GTAI: Geo-centric Technique for Advertisement and Improvisation using Sentiment Analysis

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering
in
Software Engineering

Submitted By
Neetika Bansal
(Roll No. 801431015)

Under the supervision of:
Ms. Ashima Singh
Assistant Professor, CSED

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004

July

Certificate

I hereby certify that the work which is being presented in the thesis entitled, ***"GTAI: Geo-centric Technique for Advertisement and Improvisation using Sentiment Analysis"***, in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Software Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of ***Ms. Ashima Singh*** and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

NBansal

(Neetika)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Ms. Ashima Singh)
Assistant Professor
CSED, Thapar University
Patiala

Countersigned by

(Dr. Deepak Garg)
Head
Computer Science and Engineering Department
Thapar University
Patiala

(Dr. S. ~~S. Bhatia~~)
Dean (Academic Affairs)
Thapar University
Patiala

Acknowledgement

No volume of words is enough to express my gratitude towards my guide, **Ms. Ashima Singh,** Assistant Professor, Computer Science and Engineering Department, Thapar University, Patiala, who has been very concerned and has aided for all the materials essential for the preparation of this thesis report. She has helped me to explore this vast field in an organized manner and provided me with all the ideas on how to work towards a research oriented venture.

I am also thankful to **Dr. Deepak Garg**, Head of Computer Science and Engineering Department and, P.G. Coordinator, **Ms Rupali Bhardwaj** for the motivation and inspiration that triggered me for the thesis work.

I express my gratitude to all the staff members of Computer Science and Engineering Department for providing seminars and encouraging towards research work .I also thanks my colleagues who were always there in the need of the hour and provided with all the help and facilities, which I required, for the completion of my thesis work.

Most importantly, I would like to thank my parents, friends and the almighty for showing me the right direction out of the blue, to help me stay calm in the oddest of the times and keep moving even at times when there was no hope.

Neetika Bansal

(801431015)

Abstract

The exponentially exceeding internet based user content has been taken as a call by researchers, to mine this gigantic form of user information. This kind of information has been utilized for new inventions. Especially, Social Media(SM) specific information has exploded enormously in the present era of internet. The user specific information in the form of structured and unstructured data has helped in getting familiarized with likes, dislikes of a particular user community. Moreover, it has also helped in providing knowledge regarding the prevalent trends and fashion in a particular area. Sentiment Analysis (SA) as a tool has been come to a rescue in identifying user sentiments. It has also assisted in outlying and analyzing sentiments on the basis of their polarity (i.e. on positive, negative and neutral scale) for any topic which come under experimentation. Many previous studies have been conducted analysing user reviews on the basis of polarity identification and feature extraction. However, only few studies surfaced which dealt with Sentiment Analysis from the point of view of Advertisements(Ads). Also, limited work has been done from the context of analysing sentiments on the basis of geo-locations of particular user community.

The objective of the present thesis is to target spatially distributed web users worldwide by means of Geocentric Technique for Advertisement and Improvisation (GTAI) .This approach is indented to be used for marketing and promoting various products companies. Furthermore, it can be used to recommend different advertisement agencies about different products which can be promoted or advertised in particular geographical region. It is an effective geo-centric strategy, which works in four phases: Extracting user opinions from social media based upon their geo-locations, Pre-processing of textual user information, analysing the user sentiments using machine learning algorithm, and taking decisions by the help of rule base. The strategy presented in the paper is further demonstrated by using Facebook page. In the initial phase of experimental setup of GTAI, User reviews are collected from Facebook page and after analysing sentiments and exploring there geographical-regions using Top-k rule and 50-50% rule , then Face book page is further used as a test bed for validating the proposed technique on 3 parameters i.e. Audience Building, Consistency, Maintaining Goodwill.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviations	Description
SNS	Social Networking Service
SA	Sentiment Analysis
OP	Opinion Mining
ML	Machine Learning
EWOM	Electronic Word of Mouth
ADS	Advertisements
B2B	Business to Business
GTAI	Geo-centric Technique for Advertisements and Improvisations
AEC	Advance Emoticon Classifier
NB	Naive Bayes

Chapter 1
Introduction

In this Chapter, we have studied the role of social media and its applications in various fields. Two main Social Networking Sites, Twitter and Facebook has been discussed based upon its popularity among people. Moreover, introduction to Sentiment Analysis(SA) and the generic process illustrating the main steps of SA process are also illustrated. Also, main types of Advertisements(Ads) are discussed. Last section of this chapter outlines the organization of the thesis.

1.1 Role of Social Media

For the past decades, social media have outstand in all the directions whether it is (i)the growth in its importance among the whole user society; (ii)whether it the usage of various networking sites exiting now; or (iii)whether it is popularity of Social media and many more. In spite of Social Networking Service (SNS) being an influential communication tool [1], it also serves prominently in socializing various communications that takes place among the corporate sector [2]. SNS has also played a very crucial role in helping many Business to Business (B2B) companies and product firms in targeting large group of geographically separated audience easily by the help of web advertisements. Moreover, SNS have generated huge volumes of spatial footprints and proved to be valuable corpus for knowledge dissemination. They are the large repositories of textual information, encompassing variety of user views on different aspects and events. Moreover, these networking sites are frequently used as an opinion collector for wide range of subjects that occurs worldwide. This platform is being exploited by experts of several domains like (social sciences, sciences, psychology, education, web analytics etc.) irrespective of their age groups and cultural differences . SNS also helps various business to business companies and product companies in addressing geographically distributed large group of audience easily.

Fig. 1.1 shows the information provided by dazeinfo.com[3] on basis of popularity of SNS among different business companies on the basis of most adopted and most used scale. From the diagram we can easily predict that Twitter and Facebook forms the majority group and outstands among all the other leading networking sites from the context of most used and most adopted.

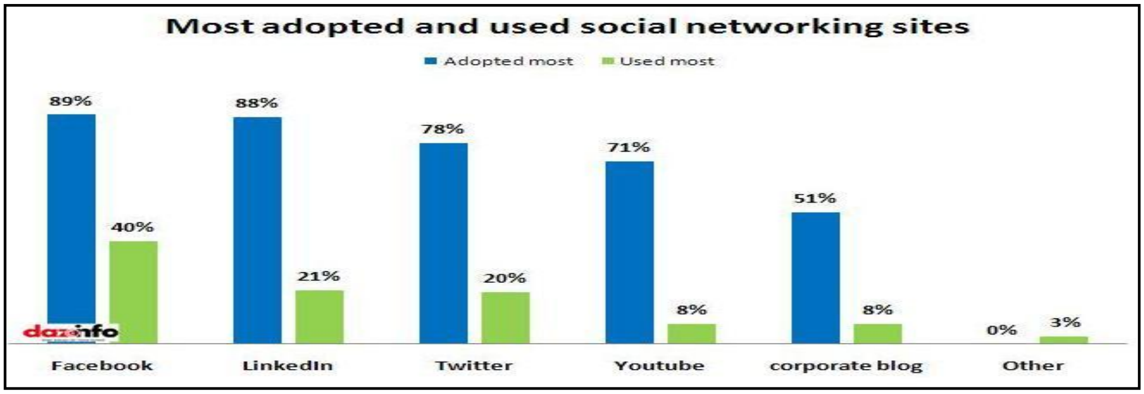


Fig.1.1:Distribution of SNS most adopted and most used by b2b. Source(dazeinfo.com)

Furthermore, user frequency is exceeding constantly for the last few years and thus SNS act as an opinion collector for wide range of subjects that occurs worldwide. Facebook Newsroom reported on February 2016[4] states that daily Facebook boasts approximately 900 million users(across the whole globe) on average basis which accounts to a subtotal of 1.2 billion Monthly Users and this stats contain approximately as monthly active users .

1.1.1 Micro-blogging Sites

Every SNS that provide users with a platform to share their feeling, emotions regarding different issues concerning both personal and social matters in form of short and recurrent posts is called as micro blogging sites. The Twitter is one of emerging micro blogging rostrum with huge repository of diverse forms of users amalgamated together as a single unit. It acts as vehicle to communicate one's opinions using messages known as twitter feeds. Theses opinions in the form of feedbacks or reviews proves to be fruitful in real world scenarios, such as product/service reviews related to movies, restaurants, hotels etc [5]. Gokulakrishna et al. described twitter as a subscribeand-publish network, providing point to point links among users to communicate information about various subjects[6]. Twitter users are exploding these days at enormous rates as pointed out by Lunden[7] , stating that twitter's active users count have reached to account 517 million accounts (as on July 1, 2014). it has been ranked as the second-biggest social networking site after Facebook. These numbers clearly define the significance and impact of social media on today's online-population. In this thesis, at the preliminary stage we have used Twitter feeds to form new classifier and after that we have devised a technique for recommending ads to different product companies/advertisement agencies by extracting users reviews form Facebook.

1.2 Sentiment Analysis

The Internet act as rostrum for web users, that is present around the globe in order to share their experiences via the means of electronic word-of-mouth (eWOM) communication media. Sentiment Analysis or Opinion Mining (OM) is the study of computationally figuring out and categorizing people's attitudes, behaviour and opinions expressed towards an entity in form of textual data. The two techniques SA and OM complement each other in meaning and are replaceable in their work. However, a slight difference was noticed by some researchers based upon different notions which are depicted in the article [8].

1.2.1 Sentiment Analysis Process

Khan presented a generic process of SA [9] as depicted in Fig. 1.2 . In this process a sentiment engine acquire feedback (data) from various internet modes like social media, emails, text messages etc .After the acquiring the required information,

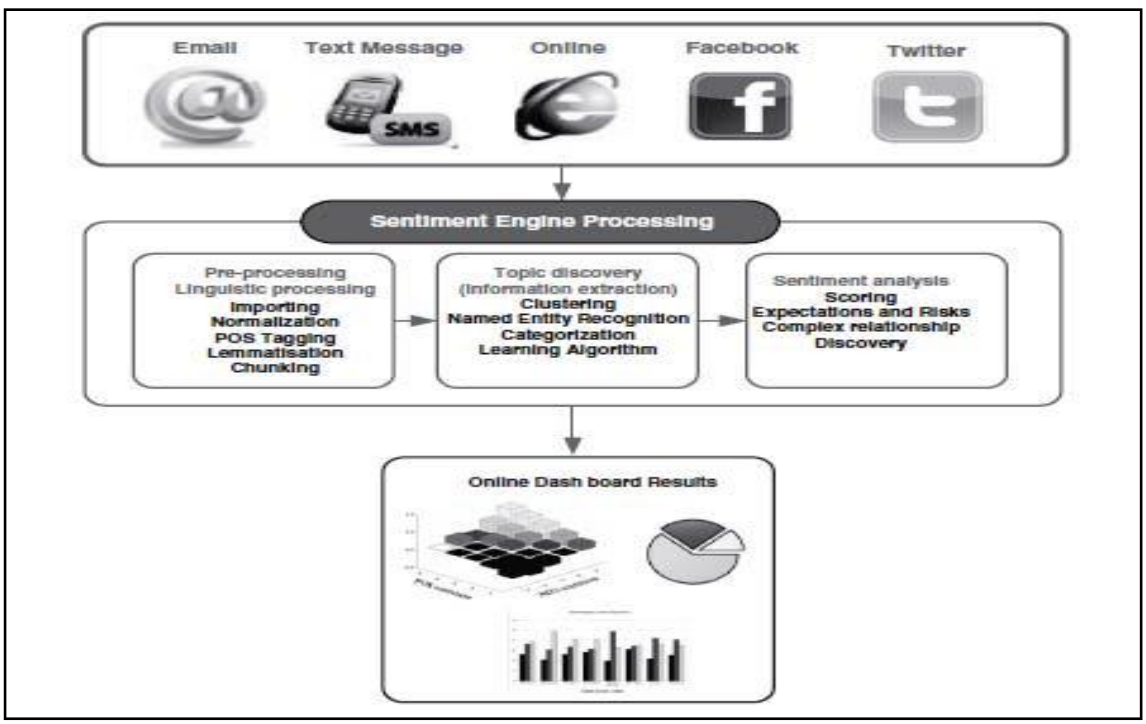


Fig. 1.2: Generic Sentimental Analysis Process

a unique algorithm/technique is applied, which further used to calculates the polarity (on positive/negative scale) of data by assigning relevant weights to it. The results can be used to draw various types of graphs for analyzing and visualizing outcomes. These outcomes can be further used to find out the overall feelings of a user towards a particular person, product or service.

1.2.2 Main fields of SA

The fields which form the major research domain of SA are: sentiment classification, feature extraction and classification and opinion Summarization. Sentiment classification (SC) is an approach in which entire document is classified towards certain entity. SC is the process of finding out the polarity of feature extracted data by first going to through the pre-processing and sentimental identification phases [10]. Feature Extraction and SC uses the opinions on different features of certain entities. Opinion summarization is significantly different than the traditional approach because in the former approach only the relevant features on which customers gave their opinions are only considered.

1.3 Online Advertisements

The subject of Web advertising has gained much academic attention and has been under public spotlight for past several decenniums. web ads are popular form then the traditional form of the Ads. The main reason for this big change is since the former type helps us to reach large group of audience. It also helps in saving the time as in the latter case we need to target individual or specific group. Moreover, online ads help in marketing the particular brand, service or product in larger context unlike the case of traditional ads.

1.3.1 Ad Definition

Ad is an important marketing and communication tool which aid various b2b companies, and product firms in promoting particular business service or specific products to diverse group of audience which are spatially distributed around the globe. Ads decision making is consider very important as it directly affects the sales and profit of the company. Fig. 1.3 shows the mechanism how the ad agency helps in targeting the users.

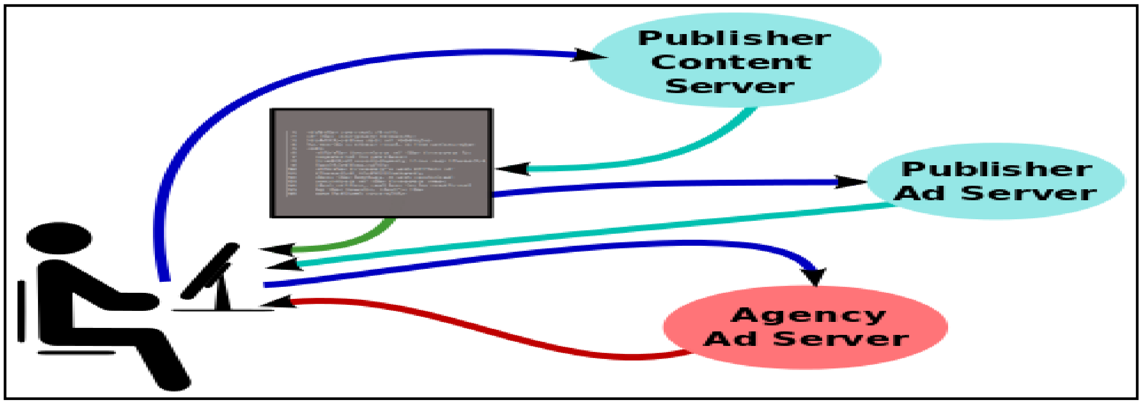


Fig. 1.3: Online advertising serving process using an ad agency

1.3.2 Types of display ads

Electronic ads is generally classified as: 1) target based advertising, which send ads to a receiver based upon their preference profiles; and 2) social media advertising, which distribute advertisements to SNS users by examining their social relationships. Target based advertising can be consider as an important application in area of recommender system and Sentiment Analysis. Fig. 1.4 shows the various ways of targeting audience based upon advertisements[11]. This approach generally works on the principle of exploiting two main approaches which are (i) content-based approach, and (ii) collaborative-based approach.

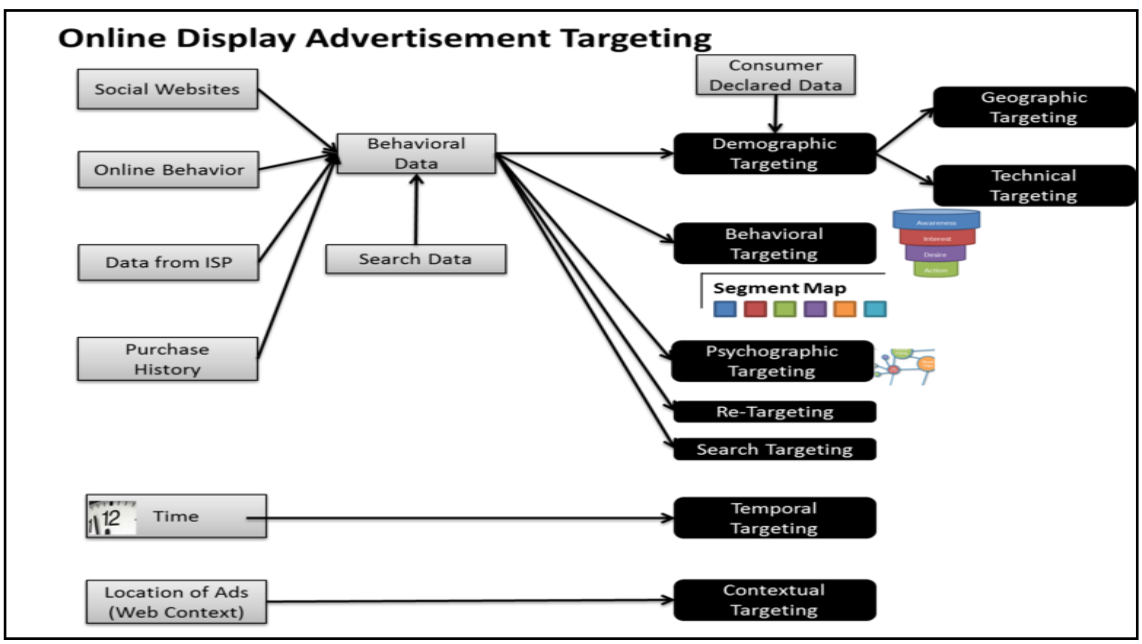


Fig.1.4: Target Based Ads : Source(Wikipedia.com)

Both the approaches are mainly use to find out the personal interests/preferences of particular group of users. The content-based approach is based on the earlier profiles of users according to their preferences, while the collaborative-based is based upon exploiting the common interests that prevalent in particular user community considering the similar kind of user profiles. Both, the discussed approaches depends completely on user ratings, which generally creates obstacle in recommending latest stuff to users when there exists no association among user reviews or rating records [12]**.**

In comparison to conventional advertising, social advertising is latest form of advertisement which generally target people using advertisements as part of a social network and uses social relations and social influences between people to sell their products or services. In other sense, social advertising uses an indirect way, such as the Electronic-Word-of-Mouth(eWOM) process or an endorsement process, to disseminate advertisements. An endorser is generally standalone entity which can be any individual or firm with recognizable public status and can represent any product/service [13]. Exploiting eWOM can prove to be fruitful for large companies to enhance their marketing drives and helps in escalating sales and profits for their products and sales. Exploiting social relationships and knowing the value of social interaction can prove to be trusted sources for social advertising. In this thesis, considering the factors like geographical locations of users, users sentiments, we have propose a geocentric technique for advertising based which works on the principle of target-based approach and user of social advertisements.

1.4 Organization of the Thesis

The rest of the thesis is organized in the following order:

- **Chapter 2 Literature Review -** In this chapter, we have done in depth study of SA domain by considering the area from different direction. Various facets of SA like Machine learning approaches used, granularity, various data sources present, polarity and studied and mapped with the related work. Also the application of SA field are discussed.
- **Chapter 3 Problem Statement –** In this chapter, we have discussed the prevent research gaps based upon this we have formulated the problem statement. Further,

the objectives of the problem statement and the methodology used to solve have been discussed.

- **Chapter 4 Design of GTAI framework –** In this Chapter, as preliminary study we have discussed existing techniques available for Micro-blogging sites and have then we have proposed a new technique by integrating different Classifiers. The approach is further used for the proposal of GTAI framework.
- **Chapter 5 Case Study -** This chapter includes the experimentation setup used to implement the GTAI and then the analysis of the results is carried out.
- **Chapter 6 Conclusion and Future work–** In this chapter brief conclusion, summary of contributions , limitations of the study and future scope of the study.

Summarizing, we have studied the various SNS and its application. We have given a brief overview of what exactly SA is; the generic steps involved in order to carry out the smooth SA process and areas which forms the major part of SA. Various available approaches have been studied to target audience by the help of ads.

Chapter 2

Literature Review

In this chapter we have presented a detailed study of SA, by considering the domain from various perspectives. Various Machine learning approaches, Categorization of Data Sources, Granularity and specific applications have been highlighted.

2.1 Sentiment Analysis Review

Sentiment Analysis (SA) or Opinion Mining (OM) is the study of computationally figuring out and categorizing people's attitudes, behaviour and opinions expressed towards an entity in form of textual data. In this section we have mapped the different items which forms the core of SA Domain to the various articles. This work has been depicted in the form of matrix where columns represent sentiment analysis studies and rows stand for various dimensions. Fig. 2.1 gives matrix formulation of data in order to get better visualisation and good understanding of the topic.

The matrix is classified in 5 broad factors which are approaches, type of data used, algorithms used, polarity, granularity of analysis. Some of the categories are further divided into sub categories for the sake of better understanding and better insights. Factor 1 i.e. Approaches is divided into machine learning, and sub divided into supervised and unsupervised learning approaches , Factor 2 i.e Type of Data Source is divided into reviews, online sources, and social media , Reviews are further divided into product reviews, movie reviews ,restaurant reviews and others reviews ,Social media partition is done in the form of Twitter, Facebook. Online sources are divided into blogs, news, forms and data sets, Factor 3 i.e. algorithms used to carry out each study. Factor 4 i.e. level of analysis is classified into 5 sublevels which are document level, sentence level , aspect level, word and concept level. Factor 5 i.e. stands for type of polarity which is sub divided as binary polarity which stands for +ve/-ve or 1/-1, polynomial polarity, general analysis which contain the articles not relevant from the point of polarity identification.

For the sake of convenience, we have used some abbreviations which are SP,US,MR, PR,RR, FB,BL,FR,DS,NE and these stand for Supervised learning, Unsupervised learning ,movie review ,product review, restaurant review, twitter, face book blogs, forums, data sets and news respectively. Where ever the article does not corresponds to any category, we have created a new category O which means other and the article is mapped with this new category devised.

DIMENSIONS			STUDIES WITH REFERENCES																												
			S.Wang[14]	J. Zhu[15]	L. Jiang[16]	C.C. Chen[17]	R. Xia[18]	X. Bai[19]	Z. Zhang[20]	M.R. Saleh[21]	A. Abbasi[22]	S.K. Li[23]	A. Reyes[24]	A. Balahur[25]	H. Kang[26]	P. Racherla[27]	A.munidas[28]	A.C.-R. Tsai[29]	Teddy[30]	E. Cambria[31]	G. Wang[32]	L. Qiu[33]	F.H. Khan[34]	Y.M.Li[35]	D. Spina[36]	Z.-H. Deng[37]	S. Krishnamoorthy[38]	H. Xu[39]	B.Chetashri[40]	B Agarwal[41]	K Gao[42]
Approaches	1.1 Machine learning	SP	*		*	*	*	*	*	*	*	*	*	*	*	*		*	*		*	*	*	*	*		*	*	*	*	*
		US		*																*						*					
Categorisation of Data Sources	2.1 Review	MR						*			*					*			*			*	*	*	*	*	*				
		PR				*	*			*	*	*	*								*					*		*	*		
		RR		*					*						*																
		O	*								*																				
	2.2 Social media	TW			*																		*		*						
		FB															*														
		O																													
	2.3 Other Online sources	BL																													*
		FR																			*										
		DS																													
		NE						*																							
		0	*											*				*		*										*	
Algorithms Used	Support Vector machine		*		*	*	*		*	*	*		*	*	*			*	*		*			*	*	*	*	*	*	*	
	Maximum Entropy			*			*														*										
	Naïve Baiye's						*	*	*				*		*							*		*							
	Neural network																*			*											
	Dictionary Tree												*			*					*										
	Regrression											*							*												
	Association Rule																														*
Granualarity	Document				*	*	*	*				*			*	*			*		*	*	*	*	*		*				
	Sentence								*							*															
	Aspect		*	*													*											*	*	*	*
	Word									*	*		*													*					
	Concept													*				*		*											

Fig. 2.1: Various dimensions of SA mapped with related work

2.1.1 Sentiment Analysis Approaches using Machine Learning

Machine Learning(ML) works on mechanism of linguistic features selections and text classification techniques. ML approach has been used in different documents. Various approaches are used in this field which are broadly classified as supervised techniques and unsupervised techniques as depicted in Fig. 2.2. In the former technique, the problem solution is carried out by using two types of documents. One of them is test data set on which validation is done and other one is the training set which acts as a classifier.

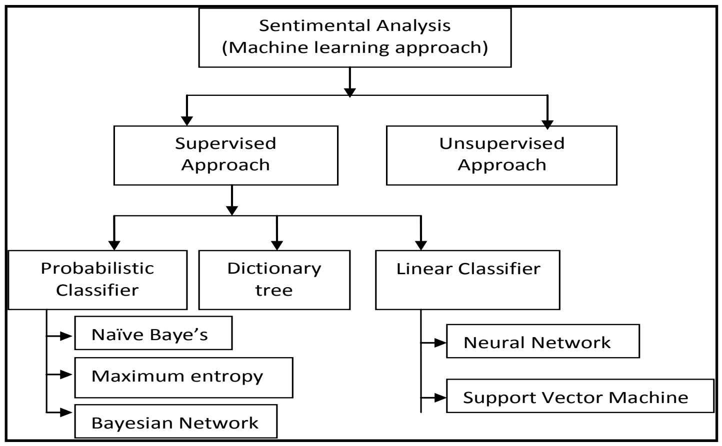


Fig. 2.2: Machine Learning techniques for Sentiment Analysis

Supervised approach can be further divided into three main types of classifiers probability , linear, and decision tree classifier. Among the probability classifier the main Naive bayes(NB) is not only simple algorithm and easy to use classifier but also a reliable algorithm. Its paramount use is in document level classification. The approach works broadly on BOW feature extraction and uses famous bayes theorem to find the probability of particular's feature's presence in a set. Kang and Yoo proposed an improved version of the approach[26] , in this approach positive sentiment accuracy can be increased by 10% to that of negative sentiment. The Maximum Entropy classifier works on the principle, the probability that a entity belongs to a class according to some context must maximize the entropy of the whole classification system. The main principle of Support Vector machines is to decide linear separators in the investigation space which can best split the diverse classes. M. Ott introduces a negative deceptive dataset and implemented spam opinion classification using SVM [43]. Experiments were carried out to find the effect of size and representations of a feature on the classification performance by Zhang et al. The classified the sentiments using NB and

SVM by using restaurant reviews. The main purpose of the process carried out using Decision tree is to provide a hierarchical disintegration of the training data set in where a condition is set up on the value of the attribute which is in turn used as a parameter to divide data. The predicate or condition is the absence or presence words. created Twitter based Lexicon using supervised reduction of features using n-grams and statistical analysis [27].The text classification is used identify and categorize documents predefined number of sub categories. In order to perform the above task, large quantity of training set of labelled documents are required in case of supervised learning, as illustrated before. In text classification, it is sometimes become complicated to make these documents, however it is easy to collect the unlabeled documents. For this purpose Unsupervised learning approaches are used which overcome this difficulty.

2.1.2 Granularity

SA is generally carried out at Document, Aspect and Sentence level as stated in document [10]. In the document level the whole document containing information to analyze is treated as one single unit. The purpose of this level is to classify the whole document as single entity expressing a positive or negative sentiment. Unlike in document level, in this granule the classification is carried out at sentence point, each one of the sentence is consider as a standalone entity and analysis is performed on it and later the results are obtained and are summarized in order to extract the precise sentiment of the document . Through research it was found out that a little difference lies between document-SA and sentence-SA, since sentence can be consider as a short form of document [5].However the outcome acquired by means of above levels seems to be inappropriate at times since they do not provide the required information regarding the entity from aspect point of observation. Thus there is a need for aspect level classification. Some Researchers have also carried work at finer granules like Word level which can be considered as the finer granule of sentence level .Concept-level is another finer grain which focus on a semantic analysis of text through the use of semantic networks, it helps in the aggregation of affective and concrete information correlated with natural language opinions.

2.1.3 Data Sources

Opinions form the basic and major factor for improving the quality of deliverables and prove to be the stepping stone in the enhancements of the services. Many different sources like Blogs, products, movies reviews sites, social media and other data sets which are available online for research purpose makes the good source of information**.** In the recent times with advancement in technology, internet usage has seen a sharp growing trend, leading to large number of online source of data. Some of them are blogs, discussions, forums etc. Most of these blogs are created to express one's opinions; reactions used for personal purpose or to express their interests for products etc. Blogs are used as a foundation source of data for various studies as shown in the matrix. Blogs plays important role in advertisements

J.H. Wang used Chinese blog posts for phrase extraction[44]. Some of the other significantly used sources are forums, news comments, discussions. Social media is package of huge amount of unstructured data in form of text ,animations, videos, posts which can be used to express one's experiences, opinions etc and can form the source for public awareness..Among the top most social media sites Twitter is a feeling, emotions. D. Spina used keyword based selection of tweets for finding the disambiguation in company names[36]. Twitter was source for many other studies as can be refer from the matrix M. Link mining technique with user centric approach to extract user opinions through social networks was carried out by Rabelo et al.[45] .Other sources which are part of this classification are you tube comments, Google scholar ,facebook posts etc**.** Online data set play a Key role in order to carry out smooth research since data acquisition is first step of sentimental analysis process. The major study is done on movie reviews, one of the movie dataset available online is(http:// www.cs.cornell.edu/People/pabo/moviereview-data). A variant-domain sentiment dataset. (http:// www.cs.jhu.edu/mdredze/datasets/sentiment) which contain different set of product reviews which are gathered from Amazon site. The different products include Books, DVDs, clothes etc. Reviews have drawn the attention of many researchers to carry out the work. Movies, product, restaurant form the source of data in recent research studies.

2.1.4 Applications

- Online advertising- Online Advertisement has become one of the largest sources of incomes in today's internet environment. Whether it is to promote your business or highlight about new invention all rely on this mode of this communication expression . More to target advertisements to large user population has always been a trying tasks for the industry people. Y.M. Li proposed a diffusion technique for advertisement purpose [35].
- Prediction Systems- Prediction is to forecast or calculate a value beforehand for various purposes whether it predict sales, movie review prediction, weather forecasting etc . Different researchers have used SA for prediction purpose. Based on opinions formed by the people about Companies, resources and services one can predict the trend of the stock price for purpose of market. Du et al. [33] carried out box office prediction based upon two features: count and content using micro blogs. The content features tag the comments into 3 categories: beneficial, harmful and neutral whereas Count based features considers several factors like quantity and quality of users, influence of expert users to other users, and time based factors.
- Competitive Analysis-Risks are crucial factor evolution of business, so it is important for companies to collect and analyze information about their competitors' products and plans. Sentiment analysis find a major role in competitive intelligence to extract and compare relations between products from customer reviews with the interdependencies among relations taken into consideration.
- Other Applications-Some of other applications like Opinions in the social and geopolitical context, Customer preference are discussed in article. S. Wang illustrated that SC be used for text filtering, online tracking opinions, analysis of public opinion poll, and chat systems[14].

Summarizing, we have studied different facets of SA like ML approaches, data type used, granularity, polarity, algorithms etc. Also the studied facets are mapped with their related work in matrix form. Also the various applications of the SA are discussed at the end of the chapter.

Chapter 3
Problem Statement

The chapter broadly concentrates on finding the gaps that exists in the literature and then analyzing them in order to formulate a problem which will act as stepping stone for the thesis . Also the key objectives of the thesis have been outlined.

3.1 Gap Analysis

Sentiment Analysis has drawn the attention of various researchers for past few decades and has continued to attract the attention of many people. SA field has always been a challenging research sphere. Previous research has mainly concentrated on improving the accuracy of exiting algorithms and developing a new and advance algorithm by integrating Machine Learning classifiers. Much work has been done on polarity identification and feature extraction of product reviews. And very less has been concluded on how to apply SA to SNS like Twitter, Facebook regarding Online Advertisements on social media and Web Advertisements of different products. Even online advertisements have not targeted the geographical areas while marketing on web for different products in order to increase sales. Much of the Advertisements that are deployed using collaborative target based approach, considering the previous user search history. However, few studies use Sentiment Analysis as an Application for the Advertisements. Moreover, the idea of accruing user sentiments from different geo-locations have been part of handful amount of studies which need to be extended in order to find out the opinions of the particular user community.

3.2 Problem Statement

Advertisements is a good marketing tool which helps product companies in targeting large group of diverse audience. It has also been a challenging task for marketers to take important decisions regarding 1. which geo-location should be targeted by the means of Ads ; 2. which geo-locations need to spanned more for maximizing Ad campaigns. Social advertisements is latest form of ads which need to looked upon and more work need to be carried out. Moreover the understanding of textual information retrieved form micro-blogging sites is also a challenging task as the text contains different syntactic symbols which need to be taken proper consideration.

3.3 Objectives

The objectives of the present thesis are as stated below:

- To study existing tools and techniques of sentiments analysis and role of online advertisement on social media.
- To compare tools and techniques available for social media, micro-blogging sites with respect to advertisements.
- To propose a technique for sentiment detection for micro-blogging sites by integrating domain independent and domain dependent classifiers.
- To propose an effective technique for advertising products on social media targeting large group of geo separated users.
- To test and validate the proposed technique.

Chapter 4

Design of GTAI Framework

In the preliminary stage of our work we have devise a technique for micro-blogging site by integrating domain independent and domain dependent Classifiers , which laid the foundation for the next work . In latter part of the chapter we have proposed a technique GTAI- geo-centric technique for advertisements and Improvisations. The key components of the proposed technique are also discussed in detail.

4.1 Existing Techniques to Analyze Twitter Feeds

Various methods and strategies was employed by researchers to monitor twitter data for occurrences of various events, stories and reactions of various users Pang et al explored few supervised machine learning methods to classify different reviews extracted from twitter [46].

Cui et al. demonstrated the reason why that sentiment analysis of twitter feeds is a difficult task[47]. The paper uses emotion tokens to tackle with the problem. The propose work is carried out in 3 parts. First, extraction of emotion token; Second, different polarities are plotted using graph propagation algorithm; lastly, sentiment detection techniques are used to analyze and classifies them .The outcomes of the approach depicts that emotion tokens are a good approach for semantic analysis provided lexicons are created which are independent of time spans. The limitation of this approach was that less overall accuracy of system was obtained.

Bahdur gave a SA mechanism for data which is collected from Twitter in form of tweets[48]. In this technique Training models are produce with the help of proposed method and the outcomes displayed fine results of classification performance. Minimal linguistic processing was applied in order to support multilingual datasets. Tweets also contains a lot of garbage words in form slangs and abbreviations, this issue was not taken by the paper. Moreover the strategy does not provide comparison of accuracy obtained with other techniques.

Ortega et al. proposed a technique with three phases; pre-processing, polarity identification and classification[49]. SentiWordNet and WordNet approach was used to find out the polarity and rule based method was also exploited..

B and Feng presented a technique by collecting syntax features of tweets (links , hashtags and punctuations) with combination features extracted using prior polarity and POS of words[50]. They used aggregation of data collected from three websites and used them to train their model and also used 1000 physically labelled twitter feeds as test set.

H and Sallis[51] gave an unsupervised approach for the purpose of analyzing sentiments. They employed Latent Dirichlet Allocation algorithm in their strategy. Their study validated their approach by using a dataset of around 10,000 tweets. The corpus was first of all pre processed and then was represented using vector space.

4.2.1 Integration of Naive Bayes and Advance Emoticon Classifier

In the propose hybrid approach we have integrated domain independent and dependent classifiers. The main classifiers used in the approach are Naive Bayes and Advance emotion classifier. The main principle on which the approach works is minimizing the Neutral tweets present which is the limitation of many studies in the approach.

1. Naive Bayes Classifier

Naive Bayes(NB) Algorithm is based on simple probabilistic classification , it work on principle of Bag of words and apply Bayes theorem with conditional Independence. It is best use for textual data and is and domain independent classifier It exploits the famous "Bayes Theorem" to calculate the probability of a given a particular feature set that belongs to some class.

$$P(Class|feature) = \frac{P(Class) * P(feature|Class)}{P(features)} \tag{4.1}$$

In Equation 1- P(Class) stands for the prior probability of the Class . P(features|label) stands for the prior probability that a given feature set is being classified by a particular Class. P(features) denotes that the prior probability that a given feature set has occurred.

Application of Naive Bayes Classifier to identify polarity of text- The working of Naive Bayes is explained with the help of following example :

Case – Consider the case that we need to find the polarity of particular tweet ie whether it falls in positive class or negative class or neutral class. Taking the assumption that we are given some count of already classified tweets, Let the frequency of positive tweets are double the count of negative tweets. Now, a tweet is found whose class is unknown and we have to find the accurate class of the new tweet(M). So we use the Bayes rule in order to find the probabilities considering the likelihood that tweet is either positive or negative.

Hence, from equation 1 we have:

$$P(n|p) = \frac{P(n)P(p|n)}{P(p)} \tag{4.2}$$

Since there are twice as many positive tweets as negative, it is rational to think that a new case (which hasn't been observed yet) is twice as likely to be positive rather than negative. In the Bayesian analysis, this belief is known as the prior probability. Prior probabilities are based on previous experience; in this case the percentage of positive tweets and negative tweets, and often used to predict outcomes before they actually happen.

Thus,

$$\textit{Prior Positive Probability tweet } P(p) = \frac{\textit{Count of positive tweets}}{\textit{Total count of tweets}} \tag{4.3}$$

$$\textit{Prior Negative Probability of tweet } P(n) = \frac{\textit{Count of negative tweets}}{\textit{Total count of tweets}} \tag{4.4}$$

Example -Consider 12k tweets in total, out of them 8k are positive and 4k negative, and where $k = 10^2$

Prior Positive Probability of tweet P(p) = 8k /12k = 2/ 3

Prior Negative Probability of tweet P(n)= 4k /12k = 1 /3

Since, the likelihood that the new tweet can exists in both the category is equal,. So likelihood of M = 0.5. Finding The Posterior Probability of M:

- Posterior positive probability of M =
  (Prior probability)×(Likelihood of M being positive)= 2/3 × 1/2 = 1/3 =33.3%.
  Thus they are 33.3 % percent changes that particular Tweet is positive.

- Posterior negative probability of M =

  (Prior probability)×(Likelihood of M being negative)=1/3×1/2 = 1/6 =16.67%.

Thus, we can conclude that M (new tweet) has more chances to fall in Positive Class

2. Emoticon Classifier

Nowadays people are using large number of emoticons in order to express their moods and emotions in order to start any kind of textual conversation. Micro-blogging sites like Twitter is An emoticon can consider as the pictorial form of facial expression which provides better visualizing of user feelings. An emotion is form by the aggregation of punctuation marks, alphabets numbers etc. It is basically formulated with the intention to convey one’s feeling to other person regarding some topic, thing or any particular person.

Read et al. proposed a new classifier which he named as emoticon classifier[52]. In this new classifier he exploited 10 emoticons. In our approach we have extended this classifier and named it as Advance Emoticon Classifier(AEC). In AEC we have used 60 emoticons. These emoticons are divided into two sets which are named as positive emoticon and negative emoticon. Positive emoticon depicts the positive feeling and negative emoticon portrays the negative feelings. Both the set of emoticons as discussed earlier is shown in the Table I. The negative emoticon set contains 23 emoticons in total and positive emoticon list contains 20 emoticons in whole and neutral list contains only 7 emoticons. Based upon this list we will use them in our proposal of new integration approach.

Table I list of Emoticons

<table>
<tr><th>Positive Emoticons</th><th>Negative Emoticons</th><th>Neutral Emoticons</th></tr>
<tr><td>:) :-) :o :] :c) :-D
:D 8-D 8D 8) :->
X-D xD :)) :3 :>
=] =-D :} :^) :c)
:-)) =) :ↄ) :P :-P
XP xp x-p</td><td>:(:-(=(:’(:((:-[
:c :’c :< :-< >:[
:-c :{ D: D8 D=
DX >:) :$ >:) >-:)
3-) 3)</td><td>;(:-O :O :o
>-O :| :-|</td></tr>
</table>

3. Integration steps

The key steps used to integrate the above discussed classifier are as following:

Step 1- Data-Extraction

Data set is the foremost part to test and validate any strategy formulated. We have used Twitter tweets to create Corpus for out study. We have use Application Programming Interfaces(API) to access Twitter data also we have used twitter streaming APIs. The details steps of how to obtain certain data from twitter is shown in Fig. 4.1. We have used link of application which is freely available to obtain authentication with Twittter. The url of application which was used to extract data is "-http://test.de/."

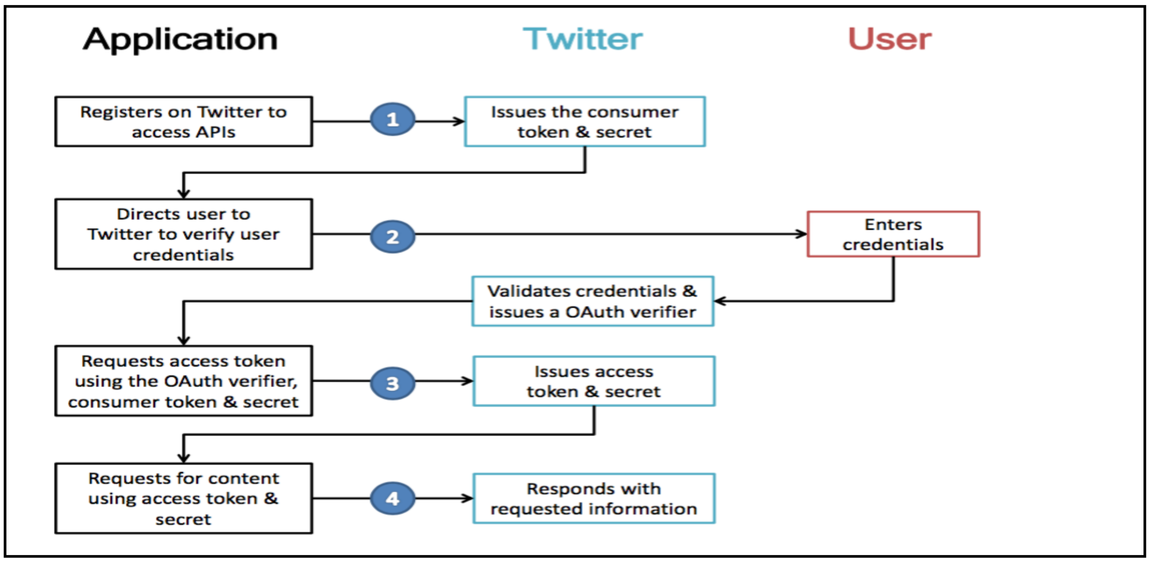


Fig. 4.1: Detail process to extract Twitter feeds

We have used test "twitter" R package to extract the tweets using R language. The query used to extract feeds is based upon keywords using the Nokia and Samsung .Fig. 4.2 gives the code to extract information using query and how to pass authentication to connect with twitter in R language.

```
install.packages("twitter")
install.packages("RCurl")              // Required Packages
install.packages("base64enc")
install.packages("httpuv")

setup_twitter_oauth(consumer_key='dchgxxx',consumer_secrect='hcd',     // Twitter authentication
                    access_token=NULL,access_secret=NULL)

tweetsforhybridApproach= searchTwitter('Nokia', lang='en')     //Applied Query
```

Fig. 4.2 : R script for twitter authentication and query applied

The query applied is based upon the keyword "Nokia" and "Samsung". For better accuracy of results the tweets for our study were restricted to English language since most of the public training are trained in English.

STEP 2- Classification

Step1- The twitter data sets used in study was distributed to students and teachers using Crowd Sourcing method to get better classification of the tweets into positive, negative and neutral. Based upon the majority classification of the tweet it is placed in particular division of classification, the classified tweet is then manually assigned the score 1for positive, 2for negative and 3 for Neutral polarity.

Step2 -After manually classifying the tweets we have applied naive baye algorithm using R language. R contains a "Sentiment" package which is used for this polarity identification purpose. Once we have got classified tweets into binary classification ie in form of positive/negative and score is assigned as stated in Step1. the neutral ones are fed to step3 since they contain wrongly classified tweets as problems exists to remove neutral data

Step3- Advance Emotion Classifier(AEC) is used in following way if we found the emoticon in negative list then it classified as negative , and if it found in positive list it is said to be positive and if dose not occur in any list it is called as neutral and score is assigned as in Step1. After that total score for sentence is calculated based on majority score of 1,0,or-1s and final classification is done and score is assigned. We have carried analysis at sentence level. After classification step is over score is allotted as explained in Step1.

STEP 3- Calculation of Overall Classification Score

After executing the first two , still if we find any tweet to be neutral then it is consider to be neutral only, else into positive or negative. The overall score is computed as by using the following equations.

$$\text{Neutral Score} = (\Sigma_N=0) \cap (\Sigma_E=0) \tag{4.5}$$

$$\text{Negative Score} = (\Sigma_N<0) \cup (\Sigma_N=0 \cap \Sigma_E<0) \tag{4.6}$$

$$\text{Positive Score} = (\Sigma_N>0) \cup (\Sigma_N=0 \cap \Sigma_E>0) \tag{4.7}$$

Σ_N,, Σ_E., Σ_A stands for scores obtained after applying NB, AEC respectively

4.2 Proposed GTAI framework

4.2.1 Problem Solution

To plug in the loopholes as discussed earlier was the key motivation to carry out this Study. We have devised a GTAI –geocentric technique for advertisement and improvisation to resolve the issues faced by the product companies. Our proposed technique will act as a middleware between advertisement agencies and SNS and between SNS and product company. Fig. 4.3 provides the diagrammatic view of SNS architecture from product perspective in which GTAI acts as a middleware.

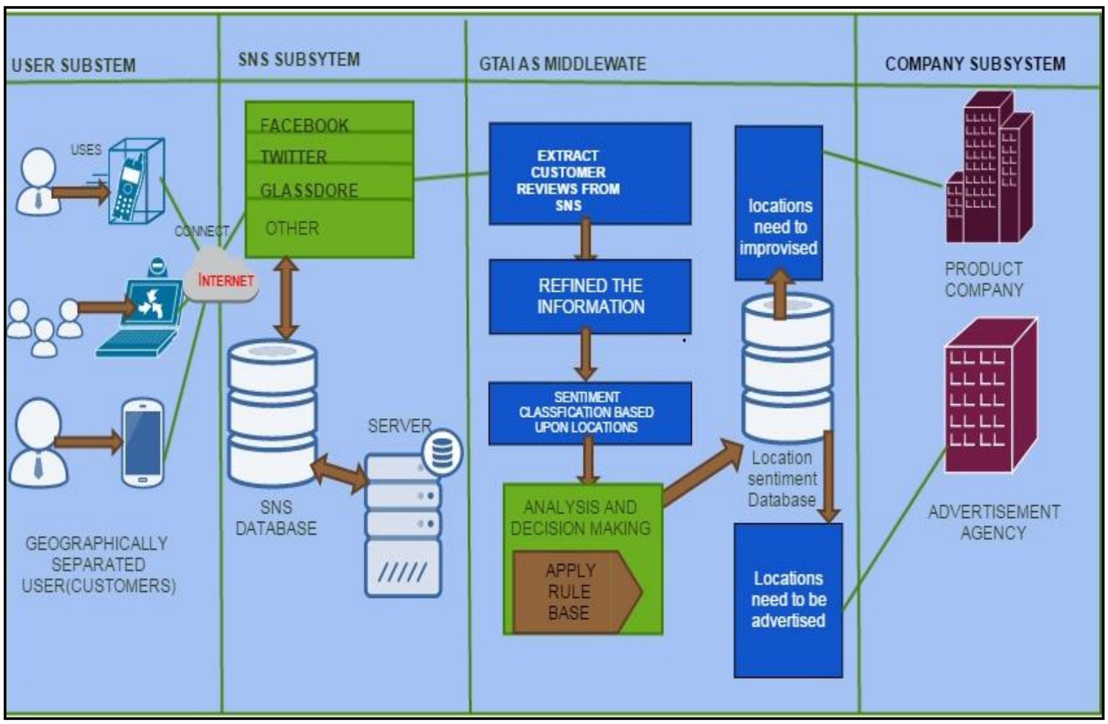


Fig. 4.3: SNS Architecture from product perspective using GTAI as a middleware

The details explanation of SNS architecture is given shown in the above diagram is illustrated below :

1.User SubSystem

In this sub-system we have consider Users (in particular customers) who use internet by the means of laptops and mobiles phones. These users are geographically distributed who share their opinions (in form of reviews, comments, feedback forms etc) about different products and services provided by products companies through the means of different social networking sites. Some of the popular Social media among customers are Facebook, Twitter, LinkedIn, Glassdore and various blogging pages. This information is a rich source for knowledge retrieval which aids various companies to interpret the opinions of users on the different services and products provided by them. Through this corpus of knowledge they can get complete picture of trends prevailing in a specific user community.

2.SNS Subsystem

Various Social networking services exists store which users opinions in forms of reviews, feed backs forms comments, personal information on their timeline etc. This information if retrieved can definitely serve an important purpose in knowing the user interest.

3.GTAI as a Middleware

GTAI acts as a middleware between Social networking sites and products companies and Between Social networking sites and advertisement agencies. GTAI extract information form SNS sites by the help of SNS application programming Interfaces(APIs), by help of many online and offline tools available which helps in directly obtaining the data from the relevant sites, by taking the help of 3rd party . The 3rd party member can constitute a individual or group of experts which help in extracting required data or by manually visiting the SNS sites and gathering the information. Once the information is obtained GTAI framework tasks are executed and then finally GTAI will provide the information to both product companies and advertisements agencies

4.Company Subsystem

Once the locations information is obtained by the product companies they could easily target these locations and give attentions to these geographical locations and similarly the advertisement agencies can forward more advertisements to locations and can even improve the frequency of ads to the relevant locations

4.2.2 GTAI Components

The GTAI technique is broadly partitioned into 4 main components which are Data Acquisition component, Data pre-processing Component, Sentiment Classification Component, Analysis and Decision making component. These modules are listed and explained in the following subsections. The block diagram of GTAI framework showing the main components of the framework is depicted below in Fig.4.4

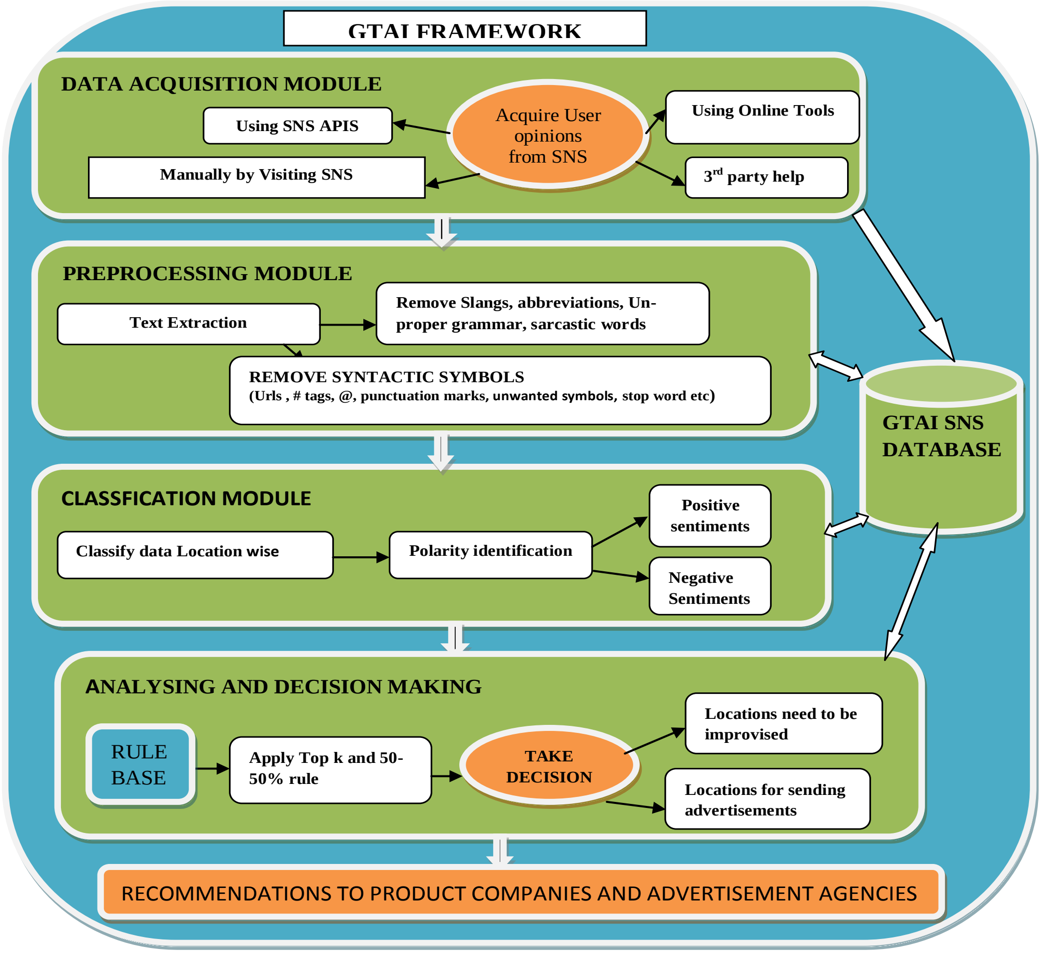


Fig. 4.4 :Block diagram of GTAI framework

The detail description of the main components of the GTAI framework with their working and purpose is illustrated below:

1. Data Acquisition Component

Data extraction is foremost and crucial step of our proposed technique. The data for GTAI can be collected from any existing Social media. However the data collected from the SNS site should be based upon the geographical distribution containing user opinions and interests pertaining to any kind of product or service. The data can be acquired by four different methods 1. Using application of APIs (application programming interface), 2.Use of online tools available, 3. By the help of 3rd party company or any individual which works on your behalf and provides you with the required information, 4. Manually collecting the user information (in form of reviews, feedback forms, comments) by visiting the relevant networking site. Out of the four methods listed above 4th method should be avoided and least used as it a time consuming method and should be used only in rare cases.

2 Data Pre-processing Component

Internet reference data is in the form of large multimedia document containing text, images, videos, audios etc. The devised GTAI strategy is applicable only for text information since it the best way to analyze and judge user opinions. The textual information obtained from any SNS sites can be collected from anyone forms like tweets, feedback forms, user reviews, posts, timeline status etc. They contain large number of unwanted data in the form of syntactic symbols and unformed Part of sentence, use of sarcastic sentences, illegal use of grammar, user sarcasm and so on. The data need to cleaned of all these so that better accuracy in the results of sentiment Analysis can be achieved. The processed information once obtained is fed to Sentiment classification component for further classification. The algorithmic steps involved to carry out this module are illustrated below.

Input: Geographically distributed user comments

Output: Pro-processed information

4.1 Algorithm for Pre-processing component

```
Procedure Pre-process (for each user opinion)
    Extract textual information (T)
    Procedure Clean (apply on text (T) data)
      Convert to Upper Case
      Procedure Remove Syntactic Symbols(S)
         Remove Urls
         Remove Hash tags
         Remove Punctuation marks.
         Remove Special symbols
      End Procedure S.
      Remove Slangs
      Remove Abbreviations
      Remove Stop words
  End Procedure P
```

4.1.3 Sentiment Classification Component

The processed information retrieved from the previous module is used to find out thepolarity of the textual information by categorizing the user opinions into binary classification i.e. on the basis of positive and negative scale. We have used Naive Bayes classifier which works on the principle of probabilistic computation .After classifying and identifying the polarity on the basis of geographical location further analyzing is done. In this algorithm once we have apply the naive bayes classifier group wise data. For all the opinions in a location if a opinion has a positive polarity we have increased the count of positive column in Sentiment file for the relevant location and if it has negative polarity we increase the count of negative column. Once the opinions for particular location is over we will repeat the Procedure Classification for other locations. The algorithm used to carry out the above steps is show below

Input: Refined user comments

Output: Classified tweets as positive and negative

```
4.2 Algorithm for Sentiment Classification

        Procedure Classification(C) [for each user opinion]
          Classify opinions by applying Naive Bayes
              If Opinion is positive
                  Increment the count of +ve sentiments in Positive column in
            Sentiment file for each location
              End if
              If  Opinion is Negative
                  Increment the count of -ve sentiments in Negative column in
            Sentiment file for each location
              End if
      End Procedure C
```

4.1.4 Analyzing and Decision-making Component

For analyzing the information (in form of sentiments obtained from previous component) and taking decisions we have formulated a rule base which is illustrated in detail in following subsection

Rule Base

We have created two rules for the purpose of be better understanding which locations need to be improvised and also to know which locations need to be promoted(in the form of advertisements). The two rule are Top k and 50-50% rule.

Rule1 – Top K rule- Pick k locations from N,

Where k is number of locations to be picked and

N is total number of locations used in the study.

Using **Condition (i)** if N <= P, Pick k= N/2 if N is even

Pick k= (N+1)/2 if N is odd

Where P is the half of the partition (granule) of that the particular boundary unit selected

Consider the case:

Boundary unit = India
Granularity level = States
N=14

Now India has 29 states
Therefore P = ((29+1)/2) i.e. 15)
Now N(14)<P(15)
So Condition (i) Is applicable

Result K =14/2 =7

Using **Condition (ii)** if N>P , Pick k= (count of all n_i >R)

Where (i belongs to [1, max]
max is maximum count of locations taken in the boundary unit)

Where (R is the range
$R=(n_{max} – n_{min})$
n_{max} is the location with maximum number of positive sentiments
n_{min} location with minimum number of positive sentiments.)

Consider the case:

Boundary unit = India

Granularity level = States

N=19

Now India has 29 states

Therefore P = ((29+1)/2) i.e. 15)

Now N(19)>P(15)

So Condition (ii) Is applicable

Note: The rule is applicable only if the locations have the same granularity (i.e. either it should be all cities, all countries or all states etc.)

In this rule among the geographical locations for which we have identified their sentiment polarity we will pick top K locations with higher frequency of positive sentiments. These locations are directly applicable for the purposes of Advertisements. Since these are locations among all the other locations have higher rate of acceptance level either it for service or product existing or newly launched in the market. These are the locations where the product is in great demand. The remaining locations which are not picked using this rule need to be further analyzed since we cannot directly take a decision whether these locations needs to be improvised or should be used for advertisement purpose , so these locations are applicable for 50-50% rule. The Algorithm to apply

Input: Classified user sentiments according to geo-locations

Output: Classified tweets in neutral, positive and negative polarity

4. 3 Algorithm for Applying Top k rule

```
Procedure Analysis (A)
    Sort location according to the decreasing order of count of +ve s
    sentiments
      Apply Top k rule on these locations
        For k locations picked for Using Top k rules
          Send Notification - Advertisements on these locations
        End For
    Remove k locations from sentiment file.
End Procedure (A)
```

Rule 2- 50-50%- This Rule is applicable to the remaining Locations which were not used in rule 1. In this rule percentage of positive and Negative sentiments for each location is calculated using the formula give below.

$$\text{Formula(+ve percentage of sentiments)} = \frac{\text{+ve sentiment count}}{\text{Total Sentiments for the location}} \quad (4.8)$$

$$\text{Formula(-ve percentage of sentiment)} = \frac{\text{-ve sentiment count}}{\text{Total Sentiments for the location}} \quad (4.9)$$

After calculating the percentages , if the percentage of positive group is more than negative percentage for a any specific location, then that location is applicable for advertisement else if percentage of negative is more it is target for Improvisations and recommendations to the web site owners of the particular product or service . The complete algorithmic steps of this module are shown below.

Input: Classified user sentiments according to geo-locations

Output: Classified tweets in neutral, positive and negative polarity

```
4.4 Algorithm for Applying Top k rule

    Procedure Analysis(AN)
            Apply 50-50% rule on the remaining locations in sentiment file
              If location contains > 50 % of +ve Tweets
                Send Notification- Advertisement on these locations
              End If
              If Location contains<50% of +ve Tweets
                 Send Recommendations for improvisations on the location to
                  marketing websites.
              End if
   End Procedure AN
```

Summarizing the chapter at the beginning we have study some of the existing techniques for Twitter feeds . After that we have proposed a new technique by integrating NB and AEC and at last we have proposed a GTAI framework with detail description of the key components.

Chapter 5
Case Study

In this chapter we have implemented the proposed GTAI by considering Facebook as the test bed . Also the various tools and technology used in the implementing are discussed before hand. Also the GTAI frame work is validating on the basis of marketing parameters.

5.1 Overview

We have created a Facebook Page " MyProducts" for the purpose of implementation and validating our proposed GTAI strategy in the real environment. The page is basically about two products- Product 1 is based upon "Avon products" and Product 2 is based upon "Amway products". Fig 5.1 shows the snapshot of MyProducts Facebook page which was develped to carry out our study. In this page various posts regarding new products launched, discounts available, links to buy the products online are provided. We have manually posted the links and have also used The page was in use in order to conduct the experiment from the period starting from 23 August 2015 till 23 April 2016. For the first six months(23 April 2015 to 23 September 2015), we have gathered user reviews in the form of comments on the various posts published on the page. The collected comments were relevant to both products. After collecting the comments for the first 6 months, then we have applied GTAI framework on the collected comments. After applying the GTAI framework we have again collected the user reviews from the page for the remaining period and then we have validated GTAI on the following three parameters i.e. Maintaining Goodwill, Consistency, and building Audiences. Fig 4 below shows the MyProduct Facebook page Snapshot

Fig. 5.1 : Facebook “MyProducts” Page

5.2 Tools and Technology Used

5.2.1 Hootsuite Dashboard: Ad Management Tool

We have used Hootsuite social media management dashboard to posts ads to MyProducts page. The dashboard help in scheduling the ads to any social network you want from one single place only. Through this tool we have schedule the relevant posts, ads, links that will posted in specific time in future. Fig. 5.2 Ads links being posted on the page.

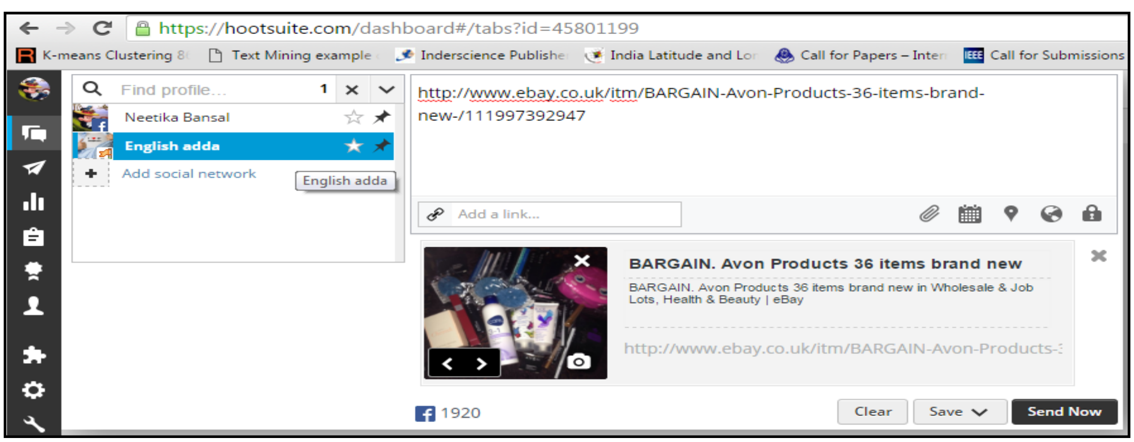


Fig. 5.2: Hootsuite dashboard used to posts ads on MyProducts page

We can see in the Fig. 5.3 the Ad link which was used to be published on the page through the use Hootsuite dashboard is posted on the MyProducts Page.

Fig. 5.3: Ad Link used via Hootsuite dashboad is published on MyProduct Page

5.1.2 MySQL

MySQL is one of the leading open-source relational database management system . In the year 2013, it was declared as the second most widely used system around the globe. It is also widely used and accepted open-source client–server model RDBMS. We have used MySQL Workbench 5.1 which provides the environment for MySQL database creation and provides better visualization of the data. We have used MySQL database to store the user comments extracted from the Facebook page.

5.1.3 R language

We have used R studio , which provides the integrated-development environment for the creation of R language. It also provides user friendly platform for easy use of R language . R is a statistical programming language which is further used for the purpose of computing. It contains various inbuilt packages which can be easily used. We have used R language for the purpose of pre-processing the extracted data. Further to carry out our work we have also carried out polarity based SA on user comments by the help of NB classifier.

5.2 Experimental Setup

The experimental setup is carried by the help of Facebook page in order to implement the proposed GTAI framework. GTAI experimentation is carried out by adhering to the following steps illustrated in the activity diagram which is shown in Fig. 5.4. There are total 12 steps in the GTAI activity diagram which are described in the subsequent subsection.

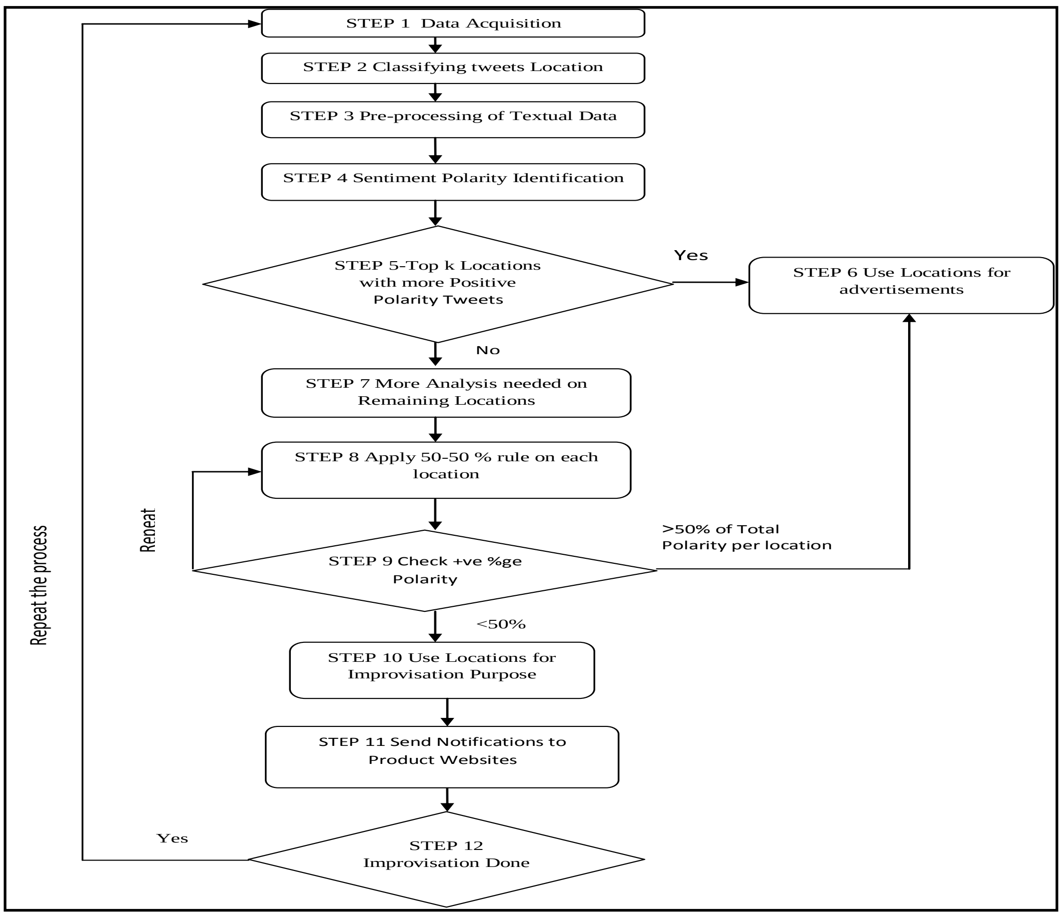


Fig. 5.4: Activity Diagram for the implementation steps of GTAI.

STEP1 – Data Acquisition - We have collected around 455 and 223 user reviews in the form of comments for Product Company 1 and Product Company 2 respectively for the period starting from 23 April 2015 until 23 September 2015 using Facebook graph APIs.

Problem faced- At the beginning we were not getting required traffic on our Page so that we can carry out a research effectively, so we promoted our Facebook page by using Facebook promoting and boosting ads and also requested people to join our page till we get the threshold number of comments which were enough to carry out the research.

STEP 2 - Classification of tweets location wise – The next step is used to classify the tweets into finer granule based upon the boundary unit we have allocated. We have allocated boundary unit for our study as country ie India so we have partitioned data state wise. To collect the location of each review(user comment) we have visited the user timeline since facebook Graph Api Version V1.6 have put restriction on getting location information through the APIs . However, we can replicate the study using other Social media like Twitter which allows to extract location of the user(in form of geocode-longitude and latitude) through the user Twitter Apis. Out of the 455 and 223 only 394 and 198 comments respectively where useful for further classification into state wise partition. The main reason for the reduction of comments was some of the comments belong to the users of other countries; some of them were from India only but have not provided information of state on their timeline in order protect their privacy and some comments were not in English languages. Example list of comment extracted form the page is shown in Fig. 5.5.

10	When are u going to dispatch my order number Y087DTSM, poor management, Realy sad.
11	Products here available which are cheaper then the mrp?
12	Honey has lot of sugar in it..its not good for our health..plz don't cheat us
13	Al are natural, safe.. proved for use for al family member
14	My whole family is using products. we are very found of them
15	Really gd deal !!! never expected such a good deal;(
16	Can any one tell me where can I find gd quality products reality to hair oil???.. i culd not find any pls help.......!1
17	@Ownere why r u not deliovering the product on time ... really sad to say poor managements System
18	The new deal launched one on one... is grt deal to grab ..Very gd deal after all..I m loving it :)

Fig. 5.5 : Example list of comments collected form page

We have used Mysql database to store the information we have retrieved. Sql script used to create table ‘my_product_review’ is shown in Fig. 5.6. This data table consist of following columns like user_comment(to store extracted user comment from My_products page), product_company_name(to store name of the companies), date_of_comment(the date when user commented on the post), Location(user state who posted the comment).

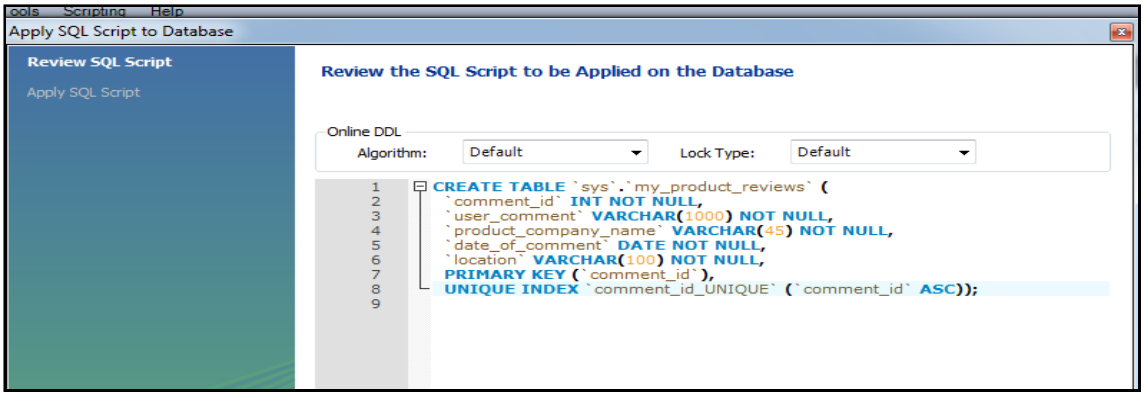


Fig. 5.6 sql script to create table 'my_prooducts_reviews'

Problem Faced- We have manually collected the location information of each and every comment by visiting user timeline. This was a bit time consuming process. The main reason for this problem was we could not extract user location information using Facebook APIs due the strict restrictions imposed by Facebook Graph API in extracting user personal information.

Geographical locations(states) from where we have received the user comments for both product 1 and product 2 is shown in Fig. 5.7.

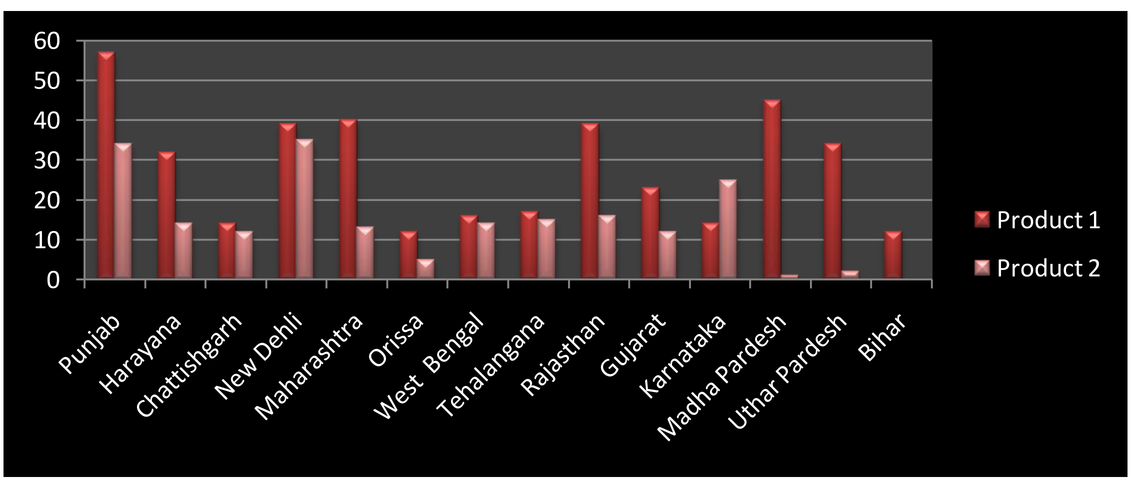


Fig.5.7: State-wise distribution of user comments on MyProducts page

5.3. STEP 3 –Pre-processing of Textual Information- The user review collected in the earlier steps is in text format only so there was need to extract any separate text data.

However, SNS users have a habit of adding symbols and using slang languages. We have remove the unwanted symbols and words to obtained accuracy in results:

- Removal of Syntactic Symbols- The collected user comments contain various punctuation marks like (,!;), special symbols like(@, #) other symbols like (Url, Http) etc we have removed these word with the help of R language. R tm(text mining) package is used for this purpose which contain "tm_map" function which helps in carrying out the above tasks.
- Removal of slangs or short forms -Users generally have a habit of using short forms or slang language. These words are not required for GTAI process as they act as extra overhead in form of time so these words need to be discarded. To discard we have used "Netlingo dictionary", we can check the meaning of words by searching through the dictionary. The words which have meaning are displayed and ones which do not have meaning or similar words is not displayed and we can remove these words.
- Removal of stopwords- The English language stop words is removed which are remove by the help of R language..

5.4. STEP 4 -Sentiment polarity Identification -We have applied naive bayes classifier on the processed information by the help of R language. "Sentiment package" is used for polarity identification of the obtained user comments. It is a 3rd party developed package which works with R . The package contains a function named 'polarity' which is used to classify text on basis of their polarity. R scrip used for applying naive bayes on data is taken from the following link[. Fig 5.8 shows how each word in the comment is assigned positive or the negative polarity .

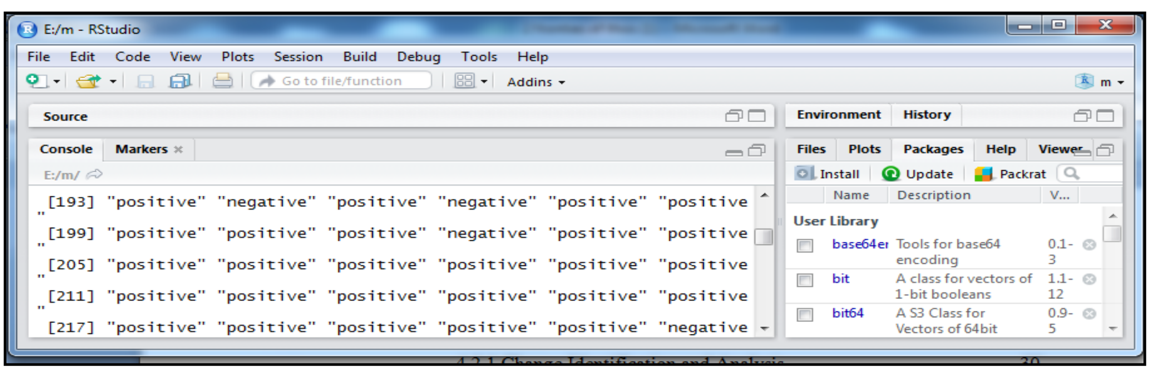
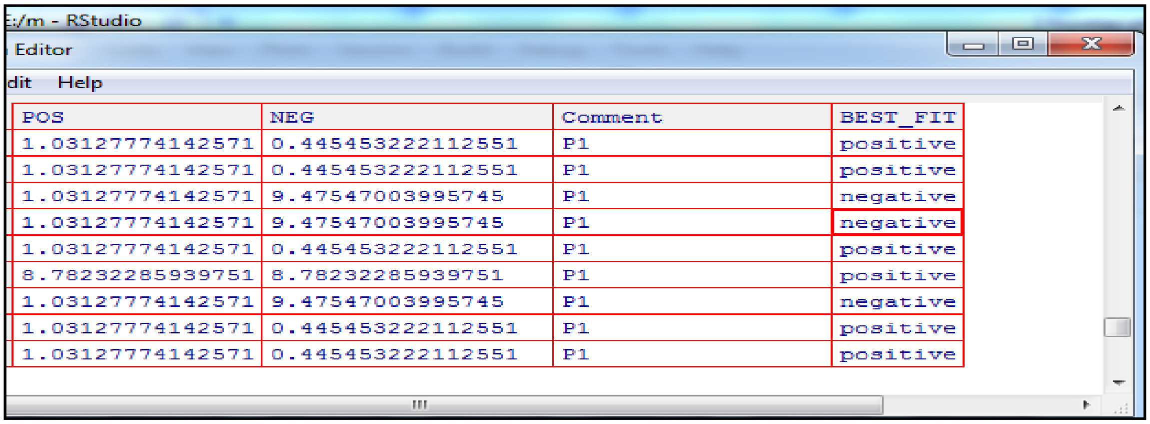


Fig 5.8 : Word wise polarity identification.

However, we have made a slight change in the R script instead of calculating ternary polarity in form of positive, negative and neutral basis we have carried out binary classification in form of positive and negative user sentiments. Fig 5.9 shows the Total probability of positive sentiments and negative sentiments total probability of probability calculated for product 1 for geo location Punjab.



POS	NEG	Comment	BEST_FIT
1.03127774142571	0.445453222112551	P1	positive
1.03127774142571	0.445453222112551	P1	positive
1.03127774142571	9.47547003995745	P1	negative
1.03127774142571	9.47547003995745	P1	negative
1.03127774142571	0.445453222112551	P1	positive
8.78232285939751	8.78232285939751	P1	positive
1.03127774142571	9.47547003995745	P1	negative
1.03127774142571	0.445453222112551	P1	positive
1.03127774142571	0.445453222112551	P1	positive

Fig 5.9 : Overall polarity identification.

In this step we have carried out polarity classification at sentence level where each comment from specific geo-location is consider as one sentence. Fig. 5.10 gives the state-wise distribution of count positive sentiments for both products i.e. product 1 and product 2.

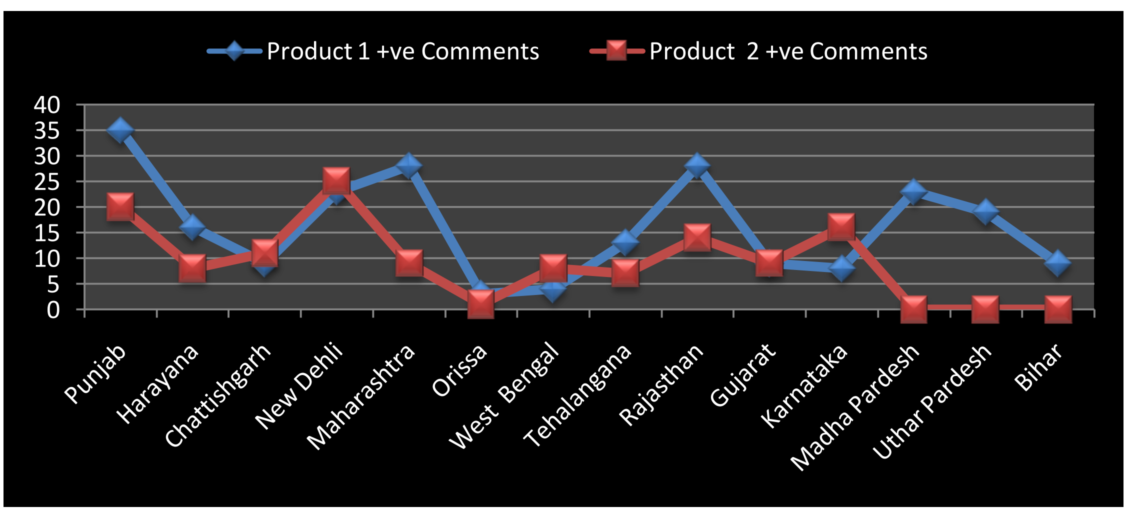


Fig.5.10 State-wise distribution of positive sentiments for both Product 1 and Product 2

Fig. 5.11 provide the information of the state-wise spatial distribution on the basis of the count of negative sentiments for both Product1 and Product 2. From the diagram we can easily depict that New Dehli have approximately same number of negative sentiments.

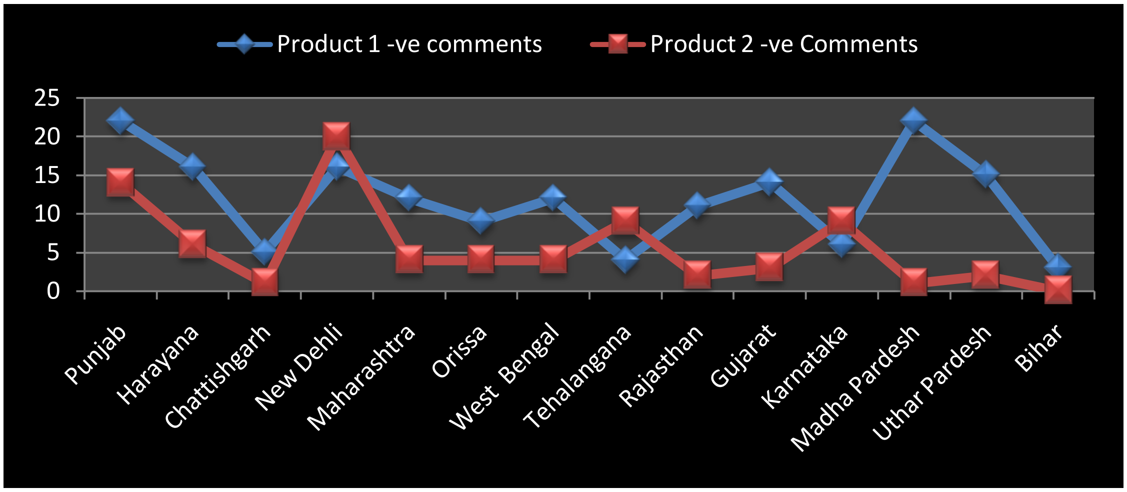


Fig.5.11: State-wise negative sentiments for both Product1 and Product 2

5.5 STEP 5- Top K locations with more positive sentiments - In this step we have applied top k rule since we have 14 states so Ist condition of rule is applied i.e. when N<=R here R=15 (number of states in country is 29)and also N is even so we will pick N/2 locations i.e. top 7 locations which have higher frequency of positive sentiments. The data of 7 locations which is retrieved for both Company 1 and Company 2 are shown in TableII and Table III respectively . The locations which are picked using this rule are straightway send to step 6(i.e. these locations are used to advertising purpose and there is no need to apply further analysis on these locations.

TABLE II- States picked for Product 1 after applying 50-50% rule

STATE
PUNJAB
MAHARASTRA
HARYANA
MADHYA PARDESH
NEW DEHLI
UTTAR PARDESH
GUJARAT

Table III- States picked for Product 2 after applying 50-50% rule

STATES
PUNJAB
NEW DEHLI
HARAYANA
TEHLANGANA
MAHARASHTRA
ORISSA
KARNATAKA

5.6 STEP 6. Use locations for advertising-The states which are applicable for this step are one which need to be more frequently advertisements since these are locations are one where customer have shown more interests(more positive sentiments) for the company's products. After the locations have been selected for advertisement notification can be send to advertisement agencies for the deploying more advertisement at these locations. Since in our case study we have not included any external Advertisement agency so we have used face book ads in form of promoting shown in Fig. 5.12 and boosting to target more people at these locations.

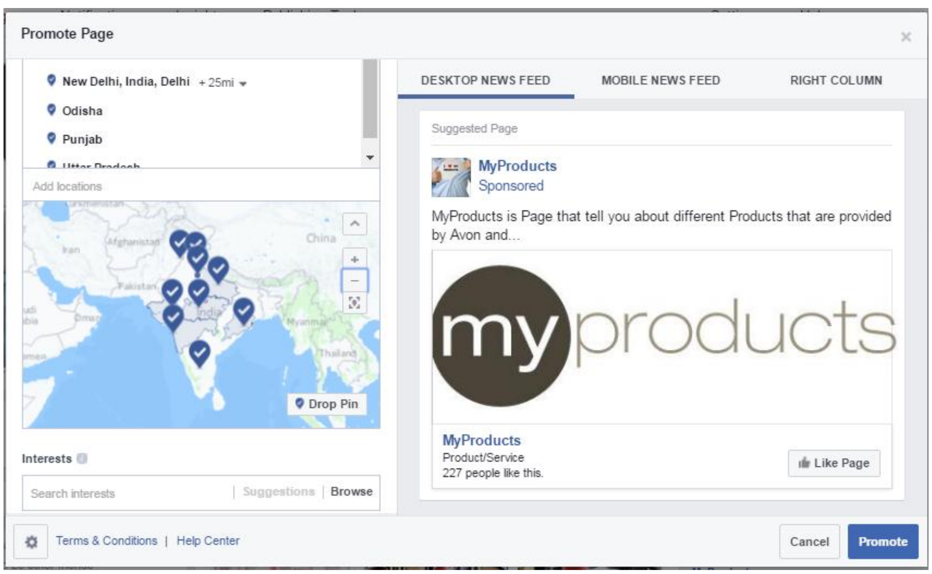


Fig. 5.12 :Promoting MyProducts Page at different locations

5.7 STEP 7 More Analysis needed- All the remaining states on which Top k rule was not applicable need to be more analysed in order to take the accurate decision. Since , we cannot directly say that these location need to be improvised or need to more advertised .

5.8 STEP 8 Apply 50-50% rule on each location- The location which needs more analysis, we have 50-50 % rule on that location. The locations for company1 and results obtained after applying 50-50% rule is shown in Table IV.

TABLE IV- Results after Applying 50-50 rule on Product 2

STATE	%age of positive sentiments	%age of negative sentiments
West Bengal	25	75
Chhattisgarh	64.2	35.8
Odisha	25	75
Tehlangana	76.4	33.6
Rajasthan	71.7	28.3
Karnataka	57.1	42.9
Bihar	75	25

The locations picked for for product 2 and the results obtained after applying 50-50% rule are shown in Table V.

TABLE V- Results after Applying 50-50 rule on Product 2

STATE	%age of positive sentiments	%age of negative sentiments
West Bengal	66.6	33.4
Chhattisgarh	91.6	8.4
Gujarat	75	25
Uttar Pradesh	0	100
Rajasthan	71.7	28.3
Madhya Pradesh	0	100
Bihar	0	0

5.9. STEP 9 – Check +ve polarity- Once the 50-50% rule have been applied to the locations we have to check the percentage of positive sentiments. The locations with positive polarity greater then 50% of the total polarity for a particular location are chosen for step6 and the remaining locations are chosen for STEP 10.

5. 10. STEP 10- Use locations for Improvisations Purpose- The locations chosen for this step are the one which need to improvised i.e. at these locations more advertisements need to be send or special discounts to be offered attract the user of these locations

5. 11 STEP 11- Send Notifications to Product Companies- We have sent the notification recommending product companies to improvise the locations. We have sent the notification through the emails and message form facebook page.

5.12 STEP 12- Improvisation Done- Once the improvisations at the locations have been done again GTAI process can be applied on these locations. However in our case we could get the

5.3 Results and Analysis

After we implemented the GTAI framework, we have again collected the data from MyPage in the form of the user comments. The collected data was used to find out the effect of GTAI on 3 main factors Building Audience, Consistency, Maintaining Goodwill .These factors are analyzed in following sections

5.3.1 Building Audience

Attracting audience is very important for any kind of advertising agency or any product companies from the point of view of marketing and brand building. It has also been a challenging task for This factor can be viewed on basis the parameter i.e. Increase in the likes on the Page.

Fig. 5.13 shows the Total likes before and after the application of GTAI framework on MyProducts page. For the First six months (when GTAI was not applied) there were only 75 Likes on the page. However for the remaining for 4 months(After the application of GTAI) there were addition of 152 new likes. The new like count is approximately double the earlier like count. The increasing curve in Fig. 5.13 clearly shows that GTAI had played a significant role in building the audience in form of likes on the Facebook page

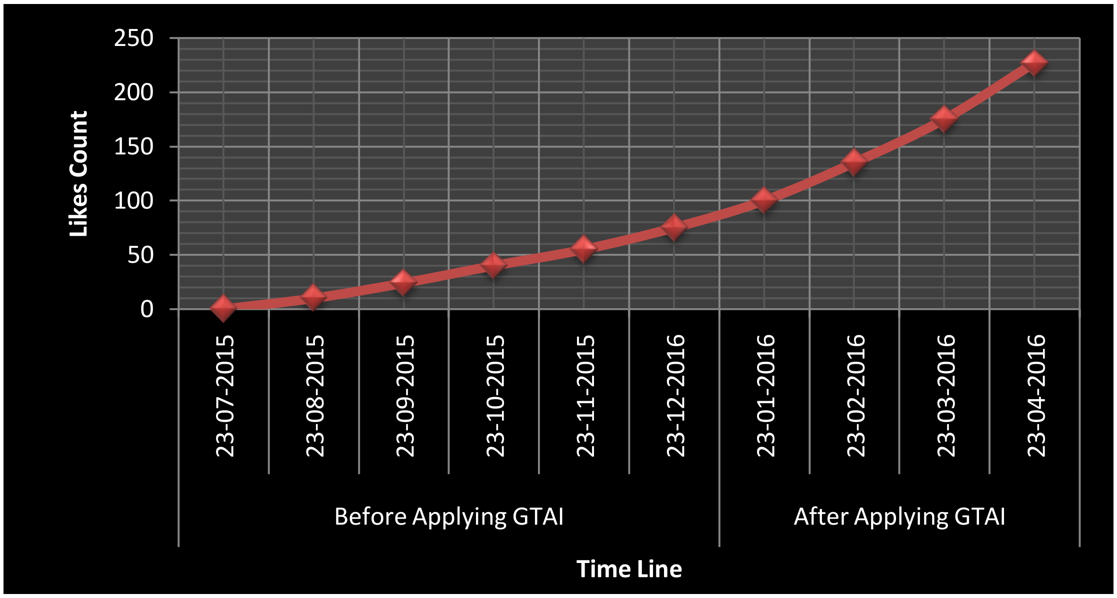


Fig. 5.13:Count of users who join before and after the applying GTAI

5.3.2. Consistency

Consistency of the user interests for the products of any business company is biggest issue till date. Most of the customers lose interests in the products for particular in short period of time. Consistency can be tested by considering the parameter frequency of users in giving their input . In this parameter we have consider the frequency of number of comments given by users for both the products based upon the relevant geolaoctions. Fig. 5.14 shows the effect of GTAI in maintaining consistency of user interests for Product 1 . From figure we can easily interpret that in majority of locations the user comments for Product 1 have significantly increased after the deploying of GTAI framework or at least have maintained previous count of user count. Only locations Uttar Pardesh and Maharashtra have seen a decreasing trend. However , the decrease at these locations are very little and can be overlooked. Not only that GTAI have helped in gaining the interests of users from area which have zero user comment count.

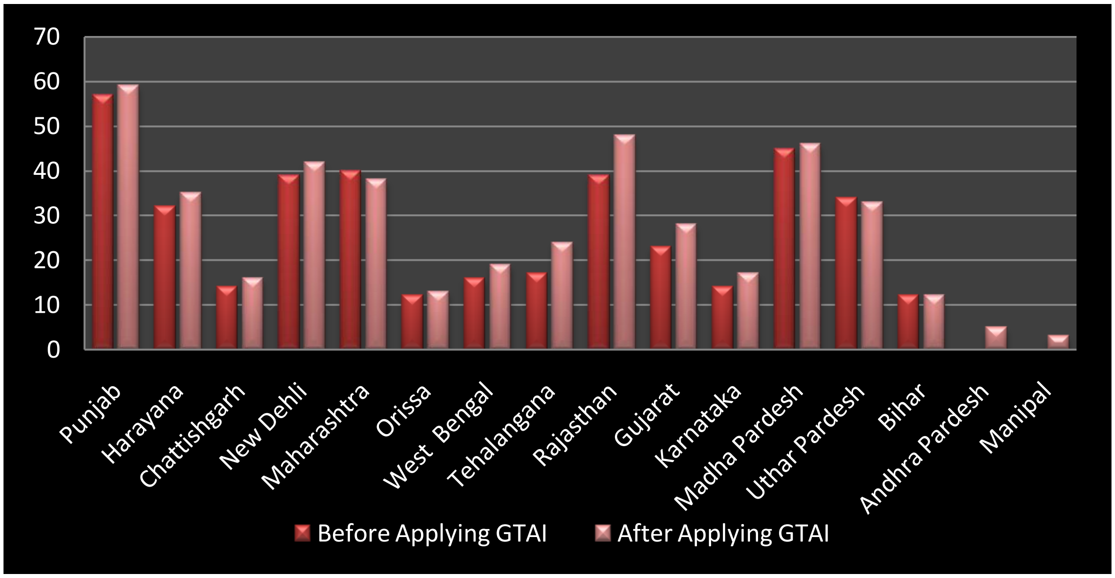


Fig. 5. 14 –State wise distribution of user comments before and after applying GTAI

Similarly from the Fig. 5. 15 we can see that product 2, GTAI has help in increasing the user comments even for those locations , for which ones their were no comment e g. consider the case of Andhra Pardesh and Manipal. Moreover, there is drastic increase in comments in some areas like Karnataka and Maharashtra. Thus we can conclude that GTAI has help in maintaining the consistency .

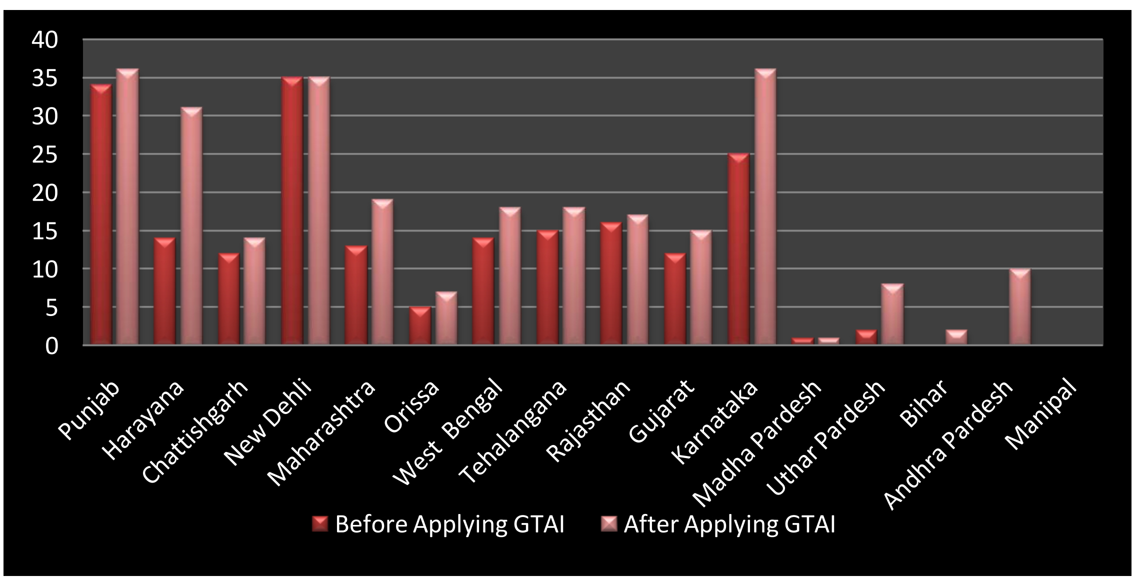


Fig. 5. 15 –State wise distribution of user comments before and after applying GTAI

5.3.3.Maintaining Goodwill

Goodwill of the company can be calculated on the basis of the parameter “positivity of user sentiments”. Fig 5.16 shows the effect of GTAI on positive sentiments of Products 1. From the figure we can easily see that in majority of states the positive sentiments have increased in significant amount. Thus we can easily conclude that GTAI have positively effective the good will of the company.

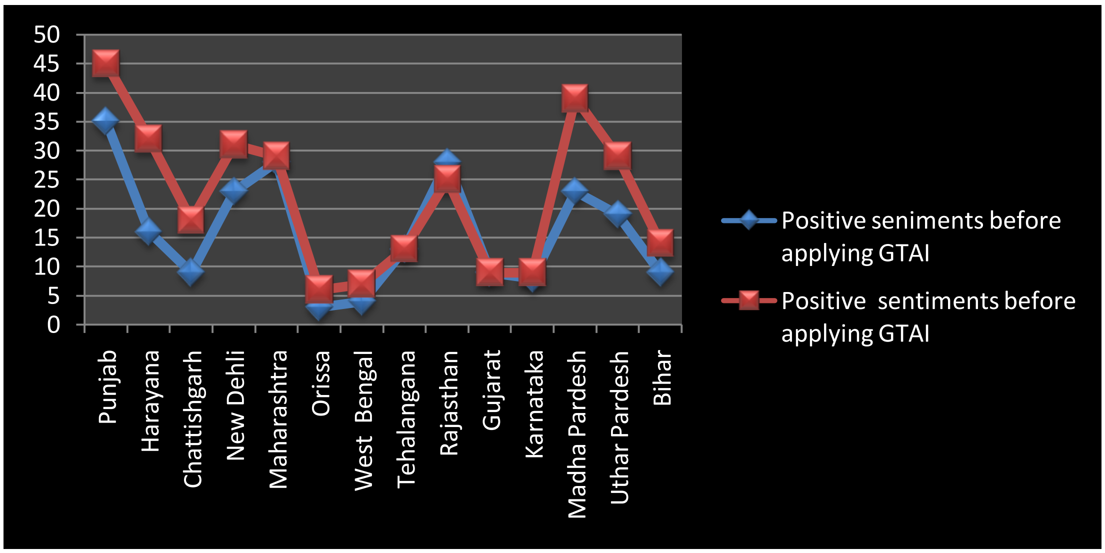


Fig. 5.16 State-wise Distribution of positive sentiments for product1 before and after GTAI framework was deployed.

Chapter 6

Conclusion and Future Scope

This chapter concludes the work carried out in the thesis by highlighting the unique contribution made by the present work. It also pans light on the limitations of the work and have also discussed the future scope of the study.

6.1 Conclusion

In this thesis, we have studied the field of SA in detail. We found that ML algorithms are used by majority of studies. Among them NB classifier is one of the simplest and reliable algorithm for text classification. Also emoticons are gaining attention from the research point of view. We have proposed a new technique for sentiment detection by integrating NB and AEC. The proposed technique works on the principle of minimizing neutral sentiments. Moreover we found lack of work done on the Ads , so we have devised GTAI which mainly works on the extracting data from SNS and using exploiting the geo-locations.

6.1 Summary of Unique Contribution

This thesis has made a commendable contribution by presenting GTAI framework. It has been a stepping stone in the area of improvising advertisements by use of user sentiments based upon their geo-locations. Furthermore, it also acted as a middleware between SNS and ad agencies and can aid various product companies in taking proper decision regarding which group of user community to be targeted based upon the exploitation of users geo-locations. Also the GTAI has been implemented using Facebook as a test bed and use of real time ads. GTAI was validated on the basis of online advertisements specific parameters such as building audience, consistency and maintaining goodwill. GTAI has impacted positively on all the listed parameters.

6.3 Limitations of the study

The limitation that existed in the study is as following:

- The GTAI framework developed is implemented by considering India as the boundary unit . The users comments are based upon only location pertaining to

Country India.

- As products are specific to particular company so it is illegal to use name of any product, so we have only concentrated on the products which are prevalent in India and the one which are easily available for retailing purpose.

6.4 Future Scope

The proposed GTAI framework was implemented by considering only those products which can be easily accessible and the strategy was implemented by considering only product reviews. However, this strategy can be used by the business companies in order to target specific geo-separated audience by the help of advertisements

.

References

[1] E. Fisher, ”*In depth social media*”, Sports Business Journal(2013), SportsBusinessDaily:http://www.sportsbusinessdaily.com/Journal/Issues/2011 /08/01/In-Depth/Socialmedia.asp

[2] J.H Kietzmann, et al. “*Social media? Get serious! Understanding the functional building blocks of social media*” Business Horizons 54 , 241—251 .

[3] http://dazeinfo.com/2013/04/01/social-media. [Accessed: 10-Apr-2016]

[4] Facebook Newsroom. (2016). Facebook: http://newsroom.fb.com/Key-Facts [Accessed: 13-March-2016]

[5] Sahayak et al., "*Sentiment Analysis on Twitter Data*." ACM SIGACT ,2015

[6] B. Gokulakrishnan et al., “*Opinion mining and sentiment analysis on a twitter data stream*”, International Conference on Advances in ICT for Emerging Regions (ICTER), 2012, pp. 182–188

[7] I Lunden, *TechCrunch*. from TechCrunch: http://techcrunch.com/2014/06/30/analyst-twitter-passed-500m-users-in-june-2014-140m-of-them-in-us-jakarta-biggest-tweeting-city/[Accessed: 10- March-2016]

[8] Kaiquan Xu , Stephen Shaoyi Liao , Jiexun Li, Yuxia Song, “*Mining comparative opinions from customer reviews for Competitive Intelligence*”, Decision Support Systems 50 (2011) 743–754

[9] F.H. Khan et al., “*TOM: twitter opinion mining framework using hybrid classification scheme”*, Decision. Support Systems. (2013), http://dx.doi.org/10.1016/ j.dss.2013.09.00

[10] W Medhat et al., "*Sentiment Analysis Algorithms And Applications: A Survey*", Ain Shams Engineering Journal, 2014

[11] http://webservices.itcs.umich.edu/mediawiki/YaffeCenter/sites/YaffeCenter/uploads/thumb/8/88/18.png/775px-18.png [Accessed: 11 March 2016]

[12] P. Massa, B. Bhattacharjee, “*Using trust in recommender systems: an experimental analysis*”, iTrust 2004, pp. 221–235

[13] C.K. Hsu, D. McDonald, "*An examination on multiple celebrity endorsers in advertising*", Journal of Product & Brand Management 11 (1) (2002) 19–29

[14] S. Wang, D. Li, X. Song, Y. Wei, H. Li, "*A feature selection method based on improved Fisher's discriminant ratio for text sentiment classification*", Expert Syst. Appl. 38 (2011) 8696–8702

[15] J. Zhu, H. Wang, M. Zhu, B.K. Tsou, M. Ma, "*Aspect-based opinion polling from customer reviews*", IEEE Trans. Affect. Comput. 2 (1)(2011

[16] L. Jiang, M. Yu, M. Zhou, X. Liu, T. Zhao, "*Target-dependent Twitter sentiment classification*", in: ACL, June 2011, pp. 151–160

[17] C. Chen, Y.-D. Tseng, "*Quality evaluation of product reviews using an information quality framework*", Decis. Support Syst. 50 (2011) 755–768

[18] R. Xia, C. Zong, S. Li, "*Ensemble of feature sets and classification algorithms for sentiment classification*", Inform. Sci. 181 (2011) 1138–1152

[19] X. Bai, "*Predicting consumer sentiments from online text*", Decis. Support Syst. 50 (2011) 732–742

[20] Z. Zhang, Q. Ye, Z. Zhang, Y. Li, "*Sentiment classification of Internet restaurant reviews written in Cantonese*", Expert Syst. Appl. 38(2011)7674–768

[21] M.R. Saleh, M.T. Martín-Valdivia, A. Montejo-Ráez, L.A. Ureña-López, "*Experiments with SVM to classify opinions in different domains*", Expert Syst. Appl. 38 (2011) 14799–14804.

[22] A. Abbasi, S. France, Z. Zhang, H. Chen, "*Selecting attributes for sentiment classification using feature relation networks*", IEEE Trans. Knowl. Data Eng. 2 (3) (2011).

[23] S.K. Li, Z. Guan, L.Y. Tang, et al., "*Exploiting consumer reviews for product feature ranking*", J. Comput. Sci. Technol. 27 (3) (2012) 635–649, http:// dx.doi.org/10.1007/s11390-012-1250-z

[24] A. Reyes, P. Rosso, "*Making objective decisions from subjective data: detecting irony in customer reviews*", Decis. Support Syst. 53 (2012) 754–760

[25] A. Balahur, Jesús M. Hermida, A. Montoyo, "*Detecting implicit expressions of emotion in text: a comparative analysis*", Decis. Support Syst. 53 (2012) 742–753

[26] H Kang, Yoo Seong Joon, Han Dongil, "*Senti-lexicon and improved Naıve Bayes algorithms for sentiment analysis ofrestaurant reviews*",Expert Syst Appl 2012;39:6000–10

[27] P. Racherla, W. Friske, "*Perceived 'usefulness' of online consumer reviews: an exploratory investigation across three services categories*", Electron. Commer. Res. Appl. 11 (2012) 548–55

[28] A Mudinas, Dell Zhang, and Mark Levene.,"*Combining lexicon and learning based approaches for concept-level sentiment analysis*". In Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, page 5. ACM, 2011

[29] A.C.-R. Tsai, C.-E. Wu, R.T.-H. Tsai, J.Y.-J. Hsu, "*Building a conceptlevel sentiment dictionary based on commons sense knowledge*", IEEE Intell. Syst. 2 (2013) 22–30

[30] Teddy, "*Measuring Political Sentiment on Twitter: Factor Optimal Design for Multinomial Inverse Regression*", Technometrics, 2013, http://dx.doi.org/10. 1080/00401706.2013.778791

[31] E Cambria, P. Gastaldo, F. Bisio, R. Zunino, "*An ELM-based model for affective analogical reasoning*", Neurocomputing 149 (2015) 443–455.

[32] G. Wang et al., *"Sentiment classification: the contribution of ensemble learning*", Decis. Support Syst. (2013), http://dx.doi.org/10.1016/j.dss.2013.08.002

[33] L. Qiu, H. Rui, A. Whinston, "*Social network-embedded prediction markets: the effects of information acquisition and communication on predictions*", Decis. Support Syst. 55 (2013) 978–987

[34] F.H. Khan et al., "*TOM: twitter opinion mining framework using hybrid classification scheme*", Decis. Support Syst. (2013), http://dx.doi.org/10.1016/ j.dss.2013.09.004

[35] Y.M. Li, T.-Y. Li, "*Deriving market intelligence from microblogs", Decis. SupportSys*t. 55 (2013) 206–217

[36] D. Spina, J. Gonzalo, E. Amigó, "*Discovering filter keywords for company name disambiguation in twitter*", Expert Syst. Appl. 40 (2013) 4986–5003

[37] Z.-H. Deng, K.-H. Luo, H.-L. Yu, “*A study of supervised term weighting scheme for sentiment analysis*”, Expert Syst. Appl. 41 (2014) 3506–3513

[38] S. Krishnamoorthy, “*Linguistic features for review helpfulness prediction*”, Expert Syst. Appl. 42 (2015) 3751–375

[39] H. Xu,F Zhang,W Wang ,”*Implicit feature identification in Chinese reviews using explicit topic mining mode*l”, Knowl.-Based Syst. 76 (2015) 166–175

[40] B.Chetashri ,Hardi Dalal, Heenal Doshi, “*Sentiment Analysis: Measuring Opinions*”, International Conference on Advanced Computing Technologies and Applications (ICACTA)(2015)

[41] A aggarwal, Dr. Y. S. Kumaraswamy, "*Opinion Mining Using Decision Tree Based Feature Selection Through Manhattan Hierarchical Cluster Measure*", Journal of Theoretical and Applied Information.Technology, 2013

[42] K. gao Jingfei Du, Hua Xu, “*Box office prediction based on microblog*”, Expert Syst. Appl. (2013), http://dx.doi.org/10.1016/j.eswa.2013.08.065

[43] M. Ott, Y. Choi, C. Cardie, J.T. Hancock, “*Finding deceptive opinion spam by any stretch of the imagination*”, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, 2011, pp. 309–319

[44] J.H. Wang, C.C. Lee, “*Unsupervised opinion phrase extraction and rating in Chinese blog posts*”, in: IEEE Third International Conference Won Privacy, Security, Risk And Trust, 2011 and IEEE Third International Conference on Social Computing (SocialCom), October, IEEE, 2011, pp. 820–823.

[45] J.C.B. Rabelo, R.B.C. Prudêncio, F.A. Barros, “*Using link structure to infer opinions in social networks”*, IEEE International Conference on Systems, Man, and Cybernetics (SMC 2012), IEEE, 2012

[46] B Pang, L Lee, S Vaithyanathan 2002,”*Thumbs up?: sentiment classification using machine learning techniques*”, Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. (2002), 79-86

[47] A. Cui, M. Zhang, Y. Liu, S. Ma, “Emotion Tokens: Bridging the Gap among Multilingual Twitter Sentiment Analysis”, Springer-Verlag, Berlin,Heidelberg, 2011,pp. 238–24

[48] A. Balahur, “*Sentiment Analysis in Social Media Texts*”, 2013, pp. 120–128, Atlanta,Georgia

[49] R Ortegga ,L. García-Moya, H. Anaya-Sánchez, R. Berlanga-Llavori,” *Retrieving product features and opinions from customer reviews*”, IEEE Intell. Syst. 3(2013) 19– 27

[50] L. Barbosa, J. Feng,” *Robust sentiment detection on twitter from biased and noisy data*”, Proceedings of COLING, 2010, pp. 36–44.

[51] Hernandez, S. and Sallis, “ *Sentiment-preserving reduction for social media analysis*”. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. (2011), 409-416

[52] J. Read, “*Using emoticons to reduce dependency in machine learning techniques for sentiment classification*”, Proceedings of the ACL Student Research Workshop, 2005, pp. 43–48.

List of Publications

- Accepted

Neetika Bansal, Ashima Singh, "*A novel technique of sentiment detection for microblogging sites*" IEEE-3rd International Conference on Microelectronics, Circuits and Systems, Micro2016

- Accepted

Neetika Bansal, Ashima Singh, "*A review on Sentiment Analysis using machine learning approaches*" International Conference on Inventive Computation Technologies, ICICT 2016.

Video Link

The link of the video is : https://youtu.be/Tn2aE8y6qGg

\