Analysis on Hinglish Opinion Using Multinomial Naive Bayes Algorithm

Jagmeet Singh[1], Dr. Shashi Bhushan[2]
[1](*Research Scholar, I.T., Chandigarh Engineering College, Landran / Punjab Technical University, India)*
[2](*Head, I.T., Chandigarh Engineering College, Landran / Punjab Technical University, India)*

ABSTRACT: *Sentimental Analysis is the study that analyses people's sentiments, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, the topics, and their attributes. These days, the latest mobile devices and websites are interpreting mashup language based keyboards; this has enabled many users to express their opinions and views about products in 'Hinglish'. This research is focused on conducting sentimental analysis of 'Tweets' written in Hinglish mashup language. The system built here is highly accurate as it is on Multinomial Naïve Bayes algorithm that uses machine learning as basis to classify positive, negative or neutral opinion of products safety. The results obtained using that system has high true positive rate and low false rate as compared to previous Naïve Bayes algorithm.*

Keywords: *Sentimental analysis, Mashup language, Multinomial naïve Bayes*.

I. INTRODUCTION

All over the world large numbers of languages are spoken, even in single country multi languages are spoken. As globalization is taking place due to migration of individuals, it leads to cultural fusion. In India, Hindi is the dominant language, majority of individuals know how to speak, read and write Hindi language. As we all know English is the dominant language of the world. It has large expanse over the earth. It has survived and thrived as it is corpus of words, nouns & verbs from other languages also and is increasing day by day. English language has assimilated in many cultures, societies with a blend or mix of native languages. These languages may be called MASHUP Languages i.e. mixture/fusion of two or more languages.

Today, there is need to research in this direction as no language remains so called native or pure now, especially, when one uses micro blogging websites like Twitter to express their feelings, views, where space/character length is restricted and people try to write and use limited number of words. Typically, one expresses feelings on micro blogging websites about some products, brands, sports, political issues, social issues etc. as routine. In modern world of technology, huge data is posted on these websites, which is to be made useful by drawing out information from that data, otherwise loads of important information could be useless. These expressions are compared for sentimental analysis.

II. ORGANISATION OF PAPER

This paper is organized in following manner, firstly it includes introduction to mashup language importance & sentimental Analysis, and secondly research gap is listed and defines the scope of work. It discusses the issue of extracting and processing sentimental information/data from the social networking websites written in mashup languages. Next is methodology, where the primary purpose of our work has been to do sentimental analysis of product safety in mashup languages. This is demonstrated in that particular section. The related work segment includes several aspects of exploring sentiment data at different levels, automatic generated reviews and spam detection in blogs etc. After the review section (which has tabular summary also). A novel method is proposed to overcome the gaps and complete the scope of work related sentimental analysis. Finally discussion conclusion and limitations are given.

III. LITERATURE REVIEW

Mao K., et.al. [1] This research paper addresses the problem of sentiment analysis for product reviews in different domains. A novel method is proposed by combining lexicon-based and learn-based techniques to analyze sentiment of cross-domain product reviews. In their work, three domain-lexicons are obtained by extracting sentiment words not in the basic lexicon, and four categories of features consisting of 16 features are extracted as the input of classifiers. Besides, the importance of different features is learnt by using the Information Gain (IG) algorithm, and the performance of different classifiers is also studied for each domain. Experimental results show that their domain-lexicons outperform the basic lexicon. Their proposed Combines Lexicon-based and Learn-based technique (CLL) achieves better results than state-of-the-art methods in the domains of books and hotels.

Saif H., et.al. [3] In this paper, they provided an overview of eight publicly available and manually annotated evaluation datasets for Twitter sentiment analysis. They found that unlike the tweet level, very few annotation efforts were spent towards providing datasets for evaluating sentiment classifiers at the entity level. This motivated them to build a new evaluation dataset, Standard Twitter Sentiment Gold (STS-Gold), which allows for the evaluation of sentiment classification models at both the entity and the tweet levels. They also provided a comparative study across all the reported datasets in terms of different characteristics including the vocabulary size, the total number of tweets and the degree of sparsity. They studied the various pair-wise correlations among these characteristics as well as the correlation between the data sparsity degree and the sentiment classification performance across the datasets. Their study showed that the large number of tweets in a dataset is not always an indication for a large vocabulary size although the correlation between these two characteristics is relatively strong. They also showed that the sparsity-performance correlation is intrinsic, where it might exists within the dataset itself, but not necessarily across the datasets.

Xing Fang and Justin Zhan, [2] in this paper they tackle the problem of sentiment polarity categorization, which is one of the fundamental problems of sentiment analysis. A general process for sentiment polarity categorization is proposed with detailed process descriptions. Data used in this study are online product reviews collected from Amazon.com. Experiments for both sentence-level categorization and review-level categorization are performed with promising outcomes. Sentiment analysis or opinion mining is a field of study that analyzes people's sentiments, attitudes, or emotions towards certain entities. Online product reviews from Amazon.com are selected as data used for this study. A sentiment polarity categorization process has been proposed along with detailed descriptions of each step. Experiments for both sentence-level categorization and review-level categorization has been performed.

Salathé M., et.al. [5] In this paper the main interest is in identifying the extent to which social contagion and homophile drive sentiment dynamics within the social network. They use the term social contagion to mean the extent to which exposure to a given sentiment is predictive of future expression of that sentiment. Previous studies have focused on binary outcomes such as the adoption (*vs*. non-adoption) of a service, and have measured exposure as the number of social contacts that have adopted the service previously. Their methodology allows considering more complex measures of exposure as they measure both the number of social contacts expressing a given sentiment as well as the intensity with which the sentiment is expressed.

Silla Jr C., et.al.[4] In this paper, they present the Latin Music Mood Database, an extension of the Latin Music Database(LMD) but for the task of music mood/emotion classification. The method for assigning mood labels to the musical recordings is based on the knowledge of a professionally trained Brazilian musician and the identification of the predominant emotion perceived in each song. They present an analysis of the mood distribution according to the different genres of the database. The Latin Music Mood Database which is an extended version of the LMD but with one emotion label, representing the predominant emotion perceived in each song of the LMD. Their data analysis of this novel database has shown that there is not a clear one genre equal to one type of emotion/mood as people might think at first. However, it was clear that the majority of the music genres present in the LMD had at least 25 % of

their songs about either love or passion. They believe that this novel Latin Music Mood Database might prove useful for the Music Information Retrieval research community.

Table I the following table gives information regarding techniques and findings of all the papers mentioned above

NUMBER	NAMES	DATASET	PRODUCT	TECHNIQUE	FINDINGS
1	Kaili Mao, Jianwei Niu , et.al	twitter.com	Chinese product i.e. hotel, books, electronic	CLL technique	Hotels - 85.8 Books – 92.9 Elec. – 81.6
2	Xing Fang and Justin Zhan	amazon.com	Beauty, Books, Electronic, Home	Naive bayes classifier, Random forest, Svm	Sentence level 85% , review level 73%
3	Hassan Saif, Miriam Fernandez, et. al.	twitter.com	Data set evaluation	Descriptive statistics	Best accuracy in STS GOLD dataset 85%
4	Marcel Salathé, Shashank Khandelwal, et.al.	twitter.com	Health behavior	Covariance analysis	---
5	Carolina L. dos Santos and Carlos N. Silla Jr	Latin music mode database	Emotions i.e. joy, love, passion, sad, excitement	Descriptive statistics	At least 25% songs are about love/passion

IV. RESEARCH GAPS

Limited work has been reported in contemporary research studies related opinion mining specially, in context of languages like ‘Hinglish’ which is mashup languages. The context written in ‘Hinglish’ has not been used to find polarity classification of branded products as such. In this field of work limited efforts has been made especially when it comes to machine learning algorithm for classification of product’s reviews/opinions. In this tweets written in Hinglish (Hindi written in English) can also be extracted to gain valuable information of branded products. With it, sentimental analysis of product can be done efficiently; which tells what does people think about it, what can be added or removed to a product? So that product/service providers may improve overall.

V. SCOPE OF WORK

Our work is limited to development of the framework which grabs the polarity of the products and its opinions, annotate, reviews mentioned in mashup languages such as Hinglish which is mixture of Hindi and English. In this polarity is found based on features related to frequency, entity relationship count and co-occurrence. Once these features are extracted machine learning is used to train the system to each feature row labeled as positive, negative or neutral for the service product. This scope of work has been delimited after conducting a systematic survey on challenges of developing frameworks to do sentimental analysis of mashup languages like Hinglish.

VI. HINGLISH TWEET

Table II The following table contains Hinglish parallel list of tweets on mobile phones.

Serial no.	User id	Tweet text	Positive	Negative	Neutral
1	lovudsharma ?@lovudsharm	Samsung phones drop test mein iphones se teen ghuna jyada survive karte hain	Yes		
2	lovudsharma ?@lovudsharma	Currently mere paas #Samsung Galaxy s6 and love it. #Samsung ki battery life iphone ki battery life se 2 times jyaada hai.	Yes		
3	ENGRAVER SETHI @EngraverSethi @lovudsharma	htc sirraa samsung de phone ghatiyyaa		Yes	
4	Ramandeep Kaur ?@ramandkaur8917	#Samsung galaxy note 5 India ka sabse mehnga aur bakwaas android phone hai.		Yes	
5	@NeerajDhaked	Ab pta chala @SamsungMobileIN samsung ke sare mobile hi ghatiya hein lelo.		Yes	
6	Ramandeep Kaur ?@ramandkaur8917	Mujhe yeh samjh main nahi aata ke #Samsung ke sabhi mobiles mein ek hi jaise features kyu hote hain. Rs.45000 and Rs. 25000 are same	Yes		
7	Navneet Bhullar ?@NavyBhullar	#IPhone da touch sab toh vdiaa hai	Yes		Yes
8	@NavyBhullar @lovudsharma	# I think HTC and samsung phones are quite similar bcoz maine dono use kiye hain..mujhe dono phones pasand hain			Yes
9	@lovudsharma@ EngraverSethi	Naye #Samsung smartphones and #iphone apni lokpriyata sirf removable battery naa hone ki wajha se kho rahe hai.		Yes	

10	@ap_pune	acha hai na ...means some one doing your marketing ... Its like I don't like apple phone doesn't mean its bad			Yes
11	@NaveenxAsad	i hate window phone tbh. Apple acha hai 5s but not like android	Yes		
12	@OyeSaaann @Mahnoor_Agha	*die hard apple boy* "android sucks! iOS for life!"			Yes
13	@Hinachki Kabachki ? @KabachiTweets	Apple bhi kamal hai. Har phone se ek acha feature utha k apne phone mai daal k aisay baichte hain jaise abhi revolutionary idea laye hon!	Yes		
14	Sir? @CriminalSingh	Mard ke hanth mein girly thing acha nahi lagta bhai __ i6 is better __ enjoy now Paisa hanth ka mayel hai	Yes		
15	@mana_apple	apple i meant "mana ki apple phones ache hai, par cost bi bahut jyada hai :P	Yes		
16	@Spamsterr	I Phone me no better app only other than official one :| Baaki ache waale sab paid			Yes
17	lovudsharma ?@lovudsharm	ZTE Axon Smartphone Antimicrobial Corning Gorilla Glass use karne wala duniya ka pehla phone hai	Yes		
18	@tamoghno	NOPE... Bekaar phone. Nothing exciting for me. Apple products are not my cup of tea.		Yes	
19	Elephant Man? @ghantahaathi	Bakwaas na kar-iPhones are the best. If I had an android, I'd be banging my head on the wall in frustration every day :p	Yes		

20	@iReenKaur	bakwaas kyun? I'm not into iphones. iphone 6 plus is too wide for my taste.		Yes	

VII. METHODOLOGY

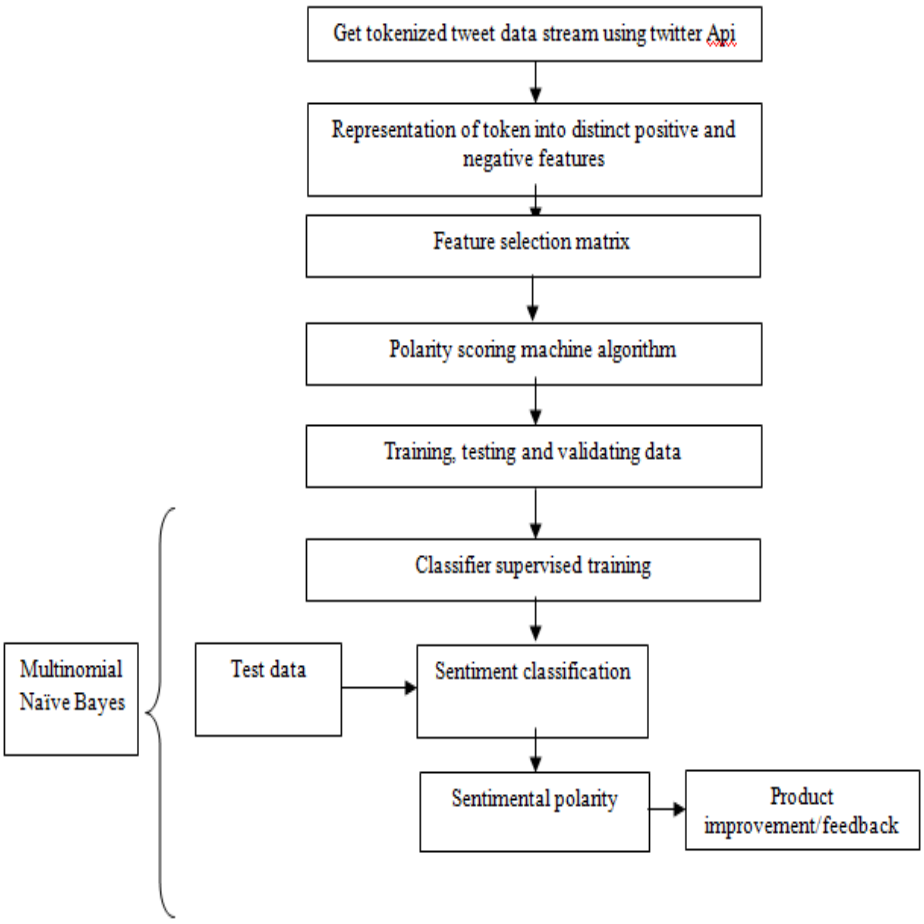

Fig 1: Flow chart showing the methodology

STEP 1 - The first step after extracting tweets using Twitter API of products/brands. We need to conduct tokenization, which is a process of splitting the strings into its desired constituent parts which is fundamental to the natural language processing. The process of tokenization is fine timed to first remove unwanted words and allow highly relevant tokens related to the product under opinion mining observation. Our tokenizer is 'sentiment aware' tokenizer. It captures emotions, as well words expressing emotions written in mashup language (Hindi & English).The tokenization also includes the topic and user markup defined by twitter e.g. usernames [@+ (\W_)]+ and hash topics etc.

Basic html markups like strong, b, em and 'i' tags can clean indicators of sentiments, our algorithm incorporates the sentiment impact by including these as tokens to be analyzed for sentiment/opinion analysis. Idioms written in Hinglish/English and multiword expression are also taken care of.

STEP 2 - In this step, we intend to bring the tokens into 'distinct word' forms. This helps to reduce the vocabulary size and helps in sharpening the results, especially in case, if we are using machine learning algorithm. Here, we had to build porter stemmer that also takes care of Hinglish word forms and distinction. The stemmer has two parts a positive distinction and negative distinction.

STEP 3 - In this step the feature rows are build in Hinglish sentiment tokenizer, Hinglish stemmer characterization, frequency count and relates extraction count values.

STEP 4 - In this step the features matrix is built as per requirement of the machine learning algorithm. Since, we are using weka [15] api to build our system using Multinomial naïve bays we organize our data set groups as follows

Table III contains brief information for table 4 & 5 and their meaning.

Serial no.	Table	Meaning
1	4	Positive
2	5	Negative

Once this supervised labeling is compiled it is time to subject the matrix to the process of learning which is explained in next step

VIII. RESULTS

In the section we will present the outcomes of the series of tests done to validate results against ground truth. These results try to evaluate and validate as per the objectives mentioned earlier and also tries to address the validity of the hypothesis taken for conducting this research. There are four possible outcomes of the research work and if we were to evaluate all these four, we first need to conduct an exercise that would design experiment to cover all these four outcomes. As per design of experiment full factorial method, we have parameters and four possible outcomes for evaluation.

Table IV includes condition name and condition and its definition of sentiment for positive sentiments.

Serial no.	Condition name	Conditions	Definition of sentiment/opinion
1	True positive	Positive opinion correctly selected	The person speaks positive opinion on product and is considered correctly.
2	False positive	Positive opinion incorrectly selected	The person speaks positive opinion on product and is considered incorrectly.
3	True negative	Positive opinion correctly rejected	The person speaks positive opinion on product and is rejected correctly
4	False negative	Positive opinion incorrectly rejected	The person speaks positive opinion on product and is rejected incorrectly

Table V: It includes condition name and condition and its definition of sentiment for negative sentiments.

Serial no.	Condition name	Conditions	Definition of sentiment/opinion
1	True positive	Negative opinion correctly selected	The person speaks negative opinion on product and is considered correctly.

2	False positive	Negative opinion incorrectly selected	The person speaks negative opinion on product and is considered correctly.
3	True negative	Negative opinion correctly rejected	The person speaks negative opinion on product and is rejected correctly.
4	False negative	Negative opinion incorrectly rejected	The person speaks negative opinion on product and is rejected incorrectly

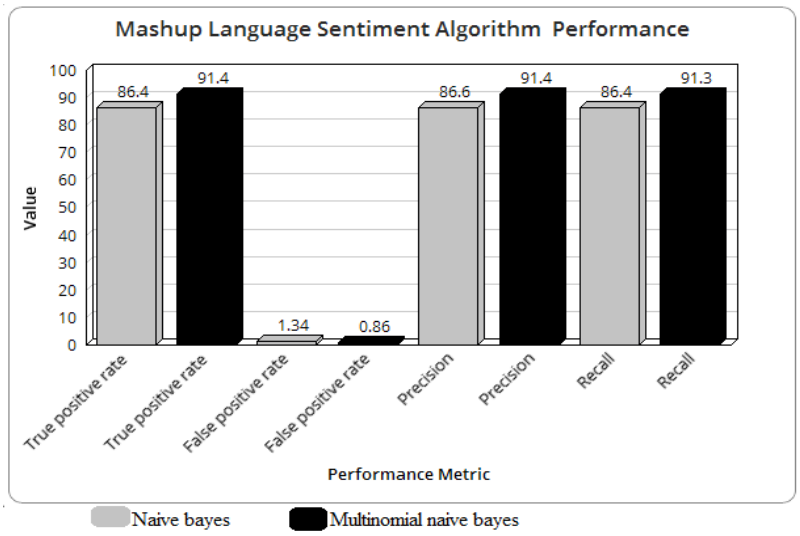

Fig 2: Bar Graph

Following are the presentations of the above four outcomes in terms of True positive rate, False positive rate, precision, Recall for previous Naïve Bayes and our Multinomial Naïve Bayes algorithms.

True Positive Rate: True positive rate is an indicator that our algorithm is correctly identifying and predicting values from actual values.

$$TP = \frac{d}{c+d} \quad \text{…equation (1)}$$

Where TP is true positive rate

d is number of correct predictions that an instance is positive.

c is number of incorrect predictions that an instance is positive.
High positive rate indicates that classifier is capable of predicting objects that are very near to the ground truth criteria. It is apparent from the above bar graph{1} that Multinomial Naïve Bayes algorithm is producing better true positive rate indicating that Multinomial Naïve Bayes algorithm is better as compared to Naïve bayes. It is calculated using equation (1). In this positive sentiments are correctly identified as positive. Hence we can say that Multinomial Naïve Bayes gives better true positive rate of positive sentiments.

False Positive Rate: False positive rate is the proportion of negative cases that were incorrectly identified as positive calculated with the equation given below

$$FP = \frac{b}{a+b} \quad \text{…equation (2)}$$

Where FP is false positive rate

b is incorrect number of prediction that an instance is negative

a is correct number of prediction that an instance is negative

False positive rate is calculated using equation (2). The high value of false positive rate indicates that algorithm is making large number of incorrect predictions. Hence, is not reliable.False positive rate in Multinomial Naïve Bayes algorithm values lower as compared to Naïve Bayes algorithm which means Multinomial Naïve Bayes is better algorithm. In this negative sentiments are classified as positive sentiment which is incorrect. As shown in the graph {1} it is clear that false positive rate of sentiment is less in Multinomial Naïve Bayes algorithm as compared to Naïve Bayes. Hence, our algorithm is clearly better.

Precision: It is the ratio of number of relevant records retrieved from the database from matching to the numbers of relevant and irrelevant records.

$$P = \frac{A}{A+C} * 100 \quad \text{…equation (3)}$$

P is precision

A is number of relevant record retrieved

C is number of irrelevant record retrieved

Closer the value of precision to one or 100 percent, less is the irrelevant proportion which means the classifier is able to filter irrelevant information more accurately for prediction with a match for actual class. Precision is calculated using equation (3). From the bar graph {1} it is clear that Multinomial Naïve Bayes algorithm gives better precision than Naïve Bayes algorithm.

Recall: It is the ratio of number of relevant records retrieved to the total number of records retrieved in a full database.

$$RECALL = \frac{A}{B} * 100 \quad \text{…equation (4)}$$

A is number of relevant records retrieved

B is number of relevant records not retrieved

Recall is calculated using equation (4). It is apparent from above bar graph {1} that recall value of Multinomial Naïve Bayes algorithm shows appropriate ratio as compared to previous Naïve bayes algorithm. Therefore, our algorithm is better.

IX. DISCUSSION

It is clear from the implementation steps that the objectives of improving accuracy and its successful usage in mashup language in Hinglish requires special knowledge on morphology and phrasal of sentences in Hinglish. Then due to non standardized state of affairs such mashup languages, it's difficult to really do polarity classification. The degree of contact and transmission between the languages determines that how much fussed the Hinglish is.
The education makes a difference, if the people are more educated they intend to use correct Hindi, English and their level of fusion is less, and if the person is less formally educated, there is dominance of Hindi in Hinglish. Therefore

this research focuses on very specific domain rather work on cross domain. Our research might have come under influence of cultural imperialism, but we have tried to be more objective than subjective in this context. The result shows the basic sentiment/opinion analysis is executed successfully.

X. CONCLUSSION

The syntax and the phonology of fussed/mashup are not standard in anyway. But routine discourse between common people across cultures, nations consist of mixed mashup vocabulary. Modern socio linguistics distinguish English in manner that it has including words of other languages due to which it's global very popular. Our research shows that people across the world are using these mashup languages especially on social websites. Results show that our approach is better to previous algorithm in all four aspects i.e. True positive rate, false positive rate, Precision, Recall.

XI. LIMITATIONS & FUTURE SCOPE

No research is without limitations, even if we have large pool of resources while conducting the research work. It was found few public repositories of tweets related to tweets are available for 'benchmarking' the algorithm in this context. Due to paucity of time, we are not able to extract large corpus in it. Other than this small pool of expert related safety we available for truly reflect finding on product safety opinion. The current research work is using machine learning algorithm with supervisor learning and corpus has not under gone review of safety experts to judge positivity or negativity of the product. Hence fore said future scope, we suggest the use of human experts should be done on small scale and along with automatic feature extraction algorithm, a hybrid approach may be more realistic and accurate in future.

REFERENCES

[1] Kaili Mao, Jianwei Niu, Xuejiao Wang, Lei Wang, Meikang Qiu, "Cross-Domain Sentiment Analysis of Product Reviews by Combining Lexicon-based and Learn-based Techniques" IEEE 17th International Conference on High Performance Computing and Communications, 2015.
[2] Hassan Saif1, Miriam Fernandez, Yulan He and Harith Alani, "Evaluation Datasets for Twitter Sentiment Analysis A survey and a new dataset, the STS-Gold" The Open University's repository of research publications, 2013.
[3] Xing Fang and Justin Zhan," Sentiment analysis using product review data" springer open Journal of Big Data, 2015.
[4] Carolina L. dos Santos and Carlos N. Silla Jr," The Latin Music Mood Database" Springer Journal EURASIP Journal on Audio, Speech, and Music Processing, 2015.
[5] Marcel Salathé, Duy Q Vu, Shashank Khandelwal and David R Hunter, "The dynamics of health behavior sentiments on a large online social network" license Springer, 2013.
[6] Haruna Isah, Paul Trundle, Daniel Neagu, " Social Media Analysis for Product Safety using Text Mining and Sentiment Analysis" IEEE, Artificial Intelligence Research (AIRe), 2014.
[7] Kaipeng Liu, Binxing Fang, Yu Zhang, "Detecting Tag Spam in Social Tagging Systems with Collaborative Knowledge**"** IEEE, Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009.
[8] Xindong Wu, Xingquan Zhu, Gong-Qing Wu,and Wei Ding, "Data Mining with Big Data" IEEE, Transactions on Knowledge and Data Engineering, 2014.
[9] Joseph S. Kong, Behnam A. Rezaei, Nima Sarshar, and Vwani P. Roychowdhury, "Collaborative Spam Filtering Using E-Mail Networks" IEEE Computer Society, 2006.
[10] Leon French, Po Liu, Olivia Marais, Tianna Koreman, Lucia Tseng, Artemis Lai and Paul Pavlidis, "Text mining for neuroanatomy using WhiteText with an updated corpus and a new web application" University of Illinois, USC Information Sciences Institutes, USA, May 2015.
[11] Vaishak Suresh,Syeda Roohi, Magdalini Eirinaki, "Aspect-Based Opinion Mining and Recommendation System for Restaurant Reviews**"** San Jose or vicinity, CA, USA, 2014.
[12] Anirban Mukhopadhyay, Ujjwal Maulik, Sanghamitra Bandyopadhyay and Carlos A. Coello. "Survey of Multiobjective Evolutionary Algorithms for Data Mining" IEEE Transactions on Evolutionary Computation, 2014.
[13] Yuchun Tang Sven Krasser Paul Judge, Yan-Qing Zhang, "Fast and Effective Spai Sender Detection with Granular SVM on Highly Imbalanced Mail Server Behavior Data ", IEEE, 2006.
[14] Fatos Xhafas Leonard Barolli, "Semantics, intelligent processing and services for big data", Future Generation Computer Systems, 2014.
[15] Mark Hall Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer Peter Reutemann, Ian H. Witten, "The WEKA Data Mining Software: An Update",SIGKDD Explorations.