of the theoretical model, then the Linear Structural Relations (LISREL) program can be used to fit the model to the data. For example, Adelodun (2008) identified the variables that tend to affect educational performance among adult learners, and developed structural equation models (SEM) for examining the relationships between the variables. He also estimated the parameters of the models and evaluated the formulated models. This was with a view to providing an appropriate framework for predicting educational performance. In their study, Adelodun and Obilade (2011) collected data from a sample of 2000 students and analyzed this data using the factor analysis tool of factor loadings, percentages, F test, test and structural equation model technique. He identified seven factors that affected educational performance. These factors were parental socio-economic characteristics (0.517), circumstances (0.604), self concept (0.647), health characteristics (0.666), marital status (0.730), training environment (0.796) and parenting style (0.817), the statistics in parentheses being factor loadings. The study alluded that structural equation model was capable of predicting educational performance using appropriate indicators". Therefore, structural equation model using LISREL Program will be used as a modelling tool in this study.

2.5 Summary

From review of literature, challenges have been noted of identifying keywords, identifying target audience in social media marketing, decline of display advertising, consumers becoming both weary and wary of advertising and clickthrough rates on banners ads dropping such that the effectiveness of display advertising is being questioned by some ignoring adverts.

2.6 Research Gaps

The research gaps identified in this study are how to utilise the various techniques highlighted in literature review to identify and remove noise in social media data and cluster users to improve digital marketing target execution in online social media networking sites.

CHAPTER THREE

METHODOL0GY

3.1 The Research Design

"Research design refers to the arrangement of the settings for the collection and analysis of the research data. It depicts the conceptual structure within which a research is carried out" (Creswell, 2012). In this study, a quantitative research design was employed. The data that was collected from social media network sites was coded into numerical form to allow for mathematical analysis. This analysis involved the application of the term frequency-inverse document frequency (TF-IDF) algorithm to identify noise in the collected data.

3.2 The Target Population

Facebook social media network site was used to provide the target population. Facebook has features allowing users to create Groups. These Groups provide a space for people to communicate about shared interests and can be created by anyone. When creating a group one can choose 3 privacy settings: Public, Closed and Secret. Table 3.1 below gives some of the privacy settings for the three Facebook groups.

Table 3.1: Some privacy settings for Facebook groups (Facebook, 2016)

	Public	**Closed**	**Secret**
Who can join	Anyone can join or be added or invited by a member	Anyone can ask to join or be added or invited by a member	Anyone, but they have to be added or invited by a member
Who can see who's in the group?	Anyone	Anyone	Only current members
Who can see what members post in the group	Anyone	Only current members	Only current members
Who can find the group in search?	Anyone	Anyone	Current and former members

In this study Soko Kuu and Soko Nyeusi Official Facebook Public groups were used as the target population. Specifically, users who liked commented or bought the goods and services that were being advertized in these two Facebook groups were considered for inclusion in the study.

3.3 Sampling techniques and illustrations

The total data collected for the study consisted of a sample of 155. This sample data was collected from a population of 90,202 Soko Kuu Facebook Group Members and 6836 Soko Nyeusi Official Group Facebook members. Total population from the two groups was therefore, 97,038 at the time of the study. Purposive sampling was used as data for the sample was selected deliberately by the researcher from the profiles of users who liked or commented on goods/services posted for sale in Facebook Public groups of Soko Kuu and Soko Nyeusi Official Group.

3.4 Data Collection Instruments

Data collection is the process of obtaining representative information from the field of study that helps the researcher answer research questions or achieve the set objectives. In this study, the primary data was collected afresh for the first time and therefore, original in character was employed. Tally sheets were used to collect data from online social media network site Facebook.

3.5 Data Collection Procedures

Two Facebook public groups, Soko Kuu and Soko Nyeusi Official Groups were used to provide the required data for this study. Privacy settings for Facebook public groups are such that anyone can; join, see who is in the group, see what members post in the group and find the group in search. These settings helped in data collection as the researcher was able to manually mine data made public from profiles of users in the groups. Users who responded to adverts for items posted had their profiles viewed and demographics and behavioral data collected manually.

The collected data was grouped into two categories: behavioral variables data and demographics data.

3.5.1 Behavioral Variables

"These are constructs that give descriptions of individuals, groups, or organizations and the method they utilize to pick, secure, employ, and dispose of products, services, experiences, or ideas to satisfy desires and the impacts that these methods have on the consumer and society" (Blackwell, Miniard & Engel, 2006).

The buyer behavioral activities were classified as shown in Table 3.2

Table 3. 2: Buyer Behavior Classification

Latent- Behavioral Variables	**Measures**
Buying occasions	Regular Occassions
	Special Occassions
Buyer - Readiness	Like
	Comment
	Buy
Attitude	Positive
	Enthusiastic
	Convinced
Benefit Sought	None
	Economy
	Quality

The following is a description of buyer behavior classification shown in Table 3. 2 above:

Buyers buying occasions were grouped into two: regular occasions and special occasions. Regular occasions' goods and services are those that are bought evenly throughout the year while special occasion goods and services are those whose demand depends on particular periods of the year.

Buyer readiness was indicated by the buyer's actions on the social media site, which could be "like" on the advertisement, "commenting" on the advertisement (for example inquiring about the commodity), or "buying: the product.

Buyer attitudes were classified as positive, enthusiastic or convinced. When a customer liked an advert, he was classified as being positive. Whenever a customer inquired or commented about an advert, he was regarded as being enthusiastic with the commodity. However, when the customer bought the product, he was regarded as being convinced by the commodity.

Buyer benefit sought was classified into none, economy and quality. Buyers who just liked the advert were classified as seeking no benefit (None). Buyers who inquired on the pricing of the commodity were said to be concerned with the economy aspect of the product (Economy). Buyers who inquired on the features of the product (for example by requesting the seller to post more photos of the commodity or available colors) were said to be seeking quality aspect of the product (quality).

3.5.2 Demographics Variables

Demographics cover the entire general public, or groups defined by criteria such as education, nationality, religion and ethnicity. After the identification of the user behavioral variables, demographic data was collected from their respective facebook profiles. The demographic data that was of interest are listed in the Table 3.3 that follows.

Table 3.3: Buyer Demographical Stratification

Latent Demographics	**Measures**
Physical Address	City
	Town
	Market
Gender	Male
	Female
Age	Youth
	Adult
Interests	Academics
	Business
	Socializing
Work	Employed
	Unemployed
Groups	Yes
	No
Religion	Christian
	Muslim
	Other
Politics	Active
	Inactive
Marital Status	Single
	Married
Leisure/Hobbies	Entertainment
	Events
Education Level	Student
	University
	College
Nationality	Kenyan
	Foreign

The following is a description of buyer demographical stratification as shown in Table 3.3 above.

Physical address was grouped into three: City for major towns with city status such as Nairobi, Kisumu and Mombasa; Town for county headquarters such as Machakos and market as any other place which is not a town or a city such as Voi.

Gender was grouped into two; male and female.

Age was grouped into two: Adult for people over 18 years and youth for people under 18 years.

Interests were grouped into three: socializing, academics and business as obtained from user's profile information.

Work was grouped into two: Employed and Unemployed.

Groups were grouped into two: yes for those who are members of groups and no those who not members to any group.

Religion was grouped into three: Christian, Muslim and Other for those who had any other religion on their profile apart from Muslim and Christian.

Politics were grouped into two: Actived for those who had political views on their profiles and Inactive for those who had no political views on their profiles.

Education was grouped into three: College, university and student as obtained from users profile information.

Nationality was grouped into two Kenyan for those were were Kenyan Citizens and Foreign for those who had any other nationality on their profile other than Kenyan.

3.5.3 Classification of goods/services

Customers buying circumstances can be influenced by the various behavioral and demographics variables for the goods and services offered for sale and not all these behavioral variables and demographics affect all the customers in all buying circumstances. There was need therefore, to group customers according to their buying trends based on their motivating factors. Nine (9) goods/services listed in Table 3.4 were

employed for this purpose due to their frequency of adverts for the same on the social media network site Facebook.

Table 3. 4: Goods/Services Stratification

S/NO	**Goods/Service**
1	Automobile
2	Clothing
3	Furniture
4	Flowers
5	Beauty Services
6	Mobile Phones
7	Computers
8	Entertainment Electronics
9	Kitchenware

Table 3. 4 above shows categorization of goods/services that were frequently advertised in Facebook social media network site groups of Soko Kuu and Soko Nyeusi Official Group. For example every advert for selling clothingwear such as dresses and shoes were categorized as clothing. Adverts for utencils and all kitchenware were categorized as Kitchenware. This was done to come up with the nine categories as shown in table 3. 4 above.

3.6 Processing and Analysis

3.6.1 Procedure

a) The nine goods/services that were identified in table 3.4 above were treated as *categories* in this study. The measures for the various demographics and behavioral variables were treated as *key words.*

b) The data related to various *Soko Kuu* and *Soko Nyeusi* Official group members who liked, inquired or bought the advertized goods/services were treated as documents.

c) Keywords were chosen such that each of them contained a unique combination of characters that were distinct from the rest of the keywords. This was important because the developed algorithm used *pattern matching* and *wildcards* to compute the term frequency (TF), the inverse document frequency (IDF) and finally the term frequency–inverse document frequency (TF-IDF). Table 3.5, 3.6 and 3.7 that follows gives examples of the documents used. In total 155 documents were prepared from the mined data.

Table 3.5: Document No 1 illustrating data mined from group members' profile

Tally Form

Facebook Number.....1..... Goods/Service.....Automobile

S/No	Face book Demographic	Details
	Physical Address	City
	Gender	Male
	Age/year of birth	Adult
	Religious views	Christian
	Political views	Involved
	Interests	socializing
	Education	College
	Work	employed
	Group(s)	Member
	Marital Status	Married
	Leisure/hobbies	Entertainment, events
	Nationalities	Kenyan
	Occasions	Regular occasion
	Buyer-readiness	Like, comment
	Attitude	Positive, enthusiastic
	Benefits sought	economy

Table 3.5 above illustrates document number 1. The document is made up of data from a group member's profile that showed interest in goods/services categorized under automobile category. The buyers' profile was viewed and information mined which was captured using tally sheets and input into this table as document number 1 to form document 1 of the total 155 documents. This procedure was repeated to come up with

155 documents under the 9 goods/services categories guided by the advertisements displayed on the Facebook Groups and the response from buyers. In table 3.5 above both behavioral and demographics variables are included.

Table 3.6 and Table 3.7 below give illustrations of documents 43 and 48 under goods/services category Furniture and Clothing respectively.

Table 3.6: Document number 43 illustrating data mined from group members' profile

Tally Form

Facebook Number.....43..... Goods/Service.....Furniture

S/No	Face book Demographic	Details
	Physical Address	City
	Gender	Lady
	Age/year of birth	-
	Religious views	Other
	Political views	Involved
	Interests	socializing
	Education	-
	Work	Jobless
	Group(s)	Member
	Marital Status	-
	Leisure/hobbies	-
	Nationalities	Foreing
	occasions	Regular occasion
	Buyer-readiness	Like
	Attitude	Positive
	Benefits sought	None

Table 3.6 above illustrates data mined from soko kuu and soko nyeusi official group member who showed interest in furniture goods/services category as document number

43 of 155 documents. The fields with a dash indicate this information was not available from the users profile, therefore this field is left as a blank.

Table 3.7: Document number 48 illustrating data mined from group members' profile

Tally Form

Facebook Number.....48..... Goods/Service.....Clothing

S/No	Face book Demographic	Details
	Physical Address	City
	Gender	Lady
	Age/year of birth	youth
	Religious views	Muslim
	Political views	Docile
	Interests	socializing
	Education	High school
	Work	jobless
	Group(s)	Member
	Marital Status	single
	Leisure/hobbies	books
	Nationalities	Kenyan
	occasions	Regular occasion
	Buyer-readiness	like
	Attitude	positive
	Benefits sought	none

Table 3.7 above illustrates data mined from soko kuu and soko nyeusi user official group member who showed interest in clothing goods/services category as document number 48 of 155 documents.

d) The documents in (C) above were fed to a *Mysql* database from which data mining was done via TF-IDF algorithm designed using the *PhP* programming language. Figure 3.1 that follows gives an example of this interface.

e) *PhP* interfaces were designed to cater for each latent variable and its measures. The text boxes were used to fetch *Mysql* data, compute the TF, IDF and TF-IDF and display them as shown in figure 3.2 below:

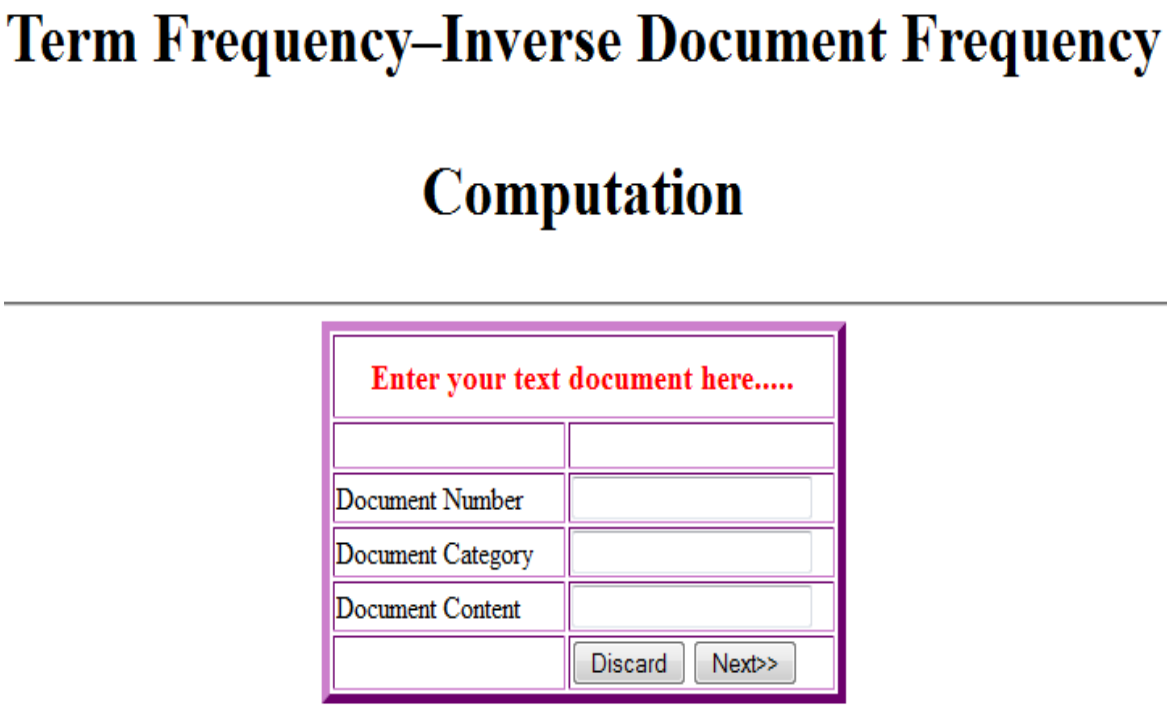


Figure 3.1: Demonstrates Interface used to feed documents into mysql database

Figure 3.1 above demonstrates the interface that was used to feed the documents into mysql database. As highlighted in tables 3.5, 3.6 and 3.7, each document in the the table was fed into the mysql database through the interface on figure 3.1. Table 3.5 for example is document number 1 of the 155 documents. When entering data on table 3.5 into this interface, Document number was entered as 1, Document Category was entered as Automobile, Document Content captured all the data for document number 1 details column in a text format and pasted to the text box as input. This procedure was repeated for all the documents totaling 155.

Figure 3.2 below demonstrate the interfaces created to fetch Mysql data, compute the TF, IDF and TF-IDF values for all the nine goods/services.

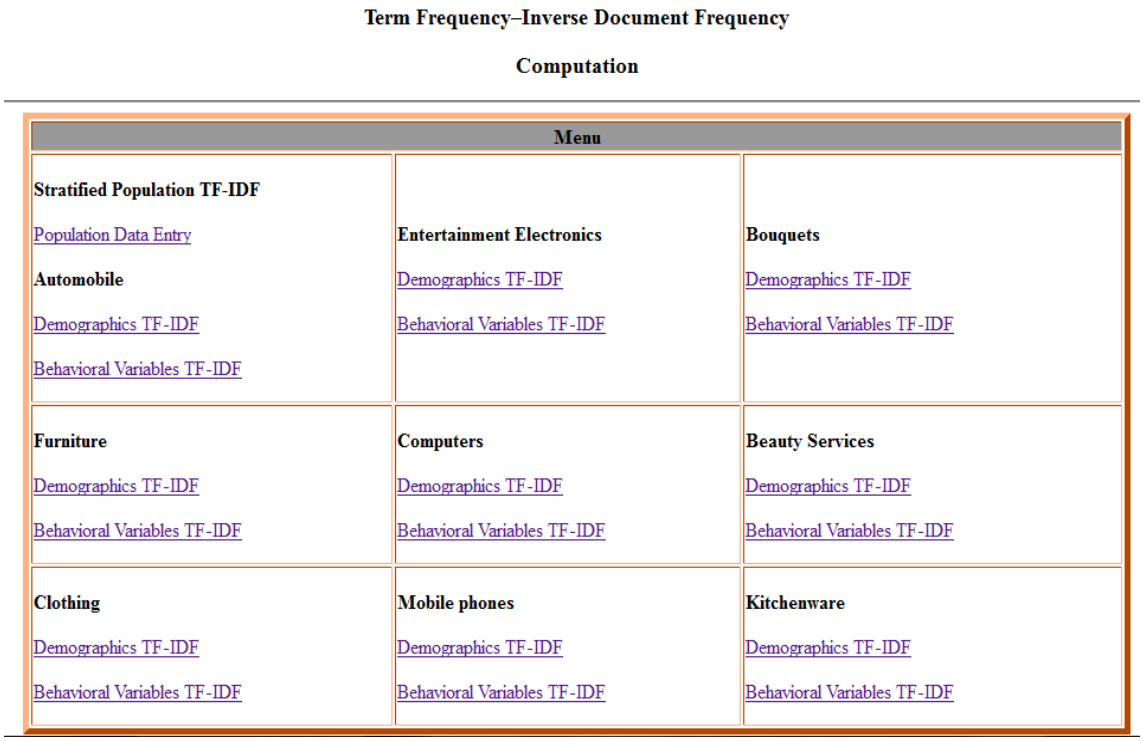


Figure 3.2: Demonstrates Webbased Interface to fetch Mysql data and compute tf-idf values

3.6.2 Noise Identification

The solution to the first objective of identify noise in data depending on the task at hand was keywords were chosen such that each of them contained a unique combination of characters that were distinct from the rest of the keywords used with the developed algorithm which used pattern matching and wildcards to compute the term frequency (TF), the inverse document frequency (IDF) and finally the term frequency–inverse document frequency (TF-IDF). Keywords are represented by the details column of the documents as illustrated in Tables 3.5, 3.6 and 3.7 above.

3.6.3 Weight Allocation

This was the solution to the second objective in the study.

Keywords were assigned a weight that expressed their importance for a particular document. This involved assigning a high weight to a term that occurred frequently in the document category but rarely in the whole document collection. Contrarily, a term that occurs in nearly all documents has hardly any discriminative power and is given a low weight, which is usually true for stopwords like determiners, but also for collection-specific terms. This meant that keywords that appeared in nearly all documents were given a low weight.

This research utilized the term frequency-inverse document frequency (TF-IDF), which is a measure of word importance in document (William, 2010). The term frequency (TF) is basically the number of times a given term appears in a document of interest. The inverse document frequency (IDF) is the logarithm of the number of all documents divided by the number of documents containing the term of interest, which is the general importance of the term in corpus.

To assign weights to document data, term frequency-inverse document frequency (TF-IDF) was employed. The relationship that was used to calculate TF-IDF for specific

terms in the documents is given by equation 1.

IDF=Log 10 [N/df]…….(1)

Where:

IDF=Inverse document frequency

N= the total number of documents

df= document frequency

To find the TF-IDF of a given word in a document, equation (1) above is multiplied by the term frequency (TF) of that word as illustrated in equation 2.

TF-IDF=TF*IDF……..…….(2)

To compute TF-IDF values of a good/ service, the good/service category is selected. Once a selection is made on the hyperlink as illustrated in figure 3.2 above, the algorithm computes and displays the TF and IDF as shown in figure 3.3 below. To compute the TF-IDF a user selects evaluate as shown in Figure.3.3 and results displayed in figure 3. 4.

Term Frequency–Inverse Document Frequency

Computation [Document: - Automobile]

Latent-Demographics	Measures	TF	IDF	Latent-Demographics	Measures	TF	IDF
Physical Address	City	16	0.98621171551437	*Interests*	Academics	3	1.7132104434506
	Town	13	1.0763883458635		Business	3	1.7132104434506
	Market	1	2.1903316981703		Socializing	18	0.93505919306699
Gender	Male	23	0.8286038621527	*Leisure/Hobbies*	Entertainment	28	0.74317366682807
	Female	7	1.345233658156		Events	7	1.345233658156
Age	Youth	1	2.1903316981703	*Education Level*	Student	1	2.1903316981703
	Adult	26	0.77535835019947		High School	2	1.8893017025063
Religion	Christian	24	0.81012045645869		College	11	1.1489390130121
	Muslim	3	1.7132104434506		University	10	1.1903316981703
	Other	3	1.7132104434506	*Work*	Employed	18	0.93505919306699
Politics	Active	20	0.88930170250631		Unemployed	10	1.1903316981703
	Inactive	10	1.1903316981703	*Groups*	Member	28	0.74317366682807
Marital Status	Single	11	1.1489390130121		No Membership	2	0.81012045645869
	Married	14	1.0442036624921	*Nationality*	Kenyan	29	0.72793370027134
					Foreign	1	2.1903316981703

Evaluate

Figure 3.3: Demonstrates computations for TF and IDF values for Latent Demographics Automobile goods/services

Figure 3.4 below shows computations for TF and TF-IDF values for Latent Demographics Automobile when evaluate on figure 3.3 above is selected.

Term Frequency–Inverse Document Frequency

Computation [Document: - Automobile]

Latent-Demographics	Measures	TF	TF-IDF	Latent-Demographics	Measures	TF	TF-IDF
Physical Address	City	16	15.779387448:	*Interests*	Academics	3	5.1396313303518
	Town	13	13.993048496:		Business	3	5.1396313303518
	Market	1	2.1903316981		Socializing	18	16.831065475206
Gender	Male	23	19.057888829	*Leisure/Hobbies*	Entertainment	28	20.808862671186
	Female	7	9.4166356070		Events	7	9.416635607092
Age	Youth	1	2.1903316981	*Education Level*	Student	1	2.1903316981703
	Adult	26	20.159317105		High School	2	3.7786034050126
Religion	Christian	24	19.442890956		College	11	12.638329143133
	Muslim	3	5.1396313303		University	10	11.903316981703
	Other	3	5.1396313303	*Work*	Employed	18	16.831065475206
Politics	Active	20	17.786034050		Unemployed	10	11.903316981703
	Inactive	10	11.903316981	*Groups*	Member	28	20.808862671186
Marital Status	Single	11	12.638329143		No Membership	2	1.6202409129174
	Married	14	14.618851274	*Nationality*	Kenyan	29	21.110077307869
					Foreign	1	2.1903316981703

Figure 3.4: Demonstrates computations for TF andTF- IDF values for Latent Demographics Automobile

As an illustration on how IDF and TF-IDF computations were done, let us consider the word City that appears 16 times in a document category called ‘Automobile latent demographics’ in figure 3.4 above. Given that the total number of documents is 155 the TF-IDF can be computed for this word. With this information, our first task is to calculate IDF. This is done as follows.

IDF=Log 10[N/df]

Where N=155

df=16

IDF=Log 10[155/16]

= Log 10[9.6875]

=0.9862

TF-IDF computation is the second task after computing IDF values and will be computed as follows:

TF-IDF=TF*IDF

Where TF=16

IDF=0.9862

TF-IDF=16*0.9862

TF-IDF=15.779387

The formular for calculating TF-IDF above was employed in a PhP code to achieve this. The PhP snippet that was used is given in Figure 3.5 below:

```
$result1 = mysql_query("SELECT count(Text1) as TWtotal1 FROM tfidftable where Text1 like '%City%' and DocName='Beauty services'");

while($row = mysql_fetch_array($result1))

{

$rw1=$row['Twtotal1'];

$rw2= $rw/$rw1;

$rw3=log10($rw2);

echo $rw3;

}

……..

$tfi=$tf*$df;

echo $tfi;

……..
```

Figure 3. 5: Demonstrates Snippet for TF-IDF Computations for goods/services beauty services

As shown in Figure 3.5 above, the snippet first establishes a connection to the database via the structured query language (SQL). This SQL fetches the term frequency (TF) as well as the document frequency (DF), the key components that are used in the TF-IDF computations.

Figure 3.5 above also confirms that the calculated TF-IDF values are products of the TF and IDF values as evident from the lines that read:

$tfi=$tf$df;*

echo $tfi;

The above procedure was followed to compute all the TF-IDF values for the nine goods/services for both behavioral and demographics variables.

3.6.4 Model for noise reduction

This was the solution to the third objective in this study.

The definitive objectives in modeling are shaped by the value accredited to the system, its anticipated problems or changes, and what can be done with the available resources, data, and technology. This study aimed to develop a noise reduction model for social media networks improved digital marketing. Lisrel software was used for this respect. "A Lisrel model is concerned with the approximation of the observed correlation (covariance) matrix. Fit statistics all measure the distance between the observed and expected matrix" (Barbara, 2009).

The requirements for this model were obtained from the social media networks demographics and behavioral data. Depending on the path costs, some variables are bound to be retained as the rest are dropped. A final model consisted of only those constructs that loaded higher than the threshold values obtained by a review of the literature on similar studies.

According to Nyakomitta and Omollo (2014) "the cut-off path costs between the latent and measurable constructs is 0.05".

Lisrel software employed the underlying data in the Statistical Package for Social Sciences (SPSS). This meant that the various TF-IDF values had to be coded

appropriately before being fed into SPSS. To accomplish this, the highest truncated values and the lowest truncated values for all words were considered. This applied for both demographics and behavioral variable data.

The first task was the identification of the latent and measurable variables. Whilst measurable variables can be determined directly, latent variables cannot. Therefore, latent variables were established as Behavioral Variables. The measurable variables were physical address, gender, age, religion, politics, marital status, interest, leisure/hobbies, education level, work, groups and nationality.

Once populated in Lisrel Software the variables are defined and a new path diagram drawn. The path diagrams designed, showed the connections between the measurable variables and the latent variables.

These path diagrams for the demographics and behavioral variables were built using the procedure above and then run. The outputs of these models formed the basis for the adoption or elimination of the measurable variables.

3.6.4.1 Noise reduction model development illustration

The following steps were followed to develop the noise reduction model:

Step 1: Identify target market population to mine data from Social media network

Step 2: Mine the data and record in tally sheets

Step 3: Categorise the data into behavioral and demographics variables

Step 4: Classify the goods and services offered for sale into categories

Step 5: Treat the mined data as documents

Step 6: Identify keywords with unique combinations from the documents

Step 7: Data preprocess keywords to obtain unique combinations

Step 8: Compute tf-idf weighting on keywords

Step 9: Consider the highest truncated and lowest truncated tf-idf values obtained in step 8 above

Step 10: Code the tf-idf values in step 9 above and feed them to SPSS package and save the data with a .sav file extension

Step 11: Populate the .sav SPSS data in step 10 above in to Lisrel Software

Step 12: Define latent and measurable variables

Step 13: Build path diagrams in LISREL software using the populated data in step 11 showing the connection between the measurable variables and the latent variables

Step 14: Obtain outputs of the models forming the basis for adoption or elimination of the measurable variables depending on their path costs. The cut-off path costs used between the latent and measurable constructs is 0.05.

3.6.5 Framework for social media networks marketing

This was the solution to the fourth objective in this study. The goal was derivation of deductions that would aid in the design of appropriate framework for social media networks digital marketing. To achieve this objective, two latent variables were utilized, behavioural and demographics. The basis for the elimination of measurable constructs depending on their weight allocation was key for this design.

3.6.6 Clustering Framework Testing

This was the solution to the fifth objective in the study. The first task in cluster testing was development of the clusters from the underlying SPSS data. SPSS feature of ‘Direct

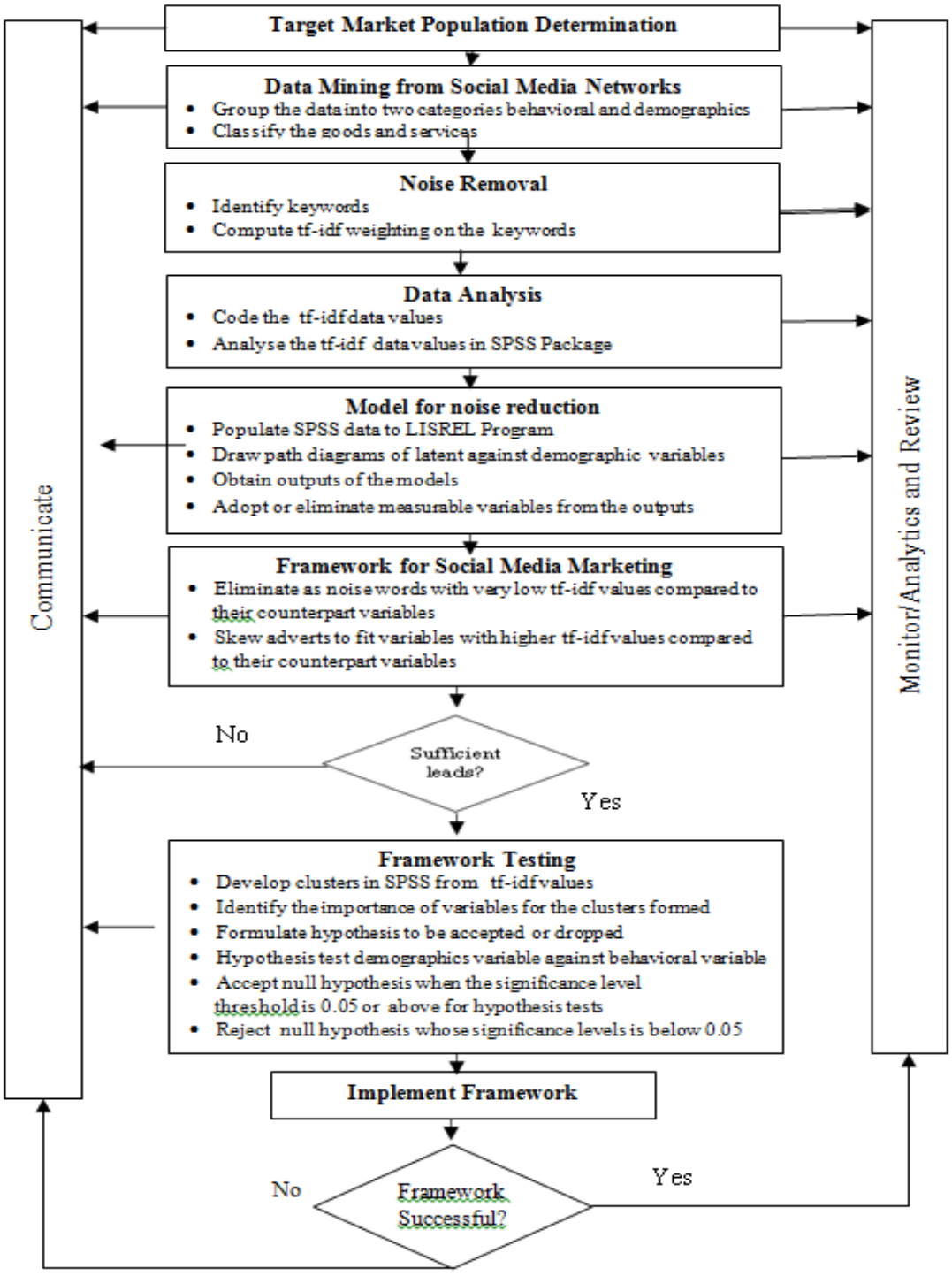
Marketing' was utilized to cluster online respondents into various groups. This SPSS feature provides a set of tools designed to improve the results of direct marketing campaigns by identifying demographic, purchasing, and other characteristics that define various groups of consumers and targeting specific groups to maximize positive response rates.

SPSS Direct Marketing provides a number of pre-packaged procedures for common direct marketing applications such as RFM analysis, segmentation, simple rule induction, segmentation, basic A/B testing and postcode selection.

IBM SPSS Statistics 20 package Direct marketing feature was used. Since this research was focused on clustering of social media network users, the cluster analysis (segmentation) option was used. Cluster analysis is an exploratory tool designed to reveal natural groupings (or clusters) within data. For example, it can identify different groups of customers based on various demographic and purchasing characteristics.

Table 3.8 below gives a visual representation of the social media marketing framework developed.

Table 3.8: Visual representation of the Social Media Marketing Framenwork



3.6.7 Ethical issues

The data collection process was guided by ethical issues and considerations of confidentiality, respect, anonymity, competence, responsibility, consent, understanding and security. The outcomes of this research are therefore to be used for academic purposes only and are specific to the dataset used.

CHAPTER FOUR

RESEARCH RESULTS AND DISCUSSION

4.1 Results

This research study was based on enhancing clustering of users in social media networks for improved digital marketing. This chapter presents the results obtained from the field of study. It also provides possible explanations for the observed phenomena and analysis of the observations.

These discussions are guided by the results obtained from the five objectives as given objective by objective below:

Results obtained from the first objective of identify noise in data depending on the task at hand were keywords that contained unique combination of characters that were distint from the rest of the keywords. Keywords that did not contain unique combination of characters underwent data-preprocessing so as to improve the overall quality of patterns mined using the the tf-idf algorithm prepared.

As an illustration for this first objective consider the keywords: employed and unemployed. We note that ‘Unemployed’ contain the pattern ‘employed’ in it. Therefore pattern matching and wildcards in the TF-IDF algorithm will give wrong values for these two keywords. This is because the TF for the pattern ‘employed’ will include the TF in the keyword ‘Unemployed’.To address this challenge, a synonym for the word ‘Unemployed’ was utilized. This was the keyword ‘jobless’. This was done for the all the other keywords as illustrated in Table 4.1 below.

Table 4.1: Synonym adopted for key words

Key Words	**Synonym adopted**
Employed and Unemployed	Jobless keyword adopted to represent unemployed
Male and Female	Lady keyword adopted to represent Female
Active and Inactive	Docile keyword adopted to represent inactive

Results from table 4.1 above ensured all the keywords used in the documents were uniquely identified and did not contain patterns of words in other keywords as this was essential in ensuring accurate computations for tf-idf values.

The second objective in the study was use weight allocation in TF-IDF environment in order to enhance detection of noise in data. TF-IDF algorithm has been successfully employed for stop-words filtering in various subject fields including text summarization and classification. Weight allocation was employed in this study to determine the importance of particular social media network advertisement words in the entire documents. The keywords (measures in this case) were assigned a weight that expressed their importance for a particular document. This involved assigning a high weight to a term that occurred frequently in the document category but rarely in the whole document collection. This meant that keywords that appeared in nearly all documents were given a low weight.

TF-IDF weighting was utilized on the data to yield the TF-IDF weights for the nine document categories for both behavioural and demographics variables.

Figure 4.1 below shows the screenshots for results obtained when TF-IDF algorithm was utilized for the document category - Automobile, latent variable-Demographics.

Term Frequency–Inverse Document Frequency

Computation [Document: - Automobile]

Latent-Demographics	Measures	TF	TF-IDF	Latent-Demographics	Measures	TF	TF-IDF
Physical Address	City	16	15.779387448	*Interests*	Academics	3	5.1396313303518
	Town	13	13.993048496		Business	3	5.1396313303518
	Market	1	2.1903316981		Socializing	18	16.831065475206
Gender	Male	23	19.057888829	*Leisure/Hobbies*	Entertainment	28	20.808862671186
	Female	7	9.4166356070		Events	7	9.416635607092
Age	Youth	1	2.1903316981	*Education Level*	Student	1	2.1903316981703
	Adult	26	20.159317105		High School	2	3.7786034050126
Religion	Christian	24	19.442890955		College	11	12.638329143133
	Muslim	3	5.1396313303		University	10	11.903316981703
	Other	3	5.1396313303	*Work*	Employed	18	16.831065475206
Politics	Active	20	17.786034050		Unemployed	10	11.903316981703
	Inactive	10	11.903316981	*Groups*	Member	28	20.808862671186
Marital Status	Single	11	12.638329143		No Membership	2	1.6202409129174
	Married	14	14.618851274	*Nationality*	Kenyan	29	21.110077307869
					Foreign	1	2.1903316981703

Figure 4.1: Demonstrates TF-IDF Computations for Latent-Demographics Automobile

The calculation of TF-IDF values for the rest of the words followed a similar format for all the nine goods/services for both behavioral and demographics variables.

Table 4.2 below shows tabulated results of computations of TF-IDF demographics values for the nine goods/services

Table 4.2: Showing TF-IDF demographical values for the nine goods/services

		AT	EE	BQ	FT	COM	BS	CLO	MOB	KIT
Physical Address	City	15.77	13.33	12.63	8.47	9.41	9.41	8.47	13.99	10.29
	Town	13.99	3.77	9.41	10.29	9.41	5.13	3.77	10.29	2.19
	Market	2.19	2.19	3.77	2.19	8.47	0	3.77	6.35	2.19
Gender	Male	19.05	12.63	10.29	10.29	13.99	5.13	2.19	13.33	6.35
	Female	9.41	6.35	13.33	9.41	9.41	9.41	11.12	13.99	8.47
Age	Youth	2.19	6.35	10.29	8.47	9.41	8.47	8.47	6.35	2.19
	Adult	20.1	11.12	12.63	10.29	13.99	6.35	6.35	17.31	9.41
Religion	Christian	19.44	11.90	15.21	11.12	15.77	11.12	7.45	18.65	11.12
	Muslim	5.13	3.77	2.19	3.77	2.19	0	5.13	2.19	2.19
	Other	5.13	3.77	6.35	5.13	5.13	2.19	3.77	3.77	0
Politics	Active	17.78	9.41	13.99	12.63	13.99	5.13	8.47	13.33	5.13
	Inactive	11.90	10.29	8.47	6.35	9.41	9.41	6.35	13.33	9.41
Marital Status	Single	12.63	10.29	15.21	7.45	13.99	9.41	8.47	15.21	5.13
	Married	14.61	7.45	6.35	9.41	8.47	5.13	6.35	9.41	9.41
Interests	Academics	5.13	5.13	6.35	3.77	2.19	3.77	2.19	6.35	2.19
	Business	5.13	7.45	3.77	5.13	0	2.19	3.77	5.13	6.35
	Socializing	16.83	9.41	14.61	11.90	16.83	9.41	9.41	15.21	7.45
Leisure/Hobbies	Entertainment	20.80	13.99	15.77	11.90	15.21	11.90	10.29	17.31	10.29
	Events	1.62	2.19	5.13	6.35	7.45	0	2.19	8.47	3.77
Educational Level	Student	2.19	2.19	5.13	0	13.99	9.41	3.77	6.35	0
	High School	3.77	0	2.19	10.29	2.19	0	3.77	8.47	0
	College	12.63	11.12	10.29	5.13	2.19	2.19	6.35	6.35	5.13
	University	11.90	5.13	9.41	3.77	5.13	3.77	3.77	9.41	5.13
Work	Employed	16.83	9.41	11.90	7.45	9.41	5.13	5.13	15.21	10.29
	Unemployed	11.90	9.41	11.90	11.90	13.33	9.41	9.41	9.41	3.77
Groups	Member	20.80	15.21	17.78	14.61	17.78	11.90	10.29	19.80	11.90
	No Membership	1.62	0	0	1.04	0	0	1.19	0	0
Nationality	Kenyan	21.11	15.21	17.78	14.61	15.77	11.90	11.12	19.44	11.90
	Foreign	2.19	0	0	0	5.13	0	2.19	2.19	0

Key to Table 4.2 AT – Automobile, EE – Entertainment Electronics, BQ – Bouquets, FT – Furniture, COM – Computers, BS – Beauty Services, CLO – Clothing, MOB – Mobile Phones, KIT – Kitchenware

The ultimate goal for noise identification was achieved by observing the TF-IDF values for the different measurable constructs. Words with very low tf-idf values compared to their counterparts were regarded as noise and eliminated from further analysis.

Inter-range difference of more than 7 values was considered to be a considerable difference (very low) to have the values ignored and eliminated as noise, while inter-range difference of 0 to 6 was considered fairly small and had their values accepted.

To illustrate how noise removal was accomplished, we consider measurable variables, *physical address, age*, *nationality, work* and *group membership* for document category automobile latent demographics. The choice of these measures is based on the relatively large range between the highest and lowest TF-IDF values.

It was observed that the inter-range difference between *'employed'* and *'Unemployed'* was fairly small (16.83-11.90, truncated to 5). The conclusion is then that both employed and unemployed were significant in the provision of the research information. On the *'Groups'* variable, the inter-range difference was fairly large (20.80-1.62, truncated to 19). Similarly, for the case of *'Nationality'*, the inter-range difference was fairly large (21.11-2.19, truncated to 19). In case of latent construct *physical address*, *city* had a TF-IDF value of 15.7, town had 13.9 and *market* had a TF-IDF of 2.1. Therefore, market becomes *noisy* in this scenario. Moreover, if we consider latent construct *age*, *adult* had a TF-IDF value of 20.1 while *youth* had 2.1. Once again, youth become *noisy* in this circumstance.

Similar deductions can be made for other constructs as well.

Figure 4.2 below illustrates data on computations for tf-idf values for computers behavioral variables.

Term Frequency–Inverse Document Frequency

Computation [Document: - Computers]

BEHAVIORAL VARIABLES	Measurable Variables	TF	TF - IDF
Occassions	Regular Occassions	18	16.831065475206
	Special Occassions	1	2.1903316981703
Buyer - Readiness	Like	12	13.333805425472
	Comment	7	9.416635607092
	Buy	3	5.1396313303518
Attitude	Positive	12	13.333805425472
	Enthusiastic	7	9.416635607092
	Convinced	3	5.1396313303518
Benefit Sought	None	10	11.903316981703
	Economy	6	8.4730826867196
	Quality	1	2.1903316981703

Figure 4.2: Demonstrates Data on Latent Behavioral Variables for Computers

Figure 4.2 above demonstrates that the TF-IDF value for *'Like'* was 13.33 while that of 'buy' was 5.13 for computers behavioral variables. Considering the fact that higher TF-IDF indicates the importance of a certain word in a document, one may be tempted to phase out variables with lower values. However, for the case of behavioral variables, this elimination is not always obvious. Constructs with lower TF-IDF values may be indications for some interesting insights. For example, lower values for *'buy'* results from the fact that very few respondents buy the advertised goods and services. This may be attributed to the fact that online fraud is rampant and therefore many people do not trust online products and services.

Therefore those who buy should not be eliminated but instead the factors that motivated them to buy should be investigated so that digital marketers can structure their adverts in such a way as to encourage more people to buy. This is in contrast to the elimination on

demographics. The deductions are that rightfully, the elimination of the noise depends on the task at hand.

Determine optimal model for reducing noise in data was the third objective. In this study, the intent was the provision of appropriate framework for social media networks marketing. This was done by adopting and dropping measurable variables depending on their path costs, which is a pictorial representation of associations. Minimum and maximum TF-IDF demographics and behavioral values calculated by the algorithm for the nine goods/services were obtained and tabulated in Table 4.3 that follows:

Table 4.3: Minimum and Maximum TF-IDF Values

Latent Variable	Document Type	Minimum TF-IDF Value	Maximum TF-IDF Value
Demographics	Auto-mobile	2	20
Behavioral	Auto-mobile	3	18
Demographics	Furniture	3	14
Behavioral	Furniture	2	13
Demographics	Clothing	2	10
Behavioral	Clothing	2	9
Demographics	Entertainment Electronics	1	13
Behavioral	Entertainment Electronics	5	10
Demographics	Computers	2	15
Behavioral	Computers	2	13
Demographics	Mobile phones	2	17
Behavioral	Mobile phones	2	15
Demographics	flowers	2	15
Behavioral	flowers	2	13
Demographics	Beauty Services	2	9
Behavioral	Beauty Services	2	8
Demographics	Kitchenware	2	10
Behavioral	Kitchenware	2	7

An appropriate scale created by the researcher covering all the above values in table 4.3 was established, taking care of all the minimum and maximum values as shown in Table 4.4 that follows:

Table 4.4: TF-IDF Scaling

TF-IDF Value	**Scale**
9 - 20	Strongly Significant
6 - 8	Moderately Significant
2 - 5	Significant
0 - 1	Insignificant

The scaling in table 4.4 above was used to code all the TF-IDF values into SPSS. This data was then saved with a *.Sav* extension and fed into Lisrel software to be utilized for model development.

Figure 4.3 below gives the results obtained when the path diagrams were built and ran using the TF-IDF SPSS coded data.

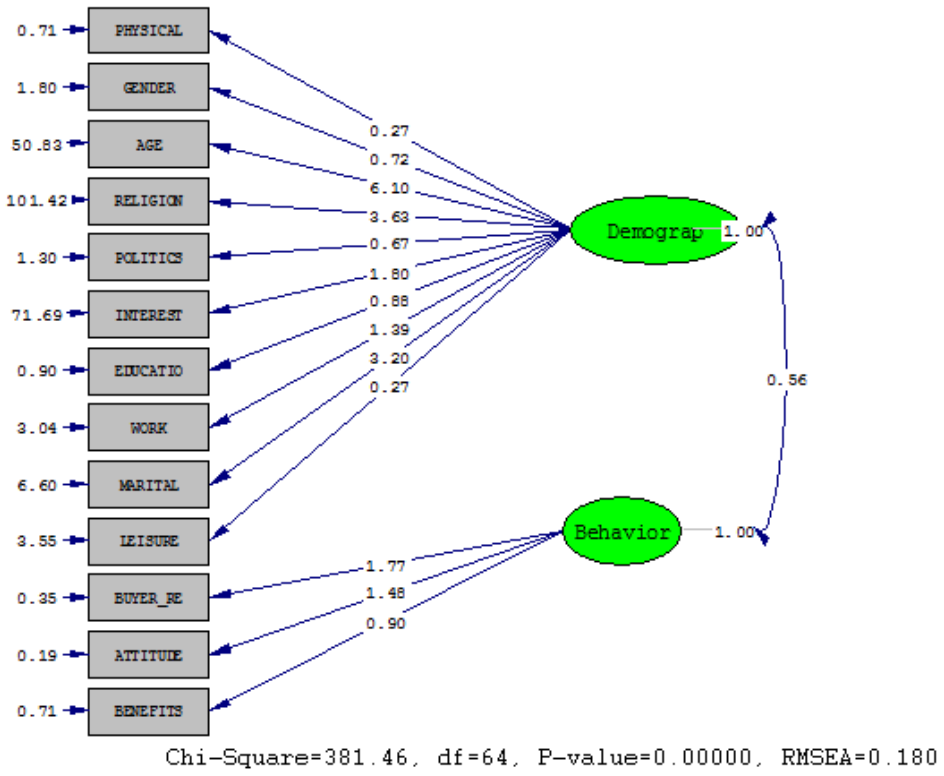


Figure 4.3: Demonstrates Path Diagrams with Statistical Estimates Loadings

Figure 4.3 above shows the connections between the measurable variables and the latent variables. The loadings to the left of each measurable variable represent the error variances. The straight lines connecting the measurable variables and the latent variables are the path costs.

Results from the model in Figure 4.3 above reveal that all the measurable variables' path costs were significantly above the threshold value of 0.05. The consequence of this observation is that this developed model is the attuned one for the given dataset.

The SEM modeling used in this perspective adopted all the measurable constructs, meaning that all of them were good indicators, although their effects differed as demonstrated by the varying values on their path costs.

Therefore, no further optimization was required, which means that the two latent variables (behavioral constructs and demographics) were sufficiently related to their measures.

Make deductions that will lead to design of appropriate framework for social media networks marketing was the fourth objective in the study. This study sought to establish the indicators that could be used to predict customers buying patterns for items advertized through digital social network platforms. These indicators were hypothesized to be consisting of both behavioral and demographics nature.

Following the large interange difference between groups, youth, market and nationality as per results obtained from tf-idf calculations in table 4.3 above, the deduction that follow is that *'Groups', 'youth', 'market'* and *'Nationality'* were *'noisy'* in this dataset and therefore should not be considered in the design of an appropriate framework for social network marketing. In other words, it was obvious that respondents were members of various online groups. Moreover, it was obvious that the population constituted of Kenyans.

For the Automobile dataset, the youths showed little interest in automobile, majorly because they could not afford them. The customers who resided in locations apart from county headquarters and places such as Kisumu, Nairobi and Mombasa (City) also exhibited little attention in automobile. This could be attributed to high cost of internet access, unavailable internet infrastructure and high cost of smart phones that could access the internet in these areas. Lack of skills on how to accomplish online surfing searching for goods and services and online buying could also be another contributing factor to poor attention exhibited by people from market physical address.

This trend is displayed in all the other remaining eight goods/services. Example two is seen in Nationality where across all the nine goods/services as per table 4.3 above TF-IDF value for "foreign" had very low values compared to its counterpart "Kenyan" these were regarded as noise hence eliminated from further analysis such as clustering and framework testing.

This also means, if the TF-IDF values for adults are much higher than that of the youth, the consequence is that adverts should be more skewed to fit adults than youth.

Figures 4.4, 4.5 and 4.6 below give results and relationships for goods/services stratified against latent demographics physical address marital status and work, without market which has been eliminated as noise.

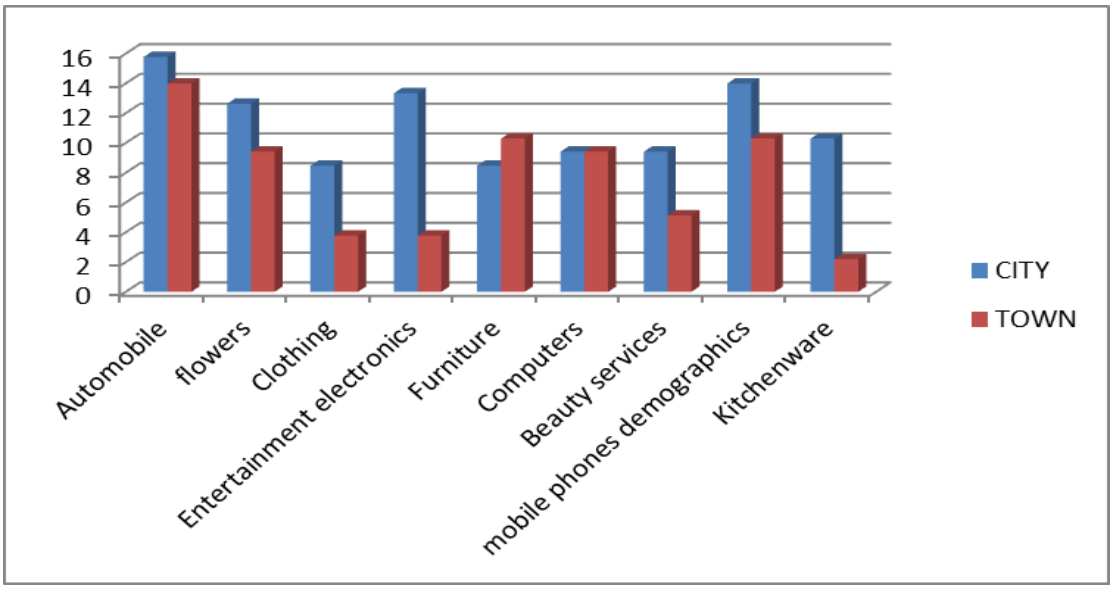


Figure 4.4: Demonstrates Relationship for goods/services against latent demographics city and town

In Figure 4.4 above tf-idf weighting values for automobile, flowers, entertainment electronics and mobile phones for city dwellers was high compared to that of town dwellers. The result is adverts for automobile, flowers, entertainment electronics and mobile phones will be skewed to fit city dwellers. Computers on the other hand has equal tf-idf weighting values for both city and town dwellers. The result therefore, is computer adverts will be skewed to fit both city and town dwellers for this dataset.

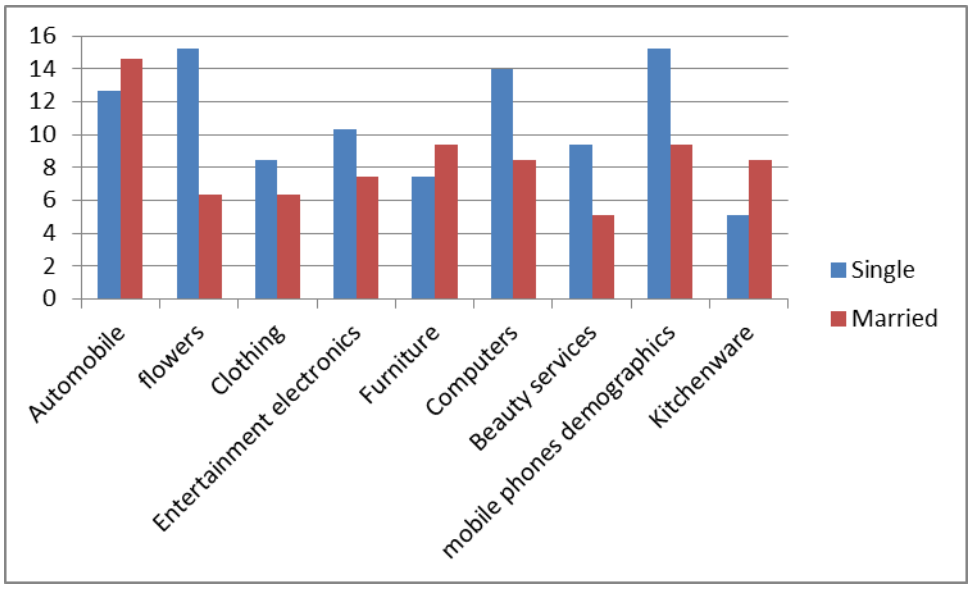


Figure 4.5: Demonstrates Relationship for goods/services against latent demographics single and married

In Figure 4.5 above tf-idf values for single users was high on flowers, clothing, entertainment electronics, computers, beauty services and mobile phones compared to that of married users. The result is adverts for flowers, clothing, entertainment electronics, computers, beauty services and mobile phones will be skewed to fit single users. Tf-idf value for married people was high on furniture and kitchenware compare to that of single users. The result is adverts for furniture and kitchenware will be skewed to fit married users for this data set.

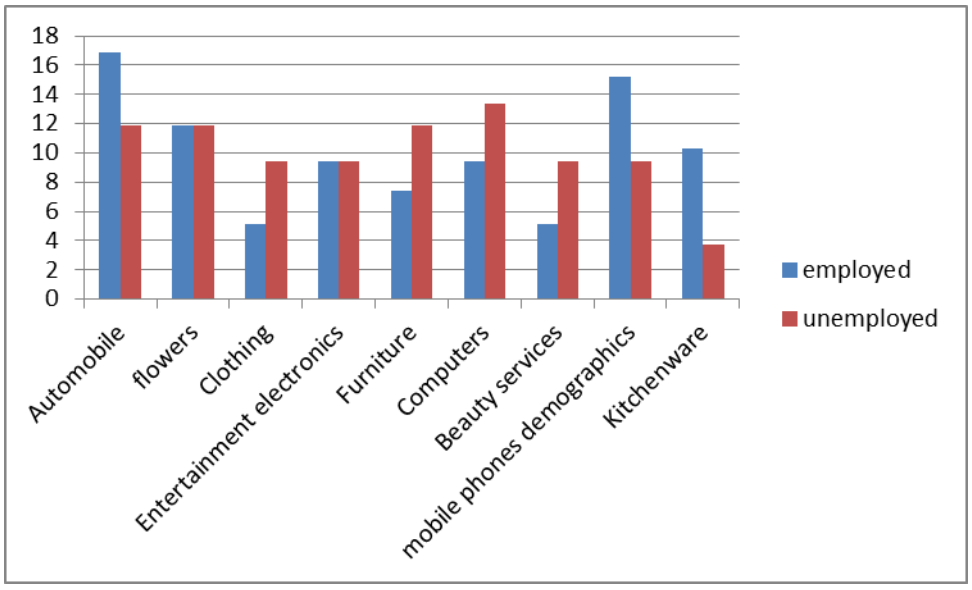


Figure 4.6: Relationship of goods/services against latent demographics employed and unemployed

In figure 4.6 above tf-idf values for employed users was high for automobile, mobile phones and kitchenware compared to that of unemployed users. The result is adverts for automobile, mobile phones and kitchenware will be skewed to fit employed users. Tf-idf values for flowers and entertainment electronics was the same for both employed and unemployed users. Result is adverts for flowers and entertainment electronics will be skewed to fit both employed and unemployed users. Tf-idf values for clothing, furniture, computers and beauty services was high for unemployed users compared to that of employed users. The result is adverts for clothing, furniture, computers and beauty services will be skewed to fit unemployed users for this data set.

In this study the fifth objective was to Test the clustering framework for social media network adverts. The first activity in cluster testing was the development of the clusters themselves. The SPSS feature of 'Direct Marketing' was utilized to cluster online respondents into various groups (clusters). According to Jay (2013), the direct marketing tool is intended to provide an easy-to-use one-stop shop for analysts working within a direct marketing context. As such, it helps users to understand their customers in greater

depth improve marketing campaigns and maximize the ROI of their marketing budget. Using this feature, the respondents were classified into two categories as shown in Figure 4.7 below. The model summary indicates clearly that based on the underlying respondents' data, two clusters were obtained.

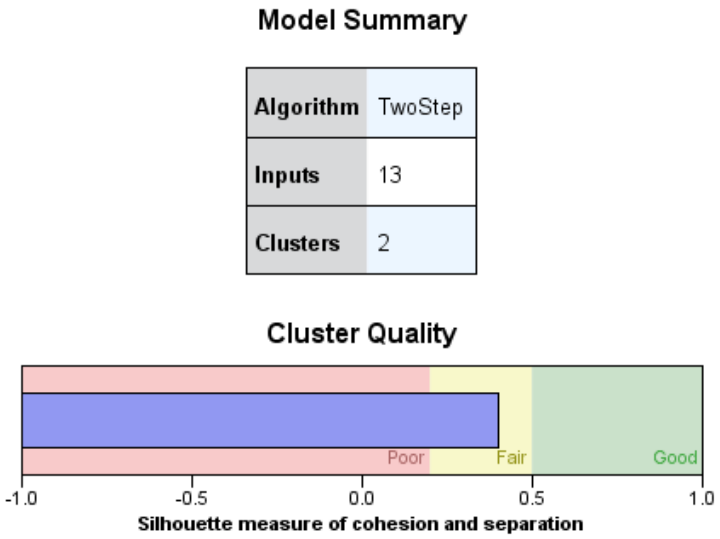
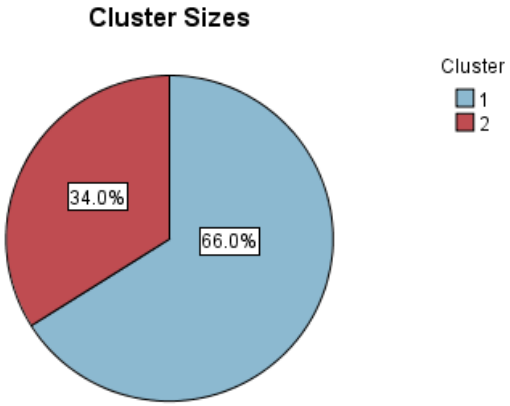


Figure 4.7: Two Clusters obtained from the data

To establish more information on these clusters, figure 4.7 above was double-clicked to display the information that the size of the largest cluster was 101, representing 66% of the population, while that of the smallest cluster was 52, which represented 34% of the entire population as represented in a pie chart as shown in Figure 4.8 below.



Size of Smallest Cluster	52 (34%)
Size of Largest Cluster	101 (66%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	1.94

Figure 4.8: Entire Population Clustering

View mode in SPSS was changed from *'Model Summary'* to *'Clusters'* to reveal more information on the importance and mean of each individual construct for figure 4.8 as shown in figure 4.9 below showing the clusters input (Predictor) importance.

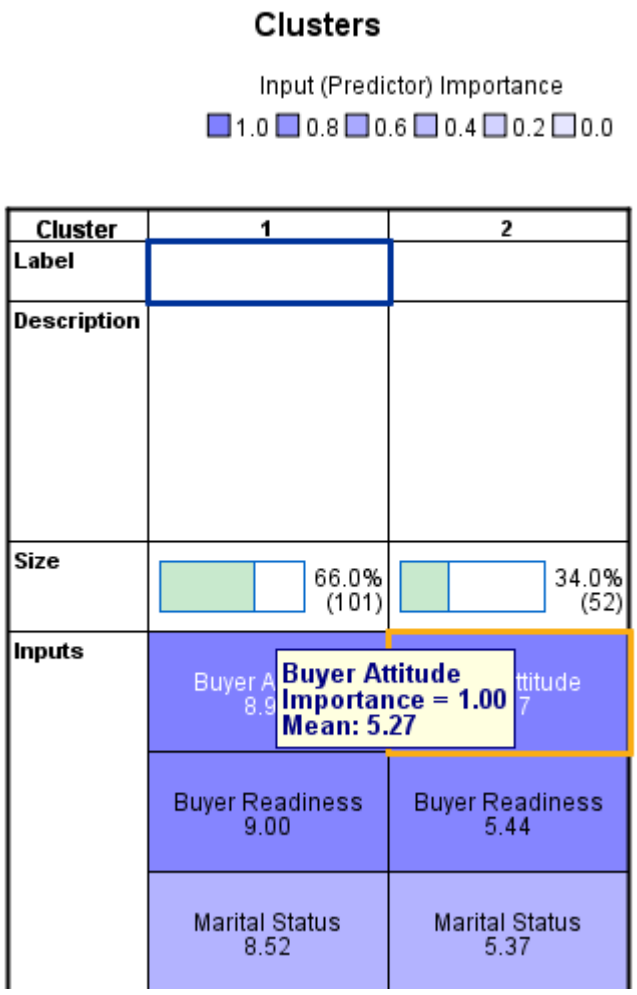


Figure 4.9: Clusters (input) predictor Importance

Figure 4.10 below shows the importance of variables for the two clusters.

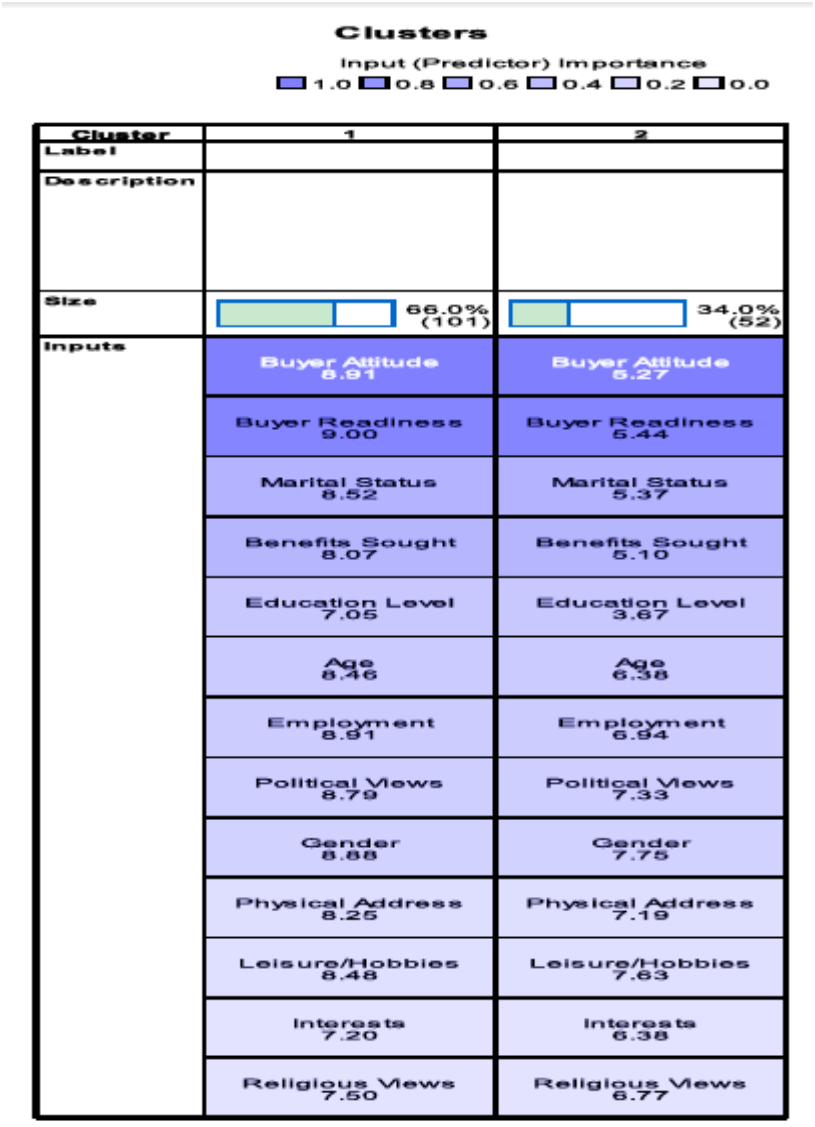


Figure 4.10: Importance and Mean of Individual Constructs

The mean for all the variables are indicated in Figure 4.10 above. Note the variables are arranged in their order of importance. Table 4.5 below shows mean of the variables for the two clusters and their importance as obtained from SPSS.

Table 4.5: Importance of Variables in the clusters

Cluster 1: 66% (101)	**Cluster 2: 34% (52)**	**Importance**
Buyer Attitude = 8.9	Buyer Attitude = 5.27	Importance = 1.00
Buyer Readiness = 9.00	Buyer Readiness = 5.44	Importance = 0.95
Marital Status = 8.52	Marital Status = 5.37	Importance = 0.50
Benefits Sought = 8.07	Benefits Sought = 5.10	Importance = 0.47
Education level = 7.05	Education level = 3.67	Importance = 0.30
Age = 8.46	Age = 6.38	Importance = 0.26
Employment = 8.91	Employment = 6.94	Importance = 0.26
Political Views = 8.79	Political Views = 7.33	Importance = 0.23
Gender = 8.88	Gender = 7.75	Importance = 0.19
Physical address = 8.25	Physcial address = 7.19	Importance = 0.08
Leisure/Hobbies = 8.48	Leisure/Hobbies = 7.63	Importance = 0.07
Interests = 7.20	Interests = 6.38	Importance = 0.03
Religious Views = 7.50	Religious views = 6.77	Importance = 0.03

The consequence of this arrangement is that among all the variables, buyer attitudes were the most important while religious views were the least important.

To test the clustering framework, non-parametric test was done on the framework. According to Singh (2013), nonparametric statistics or distribution-free tests are those

that do not depend on parameter estimates or precise assumptions about the distributions of variables. This approach is ideal because probability statements obtained from most nonparametric statistics are exact probabilities, regardless of the shape of the population distribution from which the random sample was drawn.

Specifically, Wilcoxon's Matched Pairs Signed-Ranks Test was employed to test the relationship between the behavioral and demographics. This is a nonparametric alternative to the paired-samples test. The only assumptions made by the Wilcoxon test are that the test variable is continuous and that the distribution of the difference scores is reasonably symmetric (Stephen & Mark, 2010).

The first step was the formulation of hypothesis, which were to be accepted or dropped from the results obtained. This required that combinations between the demographics and behavioral constructs be carried out. Table 4.6 below shows the possible combinations.

Table 4.6:Model testing criteria

Demographics	Behavioral
Physical Address/ Gender/Age/Religion/Politics/ interests/education/work/Marital status/leisure	Buyer readiness
	Buyer Attitude
	Benefit Sought

The first testing to be done was that of physical address against buyer readiness. Figure 4.11 below shows the data obtained. The null hypothesis was that the median of difference between buyer readiness and physical address was equal to zero.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between Buyer Readiness and Physical Address equals 0.	Related-Samples Wilcoxon Signed Rank Test	.764	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Figure 4.11: Hypothesis Test Physical Address versus Buyer Readiness

When the test was run, the significance level was identified to beat 0.764, significantly above the threshold level of 0.05 (Qian-Li Xue, 2013). Therefore, the deduction was that the null hypothesis be retained. This meant that physical address influenced the buyer readiness.

A similar approach was followed for the rest of the constructs and the information obtained is tabulated in Table 4.7 below.

Table 4.7: Construct Testing Through Significance Levels

Construct Combination	**Significance Levels**		
	Buyer Readiness	Buyer Attitude	Benefit Sought
Physical address	0.764	0.489	0.006
Gender	0.000	0.000	0.000
Age	0.969	0.785	0.025
Religion	0.045	0.076	0.933
Politics	0.020	0.007	0.000
Interests	0.03	0.07	0.342
Education	0.000	0.000	0.000
Work	0.097	0.036	0.000
Marital Status	0.194	0.376	0.157
Leisure	0.090	0.032	0.000

The constructs that had asymptotic significance levels of below 0.05 had their null hypothesis rejected while those whose asymptotic significance levels were above 0.05 had their null hypothesis accepted. In such cases that had null hypothesis rejected led to the acceptance of their alternate hypothesis. The alternate hypothesis stated that the two constructs never influenced each other while null hypothesis stated that they influenced each other.

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.1 Summary

The purpose of this research work was enhancing clustering of users in social media networks for improved digital marketing. This is in recognition of the fact that many people have social media network accounts such as Twitter and Facebook that can be effectively utilized to convey advertisement information to them. Using these platforms, digital marketers have access to behavioural and demographic data that can be mined to extract actionable patterns. However, social media data can contain a large portion of noisy data. As such, the definition of noise becomes complicated and relative because it is dependent on the task at hand. Therefore, this study sought to address the challenges of noise removal from social media data and clustering users in social media networks for improved digital marketing. The researcher formulated five objectives which were to: Identify noise in data depending on task at hand for enhanced clustering; Use weight allocation in TF-IDF environment in order to enhance detection of noise in data; Determine optimal model for reducing noise in data; Make deductions that will lead to design of appropriate framework for social media networks marketing; and to test the clustering framework for social media network adverts. To achieve these objectives, a quantitative research design was adopted. Data that was collected from Facebook users of Soko Kuu and Soko Nyeusi Official groups was coded into numerical format to allow for mathematical analysis. This analysis involved application of the term frequency-inverse document frequency (TF-IDF) algorithm to identify noise in the collected data by using this algorithm to assign weights to the social media data. TF-IDF value was dependent on the frequency that a particular word or phrase appeared in one document category as well as in the entire document collection. Lisrel software which implements structural equation modeling (SEM) was used to establish the relationship between latent variables and measurable constructs in form of a model. These relationships were in the

form of path diagrams that were drawn from the latent variables towards the measurable constructs. This meant that path analysis was necessary to determine the degree of correlation among the research variables. A significant level of 0.05 was used which meant that paths with costs of less than this value were to be dropped. However, all the constructs loaded higher than this value hence were all adopted, although their extents of associations were of varying magnitude. The path costs and TF-IDF values provided the ground for deductions to be derived concerning the design of appropriate framework for social media networks marketing. The clustering framework for social media network adverts that was developed using the direct marketing feature and was tested using non-parametric analysis, which are distribution-free tests that do not depend on parameter estimates or precise assumptions about the distributions of variables.

5.2 Conclusions

The study on enhancing clustering of users in social media networks for improved digital marketing was carried out well and achieved all it's laid down objectives.

5.3 Recommendations

The study findings that were obtained from the field of study confirm the fact that social media data contain a large portion of noisy data which needs to be eliminated before numerical computations to cluster users and get some hidden patterns within these clusters can be done. This will facilitate marketers target each of the identified clusters with relevant advertising messages. In the final analysis, this will lead to successful social media marketing campaigns. To achieve the research objectives, data was mined manually from the profiles of Facebook Soko Kuu and Soko Nyeusi Official group users who reacted to goods/services posted for sale under the nine goods/services identified. TF-IDF algorithm was employed. Therefore, there is need to explore how this data can be automatically extracted from online social media networks and computations that will determine the appropriate groups advertisers can target be done instantly when the

advertiser is creating an online social media marketing campaing. This will then inform the advertisers on which groups to target with relevant adverts at the time of selecting target audience. Future works should therefore, focus on automatically extracting data from online social media networks and performing computations to segement users for target market execution instantly at the time when an advertiser is creating a marketing campaing.

REFERENCES

Adelodun, O., & Obilade, T. (2011). Identification of Factors Affecting Educational Performance of Nigerian Adult Learners: A preliminary Study. African Research Review, *An International Multi-Discipline Journal Ethiopia, 5*(2), 225- 232.

Adelodun, O.A. (2008). *Application of Structural Equation Modeling to Latent Constructs for Evaluating Educational Performance of Adult Students*. Unpublished MSc Thesis, Ile- Ife, Nigeria: Obafemi Awolowo University.

Aggarwal, C.C, & Zhai, C. (n.d.). *Mining text data*. Boston/Dordrecht/London: Kluwer Academic Publishers.

Ayodele, O., & Olushina, A. (2013). *Using Lisrel Program for Empirical Research*

Barbara, R. (2009). *Ten Steps Applied to Development and Evaluation of Process-Based Biogeochemical Models of Estuaries*

Blackwell, M. E. (2006). *Consumer Behaviour (10th ed.).* New York: Thomson Learning.

Chaffrey, D., Ellis-Chadwick, F., Mayer, R., & Johnston, K. (2013). *Digital Marketing: Strategy, Implementation and Practice* (5th ed.). Edinburg Gate, England: Prentice Hall

Creswell, W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Upper Saddle River, NJ: Prentice Hall.

Crosby, L., & Talley, C. (2014). *How to take your retargeting to the next level, Advance strategies from the Industry experts*. Adroll

Franceschi-Bicchierai, L. (2012). *The Evolution of Digital Advertising, INFOGRAPHIC*. Retrieved from http://mashable.com/2012/09/25/the-evolution-of-digital-advertising-infographic/

Han, J., & Kamber, M. (2006). *Data Mining Concepts and Techniques*. (2nd ed.).USA: Morgan Kaufmann Publishers.

Holdren, A. (2012). *Google Adwords.* O'Relly Media

IDC Governments Insights, (2014). Breaking the barriers with Technology: A special report on the Kenyan ICT Market. *Connected Kenya ICT Authority Kenya*, 2

Jain, K.L., Chitta, R., & Jin, R. (2012). *Clustering big data.* Unpublished lecture notes, Department of Computer Science, Michigan State University

Kotler, P. & Armstrong, G. (2012). *Principles of Marketing.* (14th ed.). New Jersey: Prentice Hall

Leskovec, J., Rajaraman, A., & Ullman, D.J., (2014). *Mining of Massive Datasets*. Stanford: Stanford University.

Liu, H. (2013). *Some Computational Challenges in Mining Social Media*. Unpublished lecture notes, Niagara Falls, Canada: Arizona State University

Nyakomitta, S.P.& Omollo, V. (2014). Biometric-Based Authentication Model for E-Card Payment Technology. *IOSR Journal of Computer Engineering (IOSR-JCE), 16* (5), 137-144.

ReTargeter, (n.d.). *A Comprehensive Guide To Retargeting*. Retrieved from https://retargeter.com/presentations/A_Comprehensive_Guide_To_Retargeting.pdf

Saa, S. (2012). A Marketer's Guide to Analytics White Paper, 3-4

Singh U. (2013). Non Parametric Tests: Hands on SPSS.ICAR Research Complex for NEH Region, Umiam, Meghalaya

Stephen D., Mark S. (2010). Mining distance-based outliers in near linear time with randomization and a simple pruning rule. Proceedings of the ninth ACM SIGKDD International conference on Knowledge discovery and data mining

Stokes, R. (2011). *eMarketing, The essential guide to digital marketing* (4th ed.). South Africa: Quirk (Pty) Ltd

Taylor, M., & Colwell, B. (2014). *The State of Always-On Marketing Study*. (2014). Adobe, Razorfish

William, W. (2010). *Minorthird: Methods for identifying names and ontological relations in text using heuristics for inducing regularities from data*.

Zafarani, R., Abbasi, M.A.,& Liu, H. (2014). *Social Media Mining – An Introduction*, (Draft Version). Cambridge: Cambridge University Press.

APPENDICE

Appendix i: Publications and Presentations Related to this Thesis

Shuma, C. E., Waweru, M., & Kimwele, M. (2016). Utilization of Weight Allocation in Tf-Idf Environment for Noise Detection Enhancement. IOSR Journal of Computer Engineering (IOSR-JCE), *18*(1). Retrieved From www.iosrjournals.org