

Utilization of Weight Allocation in Tf-Idf Environment for Noise Detection Enhancement

Evelyne Chanya Shuma¹, Prof. Waweru Mwangi², Dr. Michael Kimwele³

¹(School of Computing and IT, Jomo Kenyatta University of Agriculture and Technology, Kenya)

²(School of Computing and IT, Jomo Kenyatta University of Agriculture and Technology, Kenya)

³(School of Computing and IT, Jomo Kenyatta University of Agriculture and Technology, Kenya)

Abstract: Social media data mining has gained significance in the recent past. This owes to the importance of establishing hidden patterns in the social media data that can be used in digital marketing strategies. It helps the marketers to segment customers according to their demographics and behavioral characteristics such that it becomes easier to target these customer segments with advertisement messages suitable for their cluster. As such, it becomes paramount to discover and eliminate any data that may not influence customer buying trends. This constitutes noise removal. Weight allocation then becomes crucial in identifying keywords from social media data that can help in the clustering process. In this paper, weight allocation was applied in a term frequency-inverse document frequency (TF-IDF) environment to recognize noisy data and remove it before the social media data can be exposed to further analysis. In this approach, a word that appears more frequently in a given document but rarely in the whole document collection was given a higher weight than that word that appears virtually in all documents.

Keywords: Weight allocation, Algorithm, term frequency, document frequency, TF-IDF, mySQL, data mining

I. Introduction

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 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 Online Social Networks to customers.

The utilization of social networks has grown radically over the recent past. This has been as a result of users reaching out to friends, new acquaintances and businesses. Curran & Lennon (2011) developed a model that could help demonstrate how antecedent beliefs influence a user's attitude toward a social network and how those feelings influence the user's intentions to engage in different social networking behaviors. The social media users range from tech-savvy young adults to baby boomers and older adults seeking ways to reconnect with family and friends (Anderson, 2009).

According to Ito et al., (2008), modern college students have grown up with video games and social networking. These students fancy to be connected at all times to their friends and family through texting, instant messaging, mobile phones and Facebook. Protalinski (2012) noted that popular examples of social networking sites include Twitter and Facebook. In December, 2011, Facebook surpassed having 845 Million users worldwide. Therefore, these sites provide a rich marketing platform if only the various users could be targeted with suitable advertisement messages.

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 by-product, by understanding individuals better, we can design better computing systems tailored to individuals' needs that will serve them and society better (Zafarani et al, 2014).

Blackwell et. al., (2006) define behavioral variables as those 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. On the other hand, demographics cover the entire general public, or groups defined by criteria such as education, nationality, religion and ethnicity (Perreau, 2015).

Individual online behaviour can be categorised into three: User-user behaviour. This is the behaviour individuals exhibit with respect to other individuals for example befriending someone, sending a message, playing games, following, inviting or chatting. User-Community behaviour. The target of this type of behaviour is a community for example joining or leaving a community, becoming a fan of a community or participating in its discussions. User-Entity Behavior. The target of this behaviour is an entity in social media. For instance it includes writing a blogpost or review or uploading a photo to a social media site (Zafani et al, 2014).

With the rise of social media, information sharing has been democratized. As a result, users are given opportunities to exhibit different behaviors such as sharing, posting, liking, commenting, and befriending conveniently and on a daily basis. By analyzing behaviors observed on social media, we can categorize these behaviors into individual and collective behavior. Individual behavior is exhibited by a single user, whereas collective behavior is observed when a group of users behave together. For instance, users using the same hashtag on Twitter or migrating to another social media site are examples of collective behavior. User activities on social media generate behavioral data, which is massive, expansive, and indicative of user preferences, interests, opinions, and relationships. This behavioral data provides a new lens through which we can observe and analyze individual and collective behaviors of users (Zafarani, Liu 2014)

With more than 1.5 billion users worldwide, social media offers a treasure trove of information in the form of real-time, interactive communications made available through blogs, tweets, updates, images and videos. Not surprisingly, organizations are growing more and more reliant on social media to understand and work more responsively with employees, vendors and customers, and better gauge the competition. However, mining and analyzing the huge volumes of unstructured data generated by social media is no easy task (Kumar et al, 2013).

Individuals can use mobile devices, desktop computers, and/or notebooks as the access method to reach the social networking sites, as these information and communication technologies have become necessary tools to engage in content creation and information sharing. The usage patterns of the various access methods can potentially have an impact on the usage behavior and information sharing of the various social networking sites. For example, if users believe that mobile devices are complicated, they will be less likely to use those devices (Wakefield and Whitten, 2006). Yet, if firms simplify the operations of new applications and ensure that a significant number of users have that access method, such as Smartphones or iPads, the value and use of that access method increase substantially (Wang, Lo, and Fang, 2008). It was not within the scope of this study to determine the devices used to access social media and whether these devices determined the behavior exhibited by the users on the social media sites.

According to an Experian Marketing Services study, U.S. consumers spend 27% of their total Internet time on social networking sites and forums. Facebook has more than 1.1 billion active users. Twitter, on average, records 58 million tweets every day. These statistics, combined with the millions of blog posts on the Internet and discussions that occur by the minute on forums and social networking sites, for example, can provide a rich and growing pool of data on market trends and such things as consumer interests and perceptions. Still, data is one thing; analysing it successfully to gain useful insights is quite another (Kumar et al, 2013). Social Media Mining is the process of representing, analyzing, and extracting actionable patterns from social media data. 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 behaviour. 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 et al, 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 et al, 2014).

A recent study conducted by Razorfish and Adobe reveals that the majority of marketers lack the ability to tackle behavioral data. Although more and more companies have realized the importance of using data to target consumers and deliver real-time experiences today, 76 percent of marketers have failed to use behavioral data in segmentation analysis and targeting execution, according to The State of Always-On Marketing Study. "There's no doubt that digital marketers have access to behavioral data, and there's been considerable talk and investment these last few years on 'big data,' and an executive focus on data and analytics. But many businesses are struggling to translate this data with the right technology and skills into better data-led customer-facing experiences," says Mark Taylor, vice president of analytics at Razorfish, and author of the study." (Yuyu Chen, 2014).

The challenge of unstructured data. According to Gartner, 80% of enterprise data documents, e-mails, call logs, corporate blogs and the like – is unstructured (i.e., it does not fit into any traditional database). The proliferating use of social media data (including tweets and comments in colloquial style, images, videos, blog posts, etc.) is exponentially increasing the amount of unstructured data to be sorted, analysed and used to gain meaningful insights. Yet most organizations do not have the resources or tools needed to sift through and interpret the vast quantities of social media data they have at their disposal without making considerable changes to their IT infrastructure, operational processes and organizational structure (Kumar et al, 2013).

Due to its increasing significance and quickly evolving nature, social networking research provides both scholars and practitioners with the prospect of better appreciating and leveraging the merits of online behavior. As such, it has become the most vital constituent of the marketer's toolkit in the foreseeable future.

This research paper sought to collect data from social media networks, remove noisy data from these data through the application of weight allocation in a term frequency-inverse document frequency (TF-IDF) environment. The ultimate goal was to eliminate those demographics constructs that do not directly influence users' online behaviors. The remaining variables were then proposed to be vital for the design of a digital marketing framework in social media networks. The target population was the facebook's Soko Kuu and Soko Nyeusi users. The data was collected using tally sheets and fed into MySQL database. The algorithm was designed that could establish the words' term frequency and document frequency and output the TF-IDF values for each word. The words that had comparatively low TF-IDF values were regarded as noise and this formed the basis for their elimination. This paper delimited itself to user demographics data that may influence customer buying trends. A target population of 155 was employed and TF-IDF values were employed as a basis for noise elimination.

Social Media Demographics Oriented Behavior

Curran and Lennon (2011) investigated the direct effects of beliefs to the users' attitudes toward social networks. They also tested whether beliefs indirectly or directly affected users' intentions to continue using social networks, recommend social networks, join other social networks, and stop using social networks.

Although demographic data were collected for all participants in the study, the tested models did not put into consideration the possible differences in behaviors among demographic groups. In their study, Lennon et. al.,(2015) added to the understanding of the Curran & Lennon model by examining whether demographic variables are associated with differences in antecedent beliefs, attitudes toward social networks, and reasons for choosing and using specific social networks. Their findings demonstrated that peoples' antecedent beliefs and attitudes towards social networks differ based on gender, age, marital status and parenthood. In addition, they found the differences in the reasons for choosing and using specific networks among demographic groups.

Specifically, they noted that females are more prone than males to believe that social networking is enjoyable. Also, males are more likely than females to engage in social networking due to social influences. Secondly, participants under 30 are more likely to use social media networks than those over 30 since they view them as being easy to use. For the same reasons, participants without children are more likely to utilize social media than those with children.

This study adds to the existing body of knowledge by identifying those demographics constructs that are noisy to the extent that if allowed in further market strategical targeting, they would interfere with this analysis. Therefore, they should be eliminated in order to provide a clear behavioral picture of the online users. This would enable the various groups of online users to be targeted with appropriate advertisement messages.

Social Media Behavioral Data

Tally sheets were employed to collect data from social media sites, namely Soko Kuu and Soko Nyeusi . The users' demographics were collected from their respective facebook profiles. It was then stratified as shown in Table 1 that follows.

Table 1: Demographical Stratification

Latent-Demographics	Measures	LatentDemographics	Measures
Physical Address	City	Religion	Christian
	Town		Muslim
	Market		Other
Gender	Male	Politics	Active
	Female		Inactive
Age	Youth	Marital Status	Single
	Adult		Married
Interests	Academics	Leisure/Hobbies	Entertainment
	Business		Events
	Socializing		
Work	Employed	Education Level	Student

	Unemployed		High School
Groups	Yes		College
	No		University
Groups	Yes	Nationality	Kenyan
	No		Foreign

Table 1 demonstrates the fact that various demographics were collected using different measures. However, it is important to note that not all demographics affect all the customers in all buying circumstances. Therefore, there was need to group customers according to their buying trends based on their motivating factors. Nine (9) goods/services listed in Table 2 were employed for this purpose due to their frequency of adverts in the target social media sites.

Table 2: 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

Procedure

- a) The nine goods/services that were identified 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* members who liked, inquired or bought the advertized goods/services were treated as documents.
- c) The 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).

As an illustration, we 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.

- d) The documents in (C) above were fed to a *Mysql* database from which data mining could be done via TF-IDF algorithms designed using the *PhP* programming language.
- e) The *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. Fig. 1 that follows gives an example of these interfaces.

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 but rarely in the whole document collection. This meant that keywords that appeared in nearly all documents were given a low weight.

In this study, any unwanted data was regarded as noise. To identify this noise, term frequency-inverse document frequency (TF-IDF) was employed. The relationship that was used to calculate TF-IDF for specific terms in documents is given by equation 1.

$$IDF = \log_{10} [N/df] \dots \dots \dots (1)$$

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.

$$TF-IDF = TF * IDF \dots \dots \dots (2)$$

II. Results And Discussions

The study findings that were obtained constituted of the term frequency (TF) values, the inverse document frequency (IDF) and the term frequency-inverse document frequency (TF-IDF) as shown in Fig. 1.

Fig. 1 presents data on the calculated IDF depending on the TF of specific words in the entire documents. As shown, the document of interest in this case is auto-mobile, although other documents could still be used and yield similar results. We are concerned with the calculation of IDF and TF-IDF of the physical address, *City*. Fig. 1 demonstrate that the term *City* appears 16 times in document automobile (TF=16). Note that the total number of document, N in this case is equal to the total population, which was 155(that is, N=155). With this information, our first task is to prove the value of the computed IDF. This is done as follows.

$$IDF = \log_{10} [N/df]$$

Where N=155

df=16

$$IDF = \log_{10} [155/16]$$

$$= \log_{10} [9.6875]$$

$$= 0.9862$$

Term Frequency–Inverse Document Frequency
Computation [Document: - Automobile]

Latent-Demographics	Measures	TF	TF-IDF	Latent-Demographics	Measures	TF	TF-IDF
<i>Physical Address</i>	City	16	15.779307442	<i>Interests</i>	Academics	3	5.1396313303518
	Town	13	13.992048492		Business	3	5.1396313303518
	Market	1	2.1903316981		Socializing	18	16.831065475206
<i>Gender</i>	Male	23	19.05768829	<i>Leisure/Hobbies</i>	Entertainment	28	26.30082671186
	Female	7	9.4166356070		Events	7	9.416635607002
<i>Age</i>	Youth	1	2.1903316981	<i>Education Level</i>	Student	1	2.1903316981703
	Adult	26	20.159317105		High School	2	3.7796034050126
<i>Religion</i>	Christian	24	19.442890959		College	11	12.630329143133
	Muslim	3	5.1396313303	University	10	11.903316981703	
	Other	3	5.1396313303	<i>Work</i>	Employed	18	16.831065475206
<i>Politics</i>	Active	20	17.796034050		Unemployed	10	11.903316981703
	Inactive	10	11.903316981	<i>Groups</i>	Member	28	26.30082671186
<i>Marital Status</i>	Single	11	12.630329143		No Membership	2	1.6202409129174
	Married	14	14.618851274	<i>Nationality</i>	Kenyan	29	21.110077327969
					Foreign	1	2.1903316981703

figure 1: tf-idf computations

The calculation of IDF for the rest of the words followed a similar format. To illustrate how noise removal was accomplished, we consider measurable variables, *physical address*, *age*, *nationality*, *work* and *group membership*. 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. The deduction that follows is that ‘Groups’, ‘youth’, ‘market’ and ‘Nationality’ were ‘noisy’ 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. 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 and searching for goods and services online could also be another contributing factor to poor attention exhibited by people from the markets.

Fig.2, Fig.3 and Fig.4 below give results and relationships for goods/services stratified against latent demographics *physical address* (without *market* which has been eliminated as noise), *marital status* and *work*.

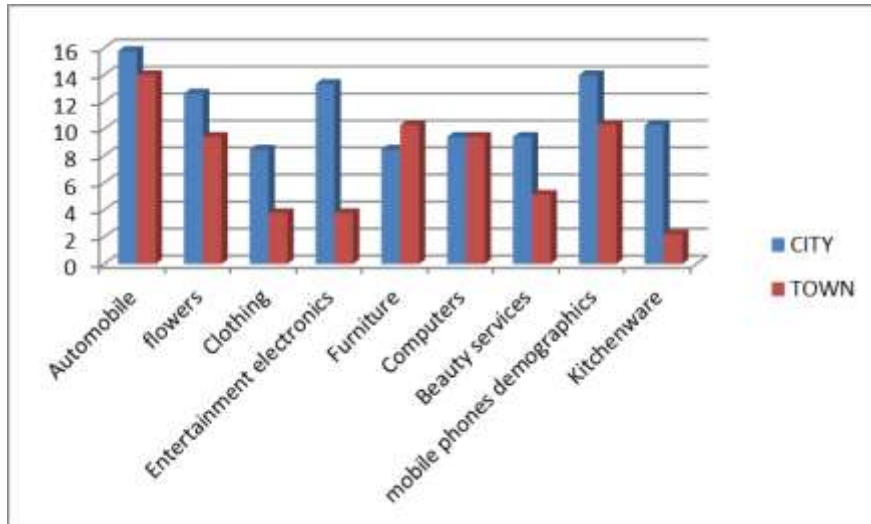


figure 2: relationship for goods/services against latent demographics city and town (physical address)

In Fig. 2 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.

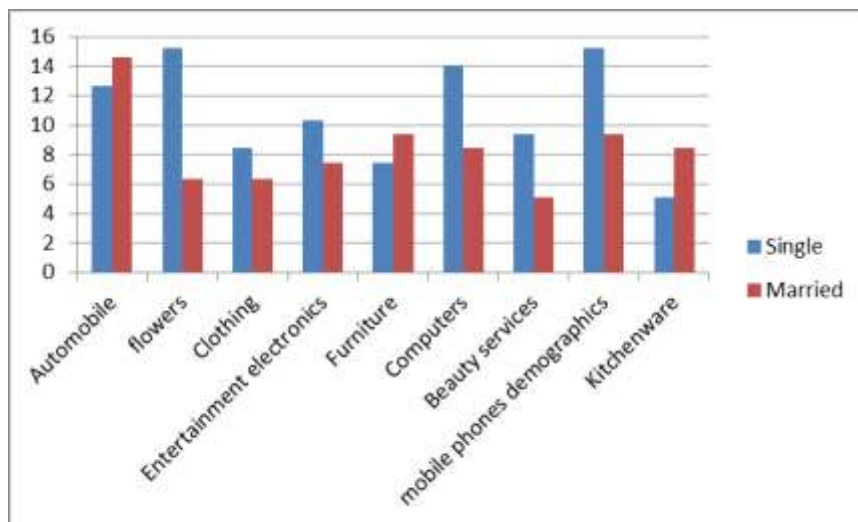


figure 3: relationship for goods/services against latent demographics single and married (marital status)

In Fig. 3 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.

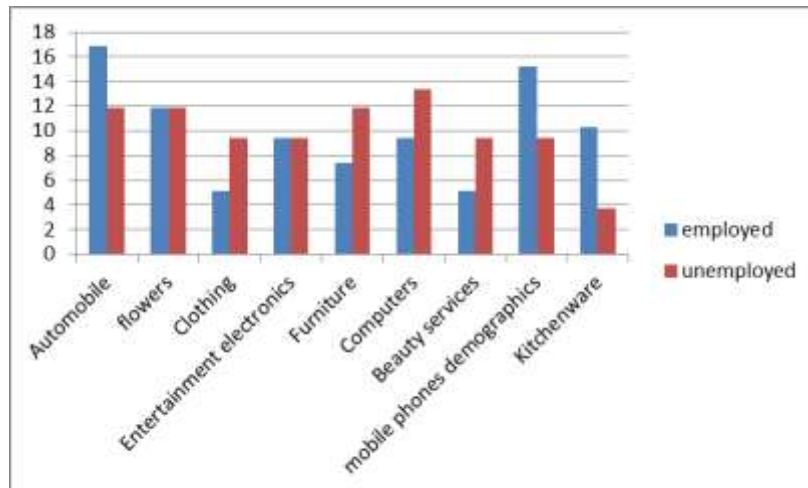


figure 4: relationship of goods/services against latent demographics (work) employed and unemployed

In fig. 4 above tf-idf values for employed users was high for automobile, mobile phones and kitchenware compared to that of single 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.

III. Conclusions And Recommendations

It is arguably true that social media data can contain a large portion of noisy data. This leads to a situation where the definition of noise is intricate and relative. This is so because the removal of this noise is dependent on the task at hand. Therefore, this study sought to address the challenges of noise removal from social media data networks for improved digital marketing strategies. To achieve this, weight allocation in TF-IDF environment was employed in order to enhance detection of noise in data. It was noted that the TF-IDF value dependent on the frequency that a particular word of phrase appears in one document as well as in the entire document collection.

The measurable variables that exhibited low values of the TF-IDF were safely regarded as noise and hence eliminated from future analysis such as non-parametric analysis, path analysis and cluster analysis. This led to the elimination of 'Groups', 'Nationality', 'Market', and 'Youth' from the framework.

For automobiles, the findings indicated that, more positive attitudes toward social networks were expressed by: males compared to females; those over 18 years than those that were under 18 years, married compared to single, city and town residents than market residents, those interested in politics than those not interested, socialites than business people and academicians, those whose leisure is entertainment than those whose leisure is events, college and university students than high school students, employed than unemployed, group members than those who do not belong to any group, and Kenyans than foreigners.

For future work in this area, it is important to note that TF-IDF algorithm was employed. Therefore, there is need to explore how other algorithms such as Learning Explicit Noise Systems (LENS), an approach for identifying corrupted fields and using the remaining non-corrupted fields for subsequent modeling and analysis, can be employed in this perspective.

References

- [1]. Curran, J., & Lennon, R. (2011). Participating in the Conversation: Exploring adoption of Online Social Media Networks. *Academy of Marketing Studies Journal* 15(1), 21-38.
- [2]. Reza Zafarani and Huan Liu, Arizona State University (2014). *Behavior Analysis in Social Media*
- [3]. Anderson, T. H.C. (December 6, 2009). College Students say Facebook is the only Social Networking site that really matters.
- [4]. Ito, M., Horst, H., Bittanti, M., Boyd, D., Herr-Stephenson, B., Lange, G., Pascoe, D.J., & Robinson, L. (2008). *Living and learning with new media; Summary of the findings from the digital youth project*. Chicago.
- [5]. Protalinski, E. (2012). Facebook has over 845 million users.
- [6]. Perreau F. (2015). *The Consumer Factor Consumer Insights, Market Research, Consumer Behavior & Neuromarketing*.
- [7]. Lennon R., Rentfro W. & Curran J. (2015). Exploring relationships between demographic variables and social networking use. *Journal of Management and Marketing Research*.
- [8]. John H. Heinrichs1, Jeen-Su Lim And Kee-Sook Lim (2011), Influence of social networking site and user access method on social media evaluation
- [9]. Yuyu Chen , Over 70 Percent of Marketers Fail to Target Consumers With Behavioral Data [Study], March 27, 2014
- [10]. Reza Zafarani, Mohammad Ali Abbasi, Huan Liu, *Social Media Mining – An Introduction* [Cambridge University Press, 2014]