TARGET AUDIENCE CLASSIFICATION ON TWITTER USING TEXTUAL FEATURE EXTRACTION AND MACHINE LEARING

Arpita Chaudhary[1] Mr. Vineet Khanna [2] Mr. Subhash Chandra Jat [3]
[123] Department of Computer Engineering, Rajasthan College of Engineering for Women, Bhankrota, Ajmer Road, Jaipur

Abstract— Social media is an increasingly important aspect of modern living. It is so much profound that it had delved into the fabric of modern living and now completely indistinguishable from it. It has affected the design of the mobile phones to a large extent, which had, until recently, used for voice based communications. Social media presents an entirely different platform for advertisements as compared to the classical methods of advertising like newspapers and television. Advertising on social media is much cheaper as compared to Newspapers and Television, with interactive graphics and multimedia options. One major advantage associated with the Social Media advertising is Target Marketing. Target Marketing refers to the approach in which the advertisement is shown only to the group of the audience which might have (possibly) interest in that category of item or service. For example, an advertisement of an Android Programming Tutorials Classes might be of interest to young enthusiastic programmer while it is of least concern to most of people. Thus, identification of the groups of people having like interest, which are potential customers of the product or service, is the key aspect of Target Marketing. Target marketing over social media is a field of active research and much of the research has been focused on Facebook and Twitter. This paper focuses upon target marketing approach with machine learning using inference mechanisms. Each of the users is labeled with feature set based on the domain of the content usually posted and shared and his "followers" and "following". The feature set is mapped to the domain of the product and service using manually classified data. The classifier then gives the optimal probability of interest of a particular user in the product of service. The simulation of the model is done over R and the results shows a significant improvement over the benchmark techniques.

Keywords: Target Marketing, Social Media, Machine Learning, Cross Domain Analysis, Topic Models.

I. INTRODUCTION

1.1 Marketing and Social Media

Social media advertising refers to the advertisements displayed to users on social media platforms. As the social networks stores user profile information and also accesses the location, mail and other user attributes, they serve highly relevant advertisements (based on specific interests, behavioral interactions and other custom targeting). In many cases when target market belongs to a particular segment of social platform, social advertising can provide huge return on investments with lower cost of acquisition. Social Media offers one of the most promising options for target marketing. It offers marketing of products and/or services in a customized manner that cannot be done using classical methods like television or print media. Advertisements can be shown in the form of still images, animated banners, interactive graphics or videos. The most popular social media application along-with the number of users is given in Table 1.1

TABLE 1.1
USER BASE OF MOST POPULAR SOCIAL NETWORKING SITES [1]

Social Media App	Number of Active Users (January 2018) (The count is specified in millions)
Facebook	2167
YouTube	1500
WhatsApp	1300
WeChat	980
QQ	843
Instagram	800
Tumblr	794
QZone	568
Sina Weibo	376
Twitter	330
Baidu Tieba	300
Skype	300
Linkedin	260
Viber	260
Snapchat	255
Reddit	250
Line	203
Pinterest	200
Yy	117
Telegram	100
VKontakte	97
BBM	63

Kakaotalk	49

The above table clearly indicates the potential base of the users at social space. The rise in the count of the active users in recent years, in billions, is shown in Figure 1.2

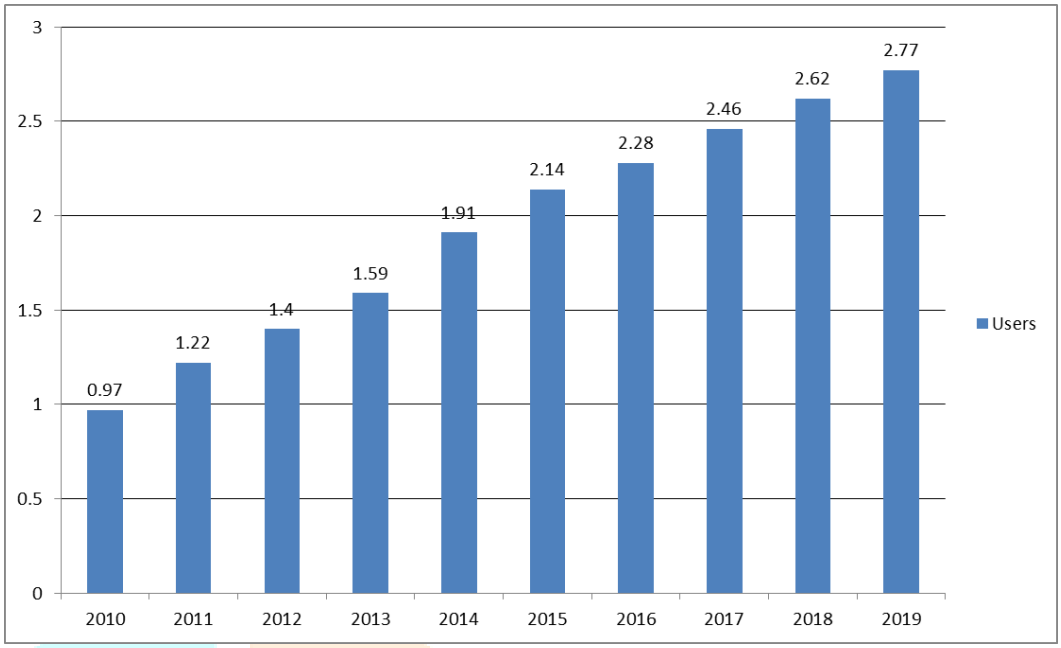

Fig 1.2 The rise in the number of users using social network applications from the year 2010, including the projected score of year 2019 (user count in Billions).

1.2 E-marketing

E marketing is a broad term which refers to online marketing which is used to address the viewer on the internet applications. E marketing is a collection of the following aspects:

1. E-mail Marketing:
   It refers to sending the description of the product or service, written in an e-mail to a set of potential customers. The content in the email is mostly in the form of text and graphics with animations for certain product categories. E-mailers are often used for sending the mail to the recipients. However, bulk mailing from a single sender is often treated as spam by most of the email service providers. Thus, e-mail marketing needs to be designed suitably. It gives best results only in the case when the user itself had shown interest in the product/ service and has subscribed for newsletters, alerts or feed data.
2. Google Search Network

The Google's Search Network is one of Google's main advertising networks. Many organizations use Google Search Network for AdWords campaigns displaying different advertisements across the network. The Network consists of search-related websites and applications where advertisement can appear. Examples include Google Partners such as Google Search, Google Maps, Google Shopping, and other partner websites that match the specific search criteria for AdWords campaigns. The Search Network is an excellent tool to reach customers explicitly searching for your product or service.

This paper is organized as follows. Section 1 gives a brief overview of the paper. It discusses the scope and motivation for the subject matter of the paper in a comprehensive way. Section 2 focuses on the problem statement and the motivation behind the research. It further discusses various categories of machine learning applications in supervised and unsupervised domains. It also covers the fundamental concepts of target marketing and its scoping in the upcoming era of digital world. Section 3 illustrates the proposed technique of target audience clustering. Section 4 discusses the results based on quality metrics and conclusions are drawn. Section 5 concludes the paper.

II. RESEARCH APPROACH AND MOTIVATION

In most cases, the product or service provider want a clear picture of exactly who the audience is – this might include customer profiles that list everything from age to favorite hobbies. Knowing as much as you can about who your buyer is will help you speak directly to them, which is integral for successful social media campaigns. On a large base of users, comprising of millions of users all across the globe, this is a typical problem and the identification of potential customers of a service cannot be done using traditional computing methods. In this paper, a clustering based machine learning approach is presented for the identification of target audience for a product or service.

Within the overall umbrella of an effective digital marketing strategy, social media campaigns can provide traffic and results in an impressive ROI. **Social media posts can be used to drive targeted traffic**. Social media ads are like the new billboards: rather than posting a sign on the side of the road and hoping people catch a glimpse as they drive by, we're posting signs on newsfeeds and hoping people stop to look as they scroll by. But in many ways, ads on social networks like Facebook and Instagram can be more effective than old fashioned roadside advertisements. Whereas billboards are put up for the general masses, social media ads can be very carefully targeted to very specific audiences.

III. PROPOSED CLUSTERING AND FEATURE SET EXTRACTION USING TOPIC MODELS AND SVM

3.1 Proposed Clustering Approach

The classification of a group of potential customers for a product or service from a raw data set is a complicated machine learning task. The integral part of this classification is the extraction of the feature set of a cluster of user. In the proposed scheme, the raw data set is first subjected to the Latent Direchlet Allocation (LDA) module to extract the topic models which are profound in the discussion. The LDA scheme extracts

the hot items of discussion in the Tweets in the form of keywords. Once the major topics are extracted, clustering can be performed using the k-means clustering approach to group the users in accordance with the topics of interest. Two of the most important aspects of this classification are:

1. The number of topics extracted from the raw data depends on the threshold value of the frequency. Typically, this is done in conjunction with the target domain under consideration. A more general domain has a large number of keywords which can be mapped to variety of topics at a fine level of granularity as compared to the target domain which is specific in nature.
2. A user can be a member of a number of clusters. Despite of cluster memberships, other attributes of the users can be of interest like the feature set based on tweet history, the age, location and other parameters related to likes and dislikes.

Figure 3.1 shows the block diagram of the basic steps involved in the accomplishment of the objective. It is important to note that the feature set extracted from users of the potential clusters is the set of positive examples for the training data. The set of negative examples can be formulated from the remaining clusters. Also, the feature set of the potential customers involves the user interests as well as other information included in the user profile as well as tweet history and the browsing preferences. However, the number of negative examples is very large as compared to the numbers of positive examples. Therefore, both positive and negative feature sets are chosen very carefully so as to ensure maximum separation in the support vector, thereby giving accurate classifier to the extent as possible.

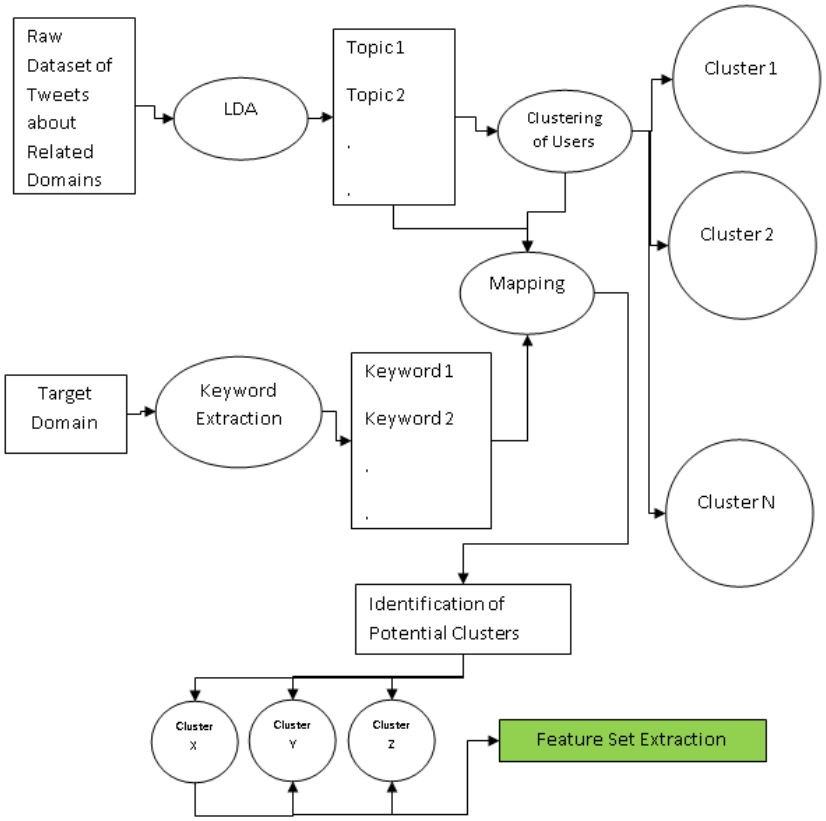

Fig 3.1 Schematic of Feature Set Extraction for the tabulation of positive examples

3.2 Topic Extraction through LDA

Latent Dircihlet Allocation is a technique for the extraction of topics from a given set of corpuses. The application of LDA can be understood by the following illustrative example.

Consider a set of sentences given below:

1. I usually went to Café Coffee Day with friends on weekends.
2. Coffee is recommended only after having some biscuits.
3. Doing something creative require coffee and peace.
4. Pugs are small dogs which are best pets.
5. My brother adopted a dog yesterday.
6. Dogs like coffee flavor biscuits.

The sentences and the topic of discussion is listed in the table 3.1

TABLE 3.1
SENTENCES AND THE TOPICS

Sentence Id	**Topic(s)**
Sentence 1	Coffee, Friends, Weekend
Sentence 2	Coffee, Biscuits
Sentence 3	Coffee and Peace
Sentence 4	Dogs, Pets
Sentence 5	Brother, Dog, Yesterday
Sentence 6	Coffee, Dog, Biscuits

TABLE 3.2
TOPIC LABELS

Topic	Label
Coffee	A
Friends	B
Weekend	C
Biscuits	D
Peace	E
Dogs	F
Pets	G
Brother	H
Yesterday	I

One can deduce the following assumptions qualitatively by analyzing the given sentences. Sentences (1), (2) and (3) are 80 percent about topic A. Sentences (4) and (5) are 80 percent about topic B. Sentence (6) is 70 percent about topic A and 30 percent about topic B. LDA is a modeling technique for this discovery.

LDA typically represents documents as **mixtures of topics. From these topics, the** words are emerged with certain probabilities. LDA assumes that the documents are produced of the words which are related to the topics being discussed. While writing any document, the user usually focuses on the length of the document. Assuming that the document consists of N works, it can be stated without loss of generality that the words follow Poisson distribution. The LDA model assumes that the document is mixture of topics according to Dirichlet distribution over a fixed set of K topics. For example, assuming that we have the two food and cute animal topics above, there might be a document to consisting of 1/3 food and 2/3 cute animals. The document can be generated by generating each word in the document. A word is generated by first picking up a topic, in accordance with the topic weight and then generating the words related to the topic in the preference order required.

This generative model is assumed for the collection of documents. LDA uses a bottom up approach, starting with the given set of documents and tries to backtrack to find a set of topics that had likely generated the text collection. This is illustrated in the figure 3.2.

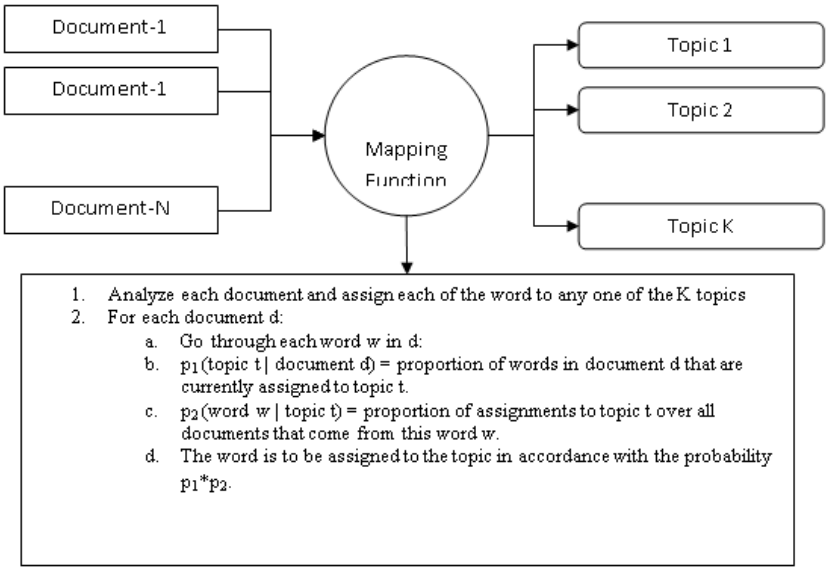

Fig 3.2 LDA Generative Model

Figure 3.2 gives the illustration of the generative model of LDA. It gives an illustration of the basic logic of LDA through which the documents can be assigned to the topics. The steps in the figure shows the LDA bag of words model through which documents can be mathematically be generated.

It is important to note that the count of the topics is identified by the users first on the basis of certain threshold value. This threshold depends upon the minimum number of documents that are related with certain topic of interest. Once the topics are identified, then a mapping is performed over the topics which are under consideration and the prominent topics in the tweets sample. Also, the clustering of the users based on the topics identified from the tweet samples can be made directly using K means clustering approach. The mapped topics from those identified through LDA are identified are isolated and the clusters of users corresponding to these topics are identified. These clusters are the potential customers for the advertisement campaign.

The feature set comprises of the words from the mapped topics to the topic related with the specific product or service. The feature set constitutes the set of positive examples of the proposed SVM classifier. The set of negative examples can be constructed from other topics excluding the one of interest in the problem under consideration. The SVM classifier is then trained on the set of positive and negative examples. After the identification of topics that are currently the topics of hot discussion in the tweets sample, clustering of the users is done on the basis of topics of interest. This is illustrated in figure 3.3.

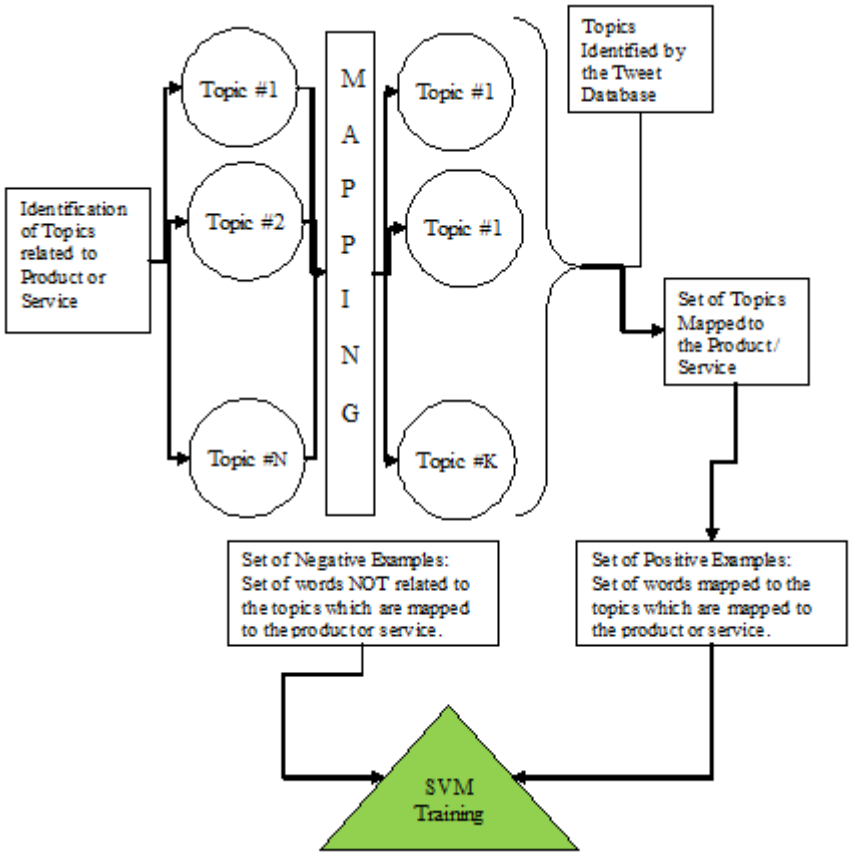

Fig 3.3 Identification of Feature set From the Bag of Words

In figure 3.3, it is illustrated how the feature set comprising of the words from topics can be constructed that forms the positive example for the proposed support vector machine based classifier.
The complete feature set comprises of the parameters from the topics of interest (i.e., words) as well as the feature set which is extracted from the potential user clusters which are mapped to those topics. The operation of the SVM classifier over any user would be performed as shown in figure 3.4.

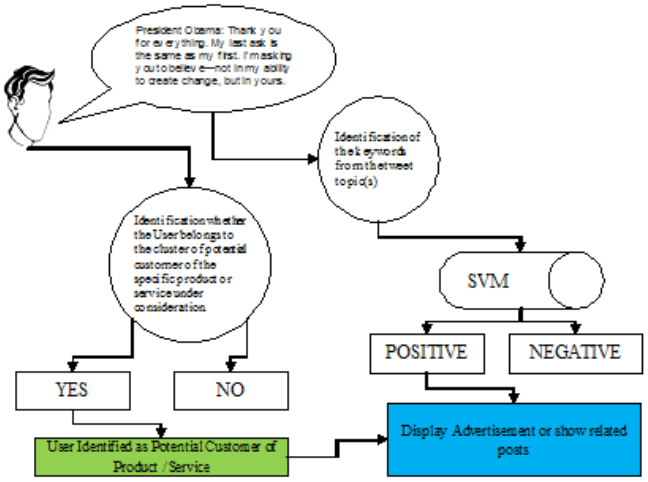

Fig 3.4 Illustration of identification of Potential User for Product or Service

It is important to note that the proposed model gives a twofold approach in contrast to the single approach based on fuzzy classifier proposed by Siaw Ling Lo et. Al []. In the base approach, the authors described the identification of potential customers targeted for a product or service based on the only the topic model extracted using LDA. In contrast the proposed model gives the identity of potential customers based on topic models using SVM classification along-with the clustering approach for the existing users of the social network platform. The existing users are clustered based on the mapped topics in the tweets to the topics related with the product or service under consideration. This results in much more efficient identification of the potential customers. In addition, if some new user, having no or a less populated tweet history, can be judged on the basis of his/her tweet(s) weather or not he/she is a potential customer for the product or service. Section 4 analyzes the methodology presented in this section over the CSV file of 204820 tweets recorded during 14 to 16 April 2016. The simulation is done using R Statistical package and the results are compared with those of the benchmark techniques.

IV. RESULTS

4.1 Tweet Database

The first few records of the CSV file tabulating the records of the tweets is shown in table 4.1.

TABLE 4.1
SAMPLE TWEETS FROM TWEET- DATABASE

User Name	Nickname	Bio	Tweet content
Bill Schulhoff	BillSchulhoff	Husband,Dad,GrandDad,Ordained Minister, Umpire, Poker Player, Mets, Jets, Rangers, LI Ducks, Sons of Anarchy, Survivor, Apprentice, O&A, & a good cigar	Wind 3.2 mph NNE. Barometer 30.20 in, Rising slowly. Temperature 49.3 °F. Rain today 0.00 in. Humidity 32%
Daniele Polis	danipolis	Viagens, geek, moda, batons laranja, cabelos coloridos, compras, sapatos, fotos e livros. E também @blogtrippolis.	Pausa pro café antes de embarcar no próximo vôo. #trippolisontheroad #danipolisviaja Pause for… https://t.co/PhcJ4oYktP
Kasey Jacobs	KJacobs27	Norwich University Class of 2017	Good. Morning. #morning #Saturday #diner #VT #breakfast #nucorpsofcadetsring #ring #college… https://t.co/dBZ7dbwX6f
Stan Curtis	stncurtis	transcendental music, art for art's sake, craftbrew aficionado, sports fanatic, go for the organic, think for yourself-question authority	@gratefuldead recordstoredayus □□□ @ TOMS MUSIC TRADE https://t.co/CURRmn6iJo
Dave Borzymowski	wi_borzo	When in doubt....Panic.	Egg in a muffin!!! (@ Rocket Baby Bakery - @rocketbabybaker in Wauwatosa, WI) https://t.co/mwfhrcxtRp
kirstin	KirstinMerrell	mars model	@lyricwaters should've gave the neighbor a buzz. Iv got ice cream and moms baked goodies □
Joshua Kosches	Jkosches86	I am a John Jay MPA Alum. I live in Miami,FL.	On the way to CT! (@ Mamaroneck, NY in Mamaroneck, NY) https://t.co/6rpe6MXDkB

It is important to note that Bio is one of the most important feature-set which is analyzed before a fan-following takes place on twitter. In the table shown above, a number of Bio keywords are shown. The Bio comprises the roles as well as adjectives that a person finds appropriate for him/her.

TABLE 4.2
SAMPLE BIO FROM TWEET DATABASE

User Name	Nickname	Bio
Bill Schulhoff	BillSchulhoff	Husband,Dad,GrandDad,Ordained Minister, Umpire, Poker Player, Mets, Jets, Rangers, LI Ducks, Sons of Anarchy, Survivor, Apprentice, O&A, & a good cigar
Daniele Polis	danipolis	Viagens, geek, moda, batons laranja, cabelos coloridos, compras, sapatos, fotos e livros. E também @blogtrippolis.
Kasey Jacobs	KJacobs27	Norwich University Class of 2017
Stan Curtis	stncurtis	transcendental music, art for art's sake, craftbrew aficionado, sports fanatic, go for the organic, think for yourself-question authority
Dave Borzymowski	wi_borzo	When in doubt....Panic.
Kirstin	KirstinMerrell	mars model
Joshua Kosches	Jkosches86	I am a John Jay MPA Alum. I live in Miami,FL.
TMJ-PA Retail Jobs	tmj_pa_retail	Follow this account for geo-targeted Retail job tweets in Pennsylvania Non-Metro. Need help? Tweet us at @CareerArc!
Vidal Alexander	Vonfandango	Just ask!
TMJ-BAL Jobs	tmj_bal_jobs	Follow this account for geo-targeted Other job tweets in Baltimore, MD. Need help? Tweet us at @CareerArc!

TMJ-MD Retail Jobs	tmj_md_retail	Follow this account for geo-targeted Retail job tweets in Maryland Non-Metro. Need help? Tweet us at @CareerArc!
Gkalkat	gkalkat	
Dr K Goodson	GodsNaturalDiva	I am an environmental engineering PhD who has recently found her career and is in hot pursuit for my passion. http://t.co/aR3MPUHgCF
Sarah Kehoe	sarahkehoe	#agencylife @boelterlincoln | Coffee Snob | MKE Lover but my ❤ is in CO | Wannabe Blogger: https://t.co/8vQULBOb0I Opinions are my own, obviously.
TMJ- BOS Health Jobs	tmj_bos_health	Follow this account for geo-targeted Healthcare job tweets in Boston, MA. Need help? Tweet us at @CareerArc!
John Hancock	JohnHancockJobs	Official #jobs handle of John Hancock Financial Services, world leader for #insurance & financial products. Follow @JohnHancockUSA for company news & updates
Raj	vaidyeah33	young metro trusts me | rutgers student | #devilsarmy #patsnation
TMJ-DTW Engin. Jobs	tmj_dtw_eng	Follow this account for geo-targeted Engineering job tweets in Detroit, MI. Need help? Tweet us at @CareerArc!
Atlanta IT Jobs666	tmj_atl_it	Follow this account for geo-targeted Software Dev. - General/IT job tweets in Atlanta, GA. Need help? Tweet us at @CareerArc!
N Master	N_Master101	#chicago born and raised #Bulls #Bears #Whitesox #Blackhawks but I'm all over the nation

The frequencies of the top ten words in the Tweets is given in table 3.3

TABLE 4.3
WORD FREQUENCIES

Word	Frequencies
Job	1754
Hire	1041
Latest	355
Open	282
Retail	257
Click	253
Great	251
Nurs	251
See	248
Fit	228

The Word-cloud of the most popular keywords in the Tweets is shown in figure 4.1

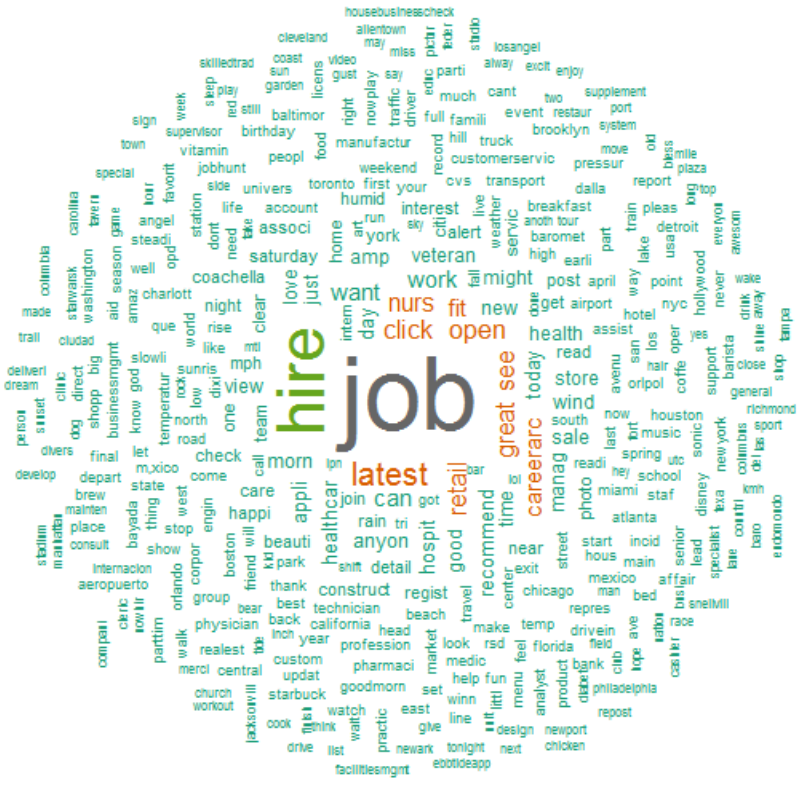

Fig 4.1 Word-Cloud of the most popular words in the Tweet Database

The most frequent terms in the corpus can be obtained from the Stop-word removal and Stemming operation on the corpus. In the dataset consisting of 5000 tweets that is considered in this paper, the most frequent words having frequency greater than 50 are shown in the figure 4.2.

```
> findFreqTerms(term.matrix1, lowfreq = 50)
 [1] "airport"     "alert"       "amp"          "anyone"      "apply"
 [6] "associate"   "avenue"      "barometer"    "beach"       "california"
[11] "can"         "care"        "careerarc"    "center"      "check"
[16] "city"        "cleared"     "click"        "coachella"   "construction"
[21] "day"         "details"     "fit"          "get"         "good"
[26] "great"       "happy"       "health"       "healthcare"  "high"
[31] "hiring"      "home"        "hospitality"  "humidity"    "interested"
[36] "job"         "jobs"        "join"         "just"        "last"
[41] "latest"      "like"        "love"         "manager"     "miami"
[46] "might"       "morning"     "mph"          "near"        "new"
[51] "night"       "now"         "nurse"        "nursing"     "one"
[56] "opening"     "park"        "photo"        "posted"      "rain"
[61] "read"        "ready"       "recommend"    "registered"  "retail"
[66] "sales"       "saturday"    "see"          "service"     "store"
[71] "street"      "sunrise"     "team"         "technician"  "time"
[76] "today"       "veterans"    "view"         "want"        "weather"
[81] "wind"        "work"        "york"         "youre"
```

Fig 4.2 Most Frequent Words having Frequency greater than 50 in 5000 tweets

The qualitative analysis of the top 100 keywords in the order of decreasing frequency in the corpus is depicted in figure 4.3.

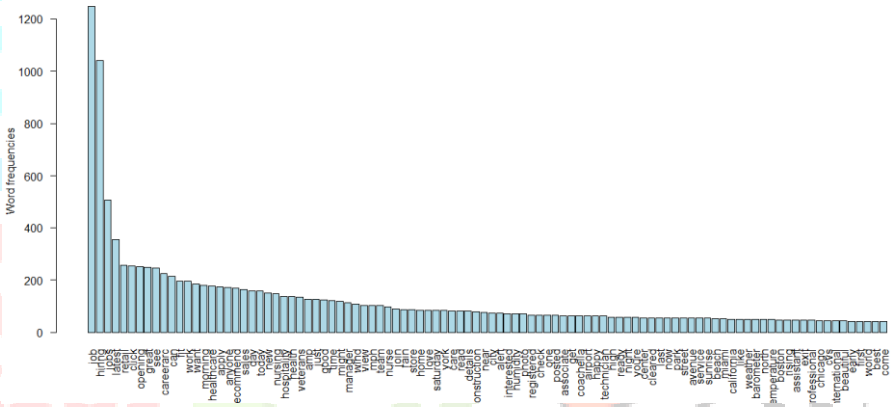

Fig 4.3 Frequency Magnitude of the top 100 frequent terms of the corpus.

4.2 Correlation between the Terms of the Corpus

The correlation of the terms related with the most frequent terms in the corpus is depicted in figure 4.4. It shows the terms most frequently used with the term "Job". The minimum correlation threshold is taken as 0.3 in the plot. However, a large number of terms are related to "Job" having correlation threshold lying in the range 0.9 to 0.1.

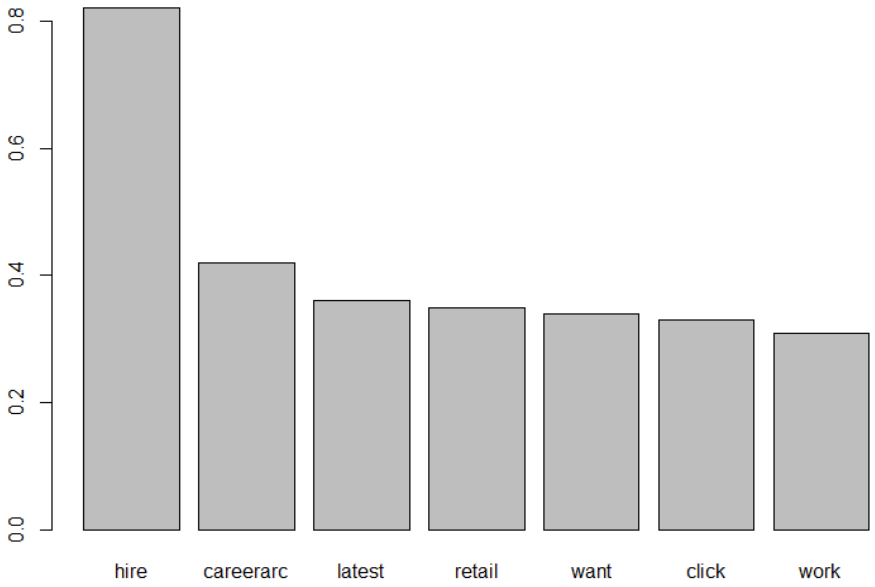

Fig 4.4 Correlation plot of the terms related with "Job"

Figure 4.5 shows the correlation plot of the terms related with the term “careerarc”. The minimum threshold considered in this graph plot is 0.15.

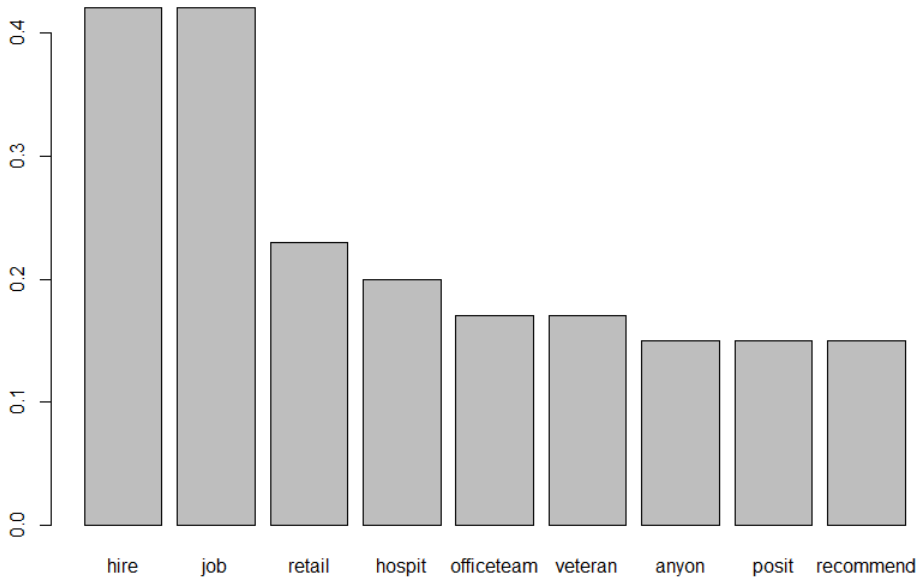

Fig 4.5 Terms related with the terms careerarc.

Figure 4.6 shows the terms correlated with the term “Hire”.

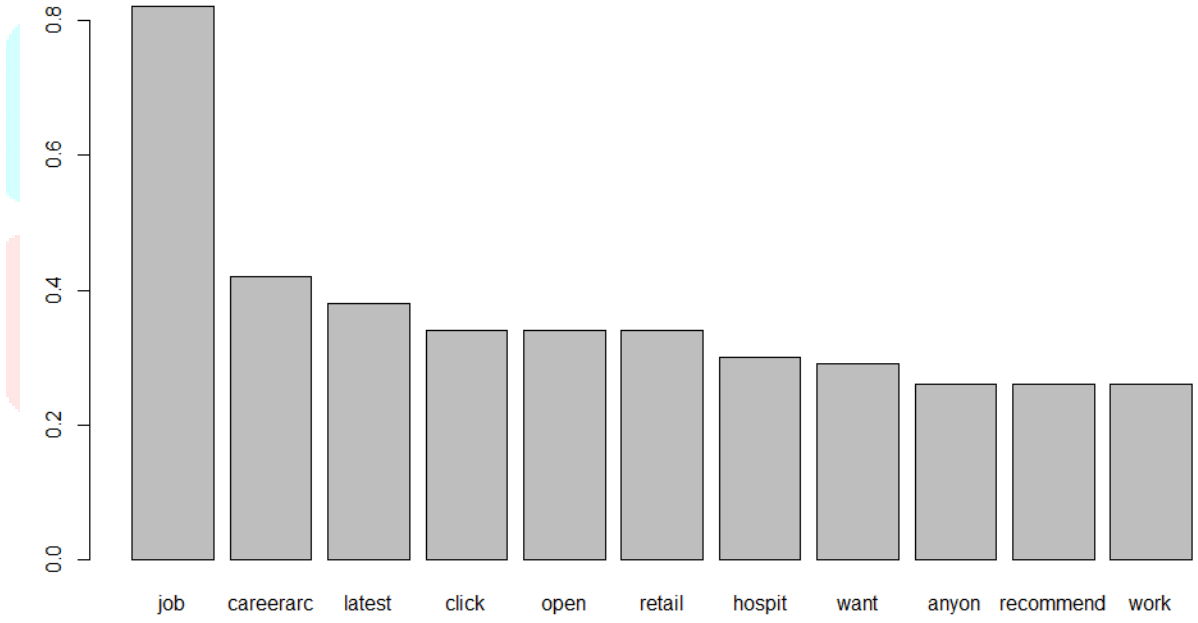

Fig 4.6 Terms Highly correlated with the Term “Hire”. Min Correlation value=2.5.

Figure 4.7 shows the terms correlated with the term “Latest”. The minimum correlation value considered in this graph is 0.3.

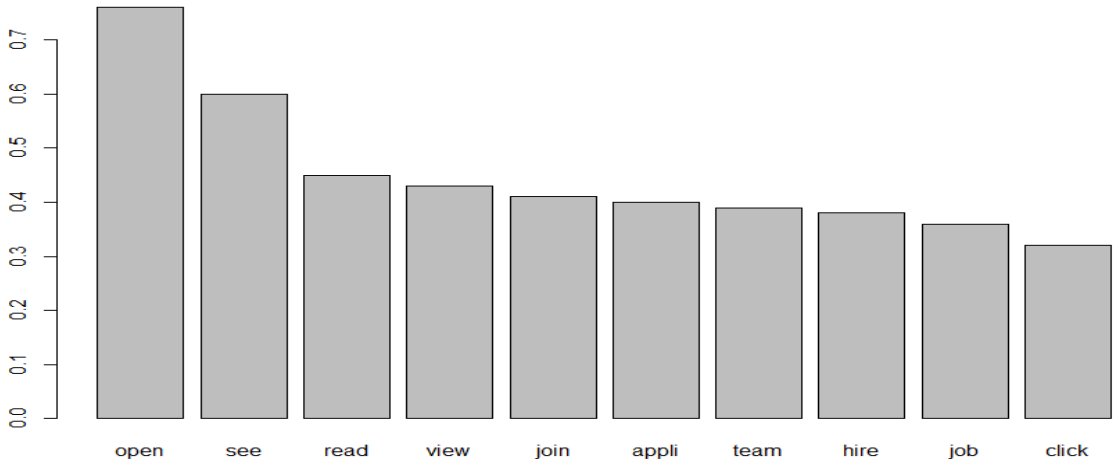

Fig 4.7 Bar Plot of the terms having high correlation with the term “Latest”

Figure 4.8 shows the terms correlated with the term "Retail". The minimum correlation value considered in this graph is 0.3.

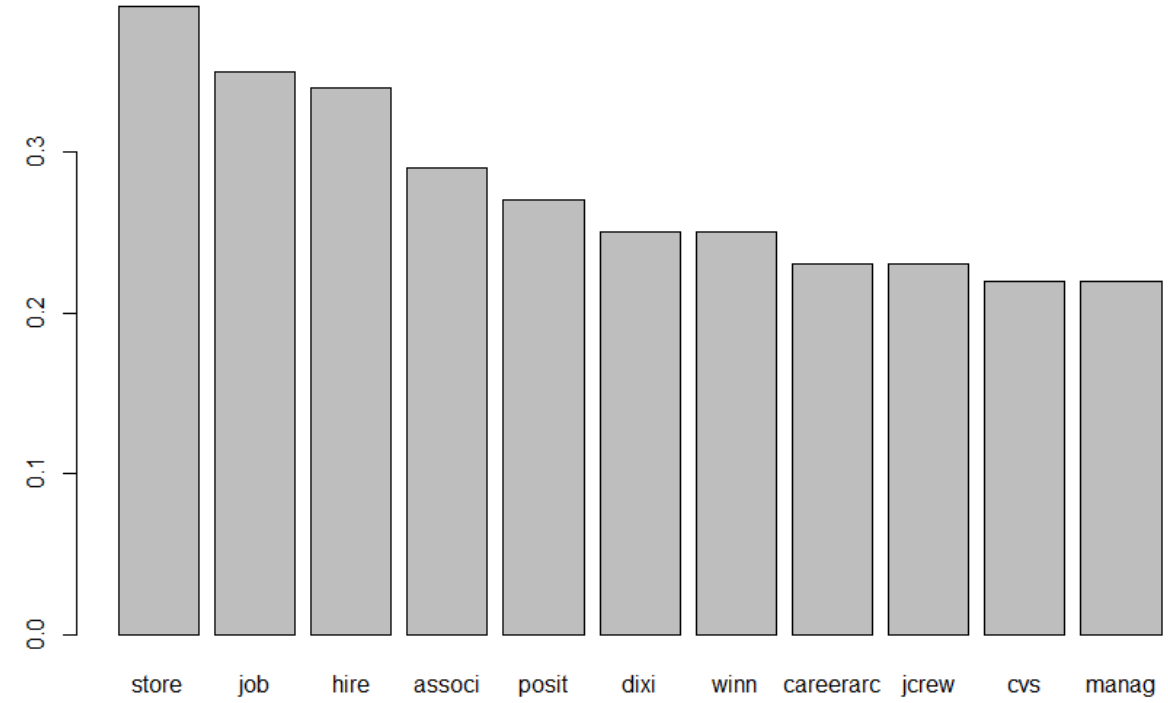

Fig 4.8 Bar Plot of the terms having high correlation with the term "Retail"

4.3 Topic Modelling

If it is considered that all the words in the Tweet Corpus arises from two topics, then the probabilities of the most frequent terms arising from the topics can be obtained from the LDA model in R. The following function can be used to generate the probabilities of the topic of the two topic model using Latent Dirichlet Allocation.

$$LDA - model \leftarrow LDA(Document - Term - Matrix, \quad K = \#topics)$$

The following results are deduced from the two topic modeling of corpus. Figure 4.9 gives the list of all the words that are supposed to be generated from Topic A in the decreasing order of probabilities. In the similar way, figure 4.10 gives the list of all those words which are supposed to be generated from Topic B in the order of decreasing probabilities.

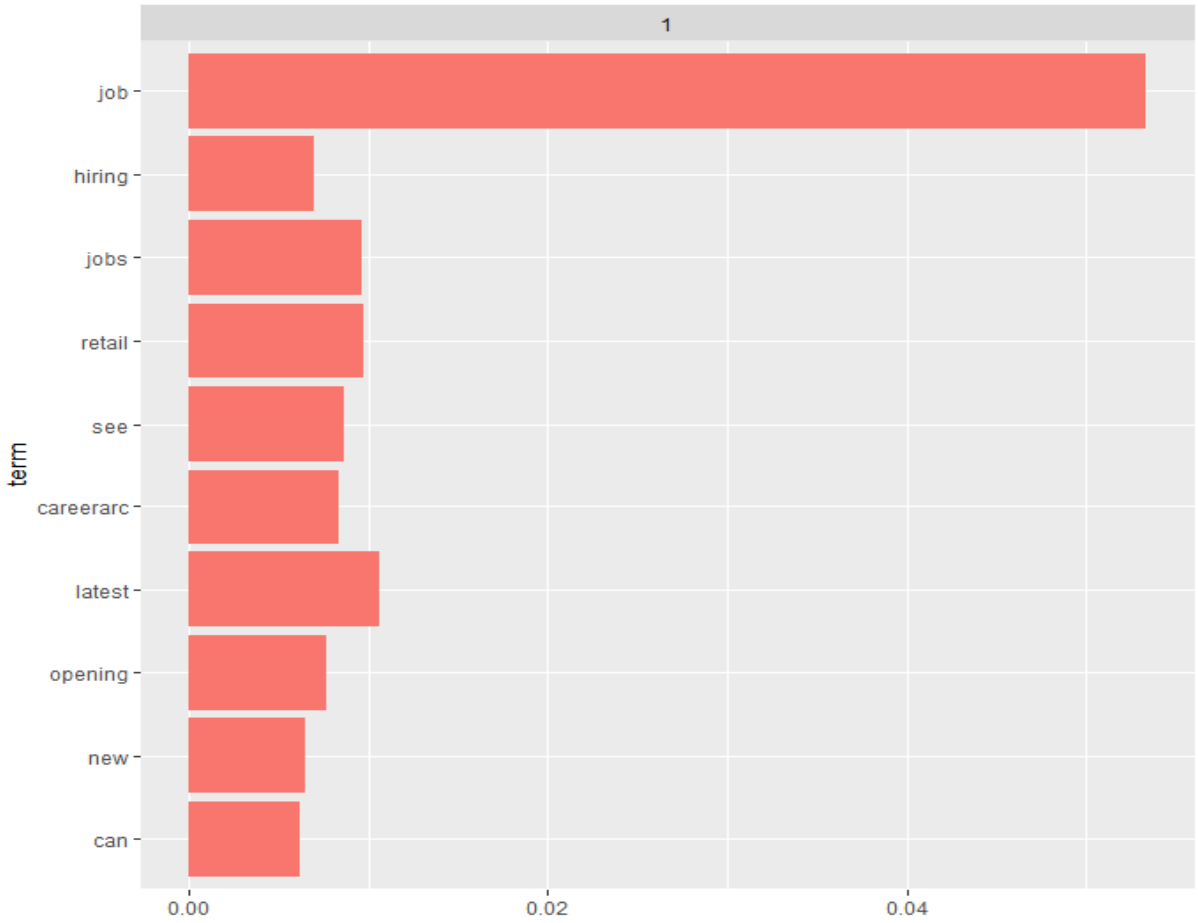

Fig 4.9 Terms Most Probable from Topic A

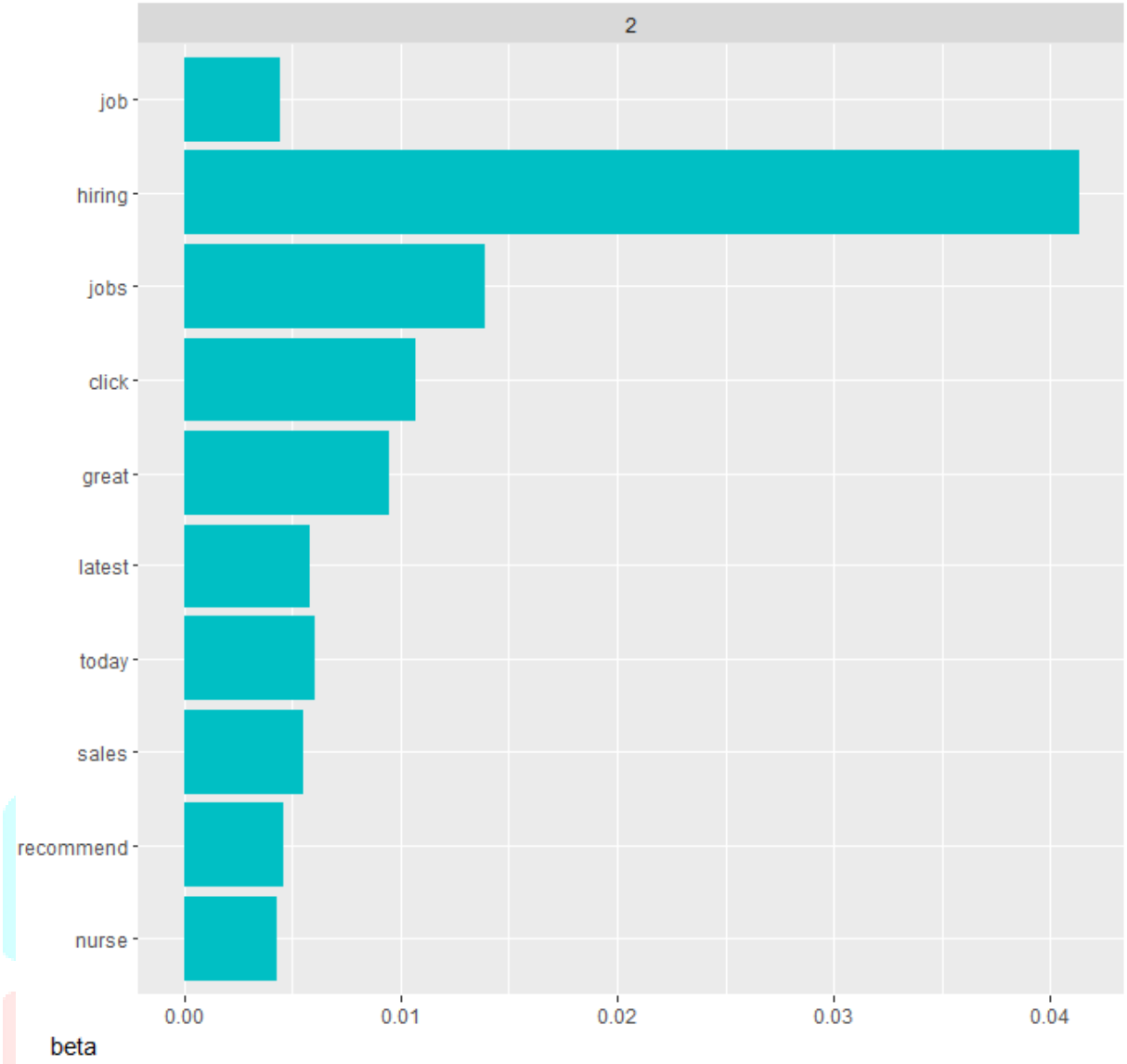

Fig 4.10 Terms Most Probable from Topic B

4.4 Clustering of the Tweet Data Set on the basis of the most Frequent Terms

The clustering of the Tweet Data set can be done using K-means cluttering algorithm on the basis of Euclidian Distance. The clustering cannot be visualized for a document consisting of 5000 records. It can be visualized by considering only a portion of the document term matrix. Considering only the first 20X30 element portion of the original document term matrix, the clustering using the k-means gives the clusters as shown in figure 4.11.

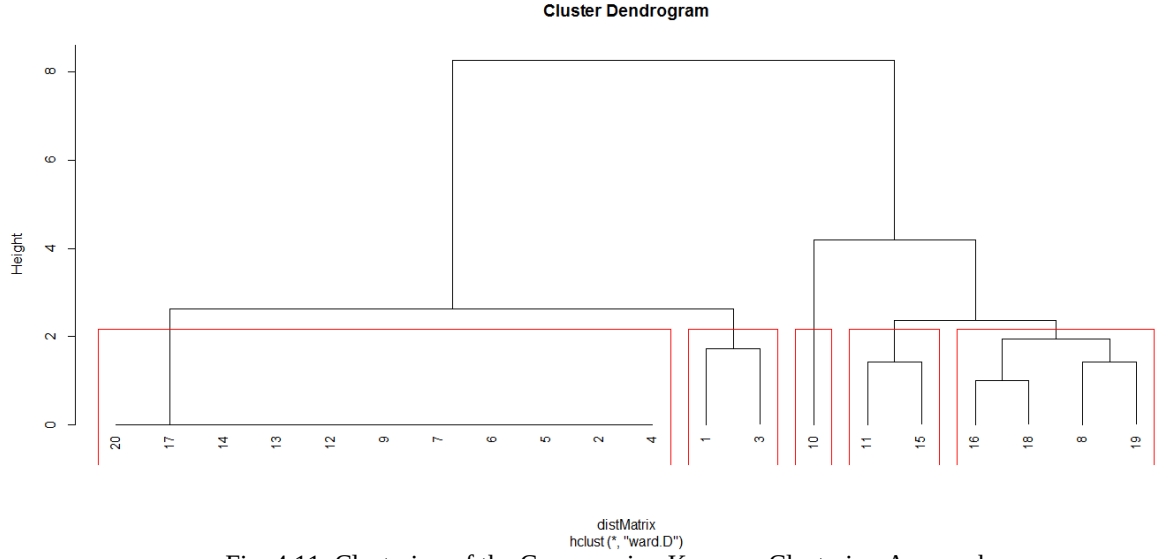

Fig. 4.11. Clustering of the Corpus using K means Clustering Approach

Figure 4.11 gives the clustering of a small portion of (20X30) of the document term matrix of the entire corpus. The entire subset of 5000 tweets is considered for the simulation model for support vector machine based classifier.

4.5 SVM Classifier for target Advertisement on the Tweet Dataset

Once clustering of Tweet dataset is performed, the clustering of the users can be tentatively made. The clustering of tweet dataset is mapped to the keywords describing the product or service. In this simulation model, we consider the target advertisement for "Job-Portal-X". The keywords related with the job portal are job, hire, salary, package, CTC etc. Considering the clusters of the keywords which are related with these terms though mapping, we can create positive and negative examples out of these keywords. The results of the classification for the first set of 100 records is tabulated as shown in table 4.4

TABLE 4.4
CLASSIFICATION FOR TARGET ADVERTISEMENT ON THE BASIS OF USERS AND TWEET CONTENTS

Tweet Id	User Name	Nickname	Classification Category for Job-Portal-X Advertisement	Classification Probability
7.21E+17	Bill Schulhoff	BillSchulhoff	Positive	0.297865
7.21E+17	Daniele Polis	Danipolis	Negative	0.111126
7.21E+17	Kasey Jacobs	KJacobs27	Positive	0.528665
7.21E+17	Stan Curtis	Stncurtis	Positive	0.146272
7.21E+17	Dave Borzymowski	wi_borzo	Negative	0.232341
7.21E+17	kirstin	KirstinMerrell	Positive	0.809384
7.21E+17	Joshua Kosches	Jkosches86	Negative	0.395611
7.21E+17	TMJ-PA Retail Jobs	tmj_pa_retail	Negative	0.365398
7.21E+17	Vidal Alexander	Vonfandango	Positive	0.955965
7.21E+17	TMJ-BAL Jobs	tmj_bal_jobs	Negative	0.292185
7.21E+17	TMJ-MD Retail Jobs	tmj_md_retail	Positive	0.847762
7.21E+17	gkalkat	gkalkat	Positive	0.404448
7.21E+17	Dr K Goodson	GodsNaturalDiva	Positive	0.063152
7.21E+17	Sarah Kehoe	sarahkehoe	Negative	0.333434
7.21E+17	TMJ- BOS Health Jobs	tmj_bos_health	Negative	0.544946
7.21E+17	John Hancock	JohnHancockJobs	Negative	0.309021
7.21E+17	raj	vaidyeah33	Positive	0.590169
7.21E+17	TMJ-DTW Engin. Jobs	tmj_dtw_eng	Negative	0.878708
7.21E+17	Atlanta IT Jobs	tmj_atl_it	Positive	0.732451
7.21E+17	N Master	N_Master101	Negative	0.582759
7.21E+17	Vicky Garcia Fitness	VickyGFitness	Negative	0.457995
7.21E+17	Andre C. Rivera	andrerivera801	Negative	0.612337
7.21E+17	Pola	PolaBonita	Negative	0.855536
7.21E+17	TMJ-VAF HRTA Jobs	tmj_VAF_HRTA	Negative	0.298287
7.21E+17	TMJ-NC Retail Jobs	tmj_nc_retail	Negative	0.340709
7.21E+17	Alan Anzo	AlanElGrande_1	Negative	0.174283
7.21E+17	GA Manufacturing	tmj_ga_manuf	Negative	0.151816
7.21E+17	New Jersey Sales	tmj_NJ_sales	Negative	0.306929
7.21E+17	Favorite Jobs	FavoriteJobs	Negative	0.442394
7.21E+17	Chris McGinn	SlimMcGinn	Negative	0.345125
7.21E+17	NJ Non-Metro Jobs	tmj_nj_usa_jobs	Positive	0.519288
7.21E+17	TMJ-ATL CstSrv Jobs	tmj_atl_cstsrv	Negative	0.834537
7.21E+17	Char Sips&Scroll	sugahNspyce_	Positive	0.643235
7.21E+17	Lisa Torre	LisaTorreFitnes	Positive	0.346986
7.21E+17	Doherty Jobs	DohertyJobs	Negative	0.816324
7.21E+17	DariusThaGreat	DariusDuhGreat	Positive	0.266366
7.21E+17	James K Barath, CMPS	jkbarath	Negative	0.613165
7.21E+17	Joe Ruffennach	firnatine633	Negative	0.513024

7.21E+17	BriannaMengel	legnemannairb	Negative	0.983952
7.21E+17	Chris S Jones	foggypaws	Positive	0.705607
7.21E+17	Patricia Suhre	SugaPiesMama	Positive	0.653903
7.21E+17	fulltimeGiGS Jobs	FTGiGSJobs	Negative	0.930805
7.21E+17	TMJ-MD HRTA Jobs	tmj_md_hrta	Positive	0.588404
7.21E+17	TMJ-SCM Health Jobs	tmj_scm_health	Negative	0.195106
7.21E+17	TMJ-NJN Pharm. Jobs	tmj_njn_pharm	Negative	0.840651
7.21E+17	TMJ-ATL Engin. Jobs	tmj_atl_eng	Negative	0.898182
7.21E+17	Jason W	jasoid	Positive	0.924361
7.21E+17	TMJ-ATL Cosmo Jobs	tmj_atl_cosmo	Negative	0.177759
7.21E+17	TMJ-MIA Retail Jobs	tmj_mia_retail	Positive	0.716565
7.21E+17	511 New York	511NY	Positive	0.626334
7.21E+17	Annie Islas	AnnieMyBoo	Negative	0.279613
7.21E+17	Ocelotemx1	Ocelotemx1	Positive	0.297301
7.21E+17	Plez	plezWorld	Negative	0.049401
7.21E+17	?KingAsshole?	biggtwon	Negative	0.939474
7.21E+17	511 New York	511NY	Negative	0.392035
7.21E+17	Bertha Balseca	BerthaBalseca	Positive	0.823649
7.21E+17	Dayton Freight Jobs	DaytonFrtJobs	Positive	0.702629
7.21E+17	Gaby Osornio ?	Gabinthehouse	Positive	0.306694
7.21E+17	Divine Appetit Co	divine_appetit	Negative	0.16181
7.21E+17	TMJ-GBR Pharm. Jobs	tmj_GBR_pharm	Negative	0.378602
7.21E+17	TMJ-NJC Acct. Jobs	tmj_njc_acct	Negative	0.54874
7.21E+17	TMJ-NYC CstSrv Jobs	tmj_nyc_cstsrv	Negative	0.743708
7.21E+17	TMJ-SC HRTA Jobs	tmj_sc_hrta	Positive	0.960829
7.21E+17	TMJ - MIA Sales Jobs	tmj_mia_sales	Positive	0.297763
7.21E+17	TMJ- WAS Health Jobs	tmj_dc_health	Negative	0.937786
7.21E+17	Lynne J. Johnson	MoJoCMO	Positive	0.034429
7.21E+17	Join BAYADA	JoinBAYADA	Negative	0.759068
7.21E+17	Carrie Cleveland	MsCarrieC	Negative	0.67834
7.21E+17	Matthew Gilmore	MattGilmore112	Positive	0.594443
7.21E+17	?Shaquana?	IndigenousSlay	Negative	0.772504
7.21E+17	'Sunkanmi	7unky	Positive	0.352259
7.21E+17	ADK310	ACEDAKIDD310	Negative	0.5987593
7.21E+17	Hannah Shields	hannahshields_	Negative	0.72091
7.21E+17	Christine Saunders	chrisandxbones	Positive	0.250988
7.21E+17	Heather Adams	HeatherAdams44	Positive	0.255863
7.21E+17	Nukdollar$	Therealnuk	Positive	0.004992
7.21E+17	TheKenyaCrooks	TheKenyaCrooks	Negative	0.385447
7.21E+17	Pablo	renteriaboxing	Positive	0.377281
7.21E+17	HCA	PracticeWithUs	Positive	0.851302
7.21E+17	Expera Jobs	ExperaJobs	Negative	0.707477
7.21E+17	TMJ-NY HRTA Jobs	tmj_ny_hrta	Negative	0.739356
7.21E+17	PA Non-Metro Jobs	tmj_pa_usa_jobs	Positive	0.003151
7.21E+17	TMJ-FL Transport.	tmj_FL_transp	Negative	0.637676
7.21E+17	Alejandro Ch ves	alejand92660840	Negative	0.619703
7.21E+17	Jobs at Oracle	JobsAtOracle	Negative	0.588015
7.21E+17	TMJ-MI Retail Jobs	tmj_mi_retail	Positive	0.48704

7.21E+17	Beej Powers	beejpowers	Negative	0.005708
7.21E+17	TMJ-MA Mgmt. Jobs	tmj_ma_mgmt	Negative	0.238315
7.21E+17	Keisha Frazier	feelgoodskin	Positive	0.653206
7.21E+17	?? Dewey Johnston ??	dewey_johnston	Negative	0.411388
7.21E+17	Jobs at Dressbarn	dressbarnjobs	Positive	0.20976
7.21E+17	Brad Hall	circabrad	Negative	0.527403
7.21E+17	Shane Barger	BargerShane	Positive	0.436844
7.21E+17	CJ Torres	cjtorres7	Negative	0.966489
7.21E+17	Mart¡n Nelson	forsomemasses	Negative	0.077789
7.21E+17	Cintas Careers	CintasCareers	Positive	0.319924
7.21E+17	Josh Spencer ??	JoshTrevil	Positive	0.301797

V. CONCLUSION

The proposed Scheme Clearly outperforms the one proposed by Siaw Ling Lo et. Al. [5]. This is because the classification for target advertisement is done on a much wider parameter based as compared to the base approach. It is indeed not possible to evaluate the outcome for a desired set of services over the same database. The target advertisement simulation consideration in this paper is evaluated for a hypothetical Job-Portal-X advertisement to the twitter users the outcome of which is shown in table 4.4. The parametric consideration between the current and the base work is depicted in table 4.5.

TABLE 4.5
COMPARISION OF THE PROPOSED AND THE BASE APPROACH

S. No.	Base Approach	Proposed Approach
1	Consideration of LDA Approach by Twitter	Consideration of LDA approach consisting of generic N topic modeling
2	--	Consideration of Generation of Keywords related with product/ service and mapping with those of tweet corpus
3	Clustering of Users	Clustering of Users
4	--	Clustering of Tweet Data Sets
5	SVM Classification on the basis of Clusters	SVM Classification on the basis of Clusters

It is evident that the proposed approach clearly outperforms that of the base approach in terms of the classification. This is because the classification of potential customers of product or service is more meticulously evaluated in the proposed approach as compared to the base approach. Section 5 gives an insight of the results and concludes the paper.

REFERENCES

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3 (March 2013), 993-1022.

[2] Liu Y, An A, Huang X. Boosting prediction accuracy on imbalanced datasets with SVM ensembles. Advances in Knowledge Discovery and Data Mining. Springer. 2006:107–118.

[3] Burges CJ. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. 1998; 2: 121–167

[4] Joachims T. Text categorization with support vector machines: Learning with many relevant features. Springer. 1998.

[5] Lewis DD, Yang Y, Rose TG, Li F. Rcv1: A new benchmark collection for text categorization research. J. Mach. Learn. Res. 2004; 5: 361–397.

[6] Zhao WX, Jiang J, Weng J, He J, Lim E-P, Yan Het al. Comparing twitter and traditional media using topic models. Advances in Information Retrieval. Springer. 2011: 338–349.

[7] Yang M-C, Rim H-C. Identifying interesting Twitter contents using topical analysis. Expert Syst. Appl. 2014; 41: 4330–4336.

[8] Lo SL, Cornforth D, Chiong R. Identifying the high-value social audience from Twitter through text-mining methods. Proceeding of the 18th Asia Pacific Symposium on Intelligent and Evolutionary Systems. Springer. 2015: 325–339.

[9] Lo SL, Cornforth D, Chiong R. Effects of training datasets on both the extreme learning machine and support vector machine for target audience identification on Twitter. Proceedings of the 5th International Conference on Extreme Learning Machines. Springer. 2015:417–434

[10] Predictive Analytics, Data Mining, Self-Service, Open Source—RapidMiner. Available: http:// rapidminer.com/. Accessed 30 April 2014.

[11] Willett P. The Porter stemming algorithm: Then and now. Program Electron. Libr. Inf. Syst. 2006; 40: 219–223.

Efron B. Bootstrap methods: Another look at the jackknife. Ann. Stat. 1979:1–26.

18. Breiman L. Bagging predictors. Mach. Learn. 1996; 24: 123–140