Social media sentiment analysis for business analytics

[1]Dr A Jammer Bhasha, [2]Athira R S

[1]Head of Department, [2]Student
Department of Computer Science & Engineering,
Hindusthan Institute of Technology, Coimbatore, India

***Abstract*: Social media plays a vital role in generating valuable data as a data source for obtaining inferences and building strategies for business marketing. The vast amount of data produced every day in social media platforms are treasures for data analytics as they can be wisely used to understand the attitude of people regarding specific products. Sentiment analysis performed on these data provides valuable business insights. Organizations tends to use more social media for getting insight into consumer behavioral tendencies, market intelligence and present an opportunity to learn about customer review and perceptions.**

Applications like Twitter, Yelp and Amazon offers organizations a fast and effective way to analyze customers' perspectives toward the critical to success in the market place. Building a system for sentiment analysis is an approach to be used to computationally measure customers' perceptions. This paper reports on the design of a sentiment analysis app by extracting a vast amount of tweets, fetching user reviews from amazon and yelp. An e-commerce prototype is also used for customer behavior prediction. Fake reviews are detected using machine learning techniques. Results classify customers' perspective via tweets into positive and negative, which is represented in a pie chart, histogram, data tables and html page. Reviews are analyzed and categorized into fake reviews and real reviews.

***Index Terms*: Social media analysis, Sentiment analysis, Business analytics, fake review detection, Amazon review analysis, Twitter sentiment analysis.**

I. INTRODUCTION

Sentimental analysis means to analyze the views and interest of people regarding celebrity, politicians or some other topic. Sentiment analysis is to classify the mood into different category like positive, negative and neutral.

Amazon reviews, Yelp reviews, twitter like social media are the areas where the tweets, reviews and comments of different users produce larger amount of data and convey the information in the form of unstructured type of data. It is difficult to understand and extract information from this data. So, we need such tools and Technologies that can efficiently store and process unstructured and big data. There are different techniques and tools are available that can handle this type of data and produce meaningful information.

Sentimental analysis of social media such as twitter is quite difficult as compared to other general sentimental analysis due to the presence of slang words, misspelling in comments & tweets, short length, and some graphics type words etc. So, R statistical computing language is used to perform sentimental analysis easily.

Online shopping sites like amazon, Flipkart allows their users to write reviews about product which can later be read by other users for making decision on whether to buy or not buy a product. Sometimes it is impossible for users to make decision on huge amount of review data. Hence a sophisticated tool is required to process huge amount of review and make inference from it. We have developed the system by keeping these things in mind.

In certain cases reviews may get faked by online spammers who are employed by firms to defame or promote some specific products. Such activities called online opinion spamming carried out by opinion spammers always lead wrong analysis and decision making and let customers to make mistakes in selecting products. Our research work also focuses on identifying spam reviews and make the right kind of business analysis.

This paper is divided into five (5) sections. Following this sec-tion is Section II which provides Background Information and Related Work of the following topics: Sentiment Analysis, Fake review detection and Amazon review analysis. Section III shows the approach that we have used for performing the procedure. Section IV demonstrates the Experimental Setup combined with the Results and Discussion and Section V presents the Conclusion and Future Work.

II. BACKGROUND AND RELATED WORK

Sentiment Analysis

Sentiment Analysis is used by companies to analyse data to under-stand the users' sentiments or opinions regarding their products or services. According to [5], Sentiment Analysis is "a process that automates the mining of attitude, opinions, views, and emotions from text, speech, tweets, and database sources through Natural Language Processing (NLP)". It is also reffered to as opinion mining and emotion analysis. It uses textual data to analytically collect, analyse, model and validate for various business intelligence applications.

Social Media

Social media as a group of Internet-based applications that create on the ideological and technological foundations of Web2.0 which is allowed to build and exchange of user generated contents. In a discussion of Internet World Start,[16] identified that a trend of inter-net users is increasing and continuing to spend more time with social media by the total time spent on mobile devices and social media

. On the other hand, businesses use social networking sites to find and communicate with clients, business can be demonstrated dam-age to productivity caused by social networking [17]. As social media can be posted so easily to the public, it can harm private information to spread out in the social world [11].

On the contrary, [18] discussed that the benefits of participating in social media have gone beyond simply social sharing to build organization's reputation and bring in career opportunities and monetary income. In addition, [15], [5] mentioned that the social media is also being used for advertisement by companies for promotions, professionals for searching, recruiting, social learning online and electronic commerce.[19]

Twitter Sentiment Analysis

Social media, such has Twitter which is convenient due to its 24-hours availability, ease of customer service and global reach. The sentiment can be found in the comments or tweet to provide useful indicators for many different purposes [20].Also, [12] and [16] stated that a sentiment can be categorized into two groups, which is negative and positive words. Sentiment analysis is a natural language processing techniques to quantify an expressed opinion or sentiment within a selection of tweets [8].Hence it helps in making valuable inferences.

Amazon Review Analysis

Amazon is one of the largest E-commerce site as for that there are in numerous amount of reviews that can be seen. A study on amazon last year revealed over 88% of online shoppers trust reviews as much as personal recommendations. Any online item with large amount of positive reviews provides a powerful comment of the legitimacy of the item. Conversely, books, or any other online item, without re-views puts potential prospects in a state of distrust. Quite simply, more reviews look more convincing. People value the consent and experience of others and the review on a material is the only way to understand others impression on the product.

Opinions, collected from users' experiences regarding specific products or topics, straightforwardly influence future customer purchase decisions [1]. Similarly, negative reviews often cause sales loss [2]. For those understanding the feedback of customers and polarizing accordingly over a large amount of data is the goal.

The system performs opinion mining over a dataset of Amazon product reviews to understand the polarized attitudes towards the products.

Fake Review Detection

As reviews are concerned, positive reviews play a stimulative role in reaping economic benefits and well-deserved reputation for merchants' businesses. Thus, it makes merchants have strong intentions to manoeuvre their fame and employ specialized imposters posting higher opinions on sites. Besides, there exists competition between online merchants. Consequently, the employed fraudsters may post negative opinions to defame their rivals, resulting in bad sales of products and services. Such individuals are called opinion spammers, and their behaviour is called opinion spamming [11]. Because of such activities, the online customers might be misled by the deceptive opinions. Therefore, opinion spam has attracted important attention from both business and research circles. The system aim to identify fake reviews and let online opinions recover reliability and facticity.

III. PROPOSED APPROACH

The In our approach we used the online datasets and analyzed it. This analyses labelled datasets using the unigram feature extraction technique. The pre-processor is applied to the raw sentences which make it more appropriate to understand. Further, the different ma-chine learning techniques trains the dataset with feature vectors and then the semantic analysis offers a large set of synonyms and similarity which provides the polarity of the content.

Data Extraction

An access token, which is used to get authentication access to ex-tract data from the Twitter database, is given upon logging into a Twitter account. Authorization details that include the API key are required to establish the connection and allow a search query. An example set of Twitter is generated when setting the parameters for the Search Twitter operator. Search parameters returned tweet and language limit of 500 English tweets, and the result type of recent or popular tweets. This contains attribute and label types including tweet ID, username, number of retweets, original text, date and time it was created, language, and many others. The amazon data is extracted by fetching page data using product code. The spam dataset is saved in an MS Excel file and is readily available online. A CSV file of Yelp data set used for fake review detection is also obtained from yelp fake data challenge page.

Pre-processing of the datasets

The data contain a lot of opinions about the data which are ex-pressed in different ways by individuals .The twitter, amazon, yelp and spam dataset used in this work is already labelled. Labelled dataset has a negative and positive polarity and thus the analysis of the data becomes easy. The raw data having polarity is highly susceptible to inconsistency and redundancy. The quality of the data

affects the results and therefore in order to improve the quality, the raw data is pre-processed. It deals with the preparation that removes the repeated words and punctuations and improves the efficiency the data.

For example, "that painting is Beauuuutifull #" after preprocessing converts to "painting Beautiful." Similarly,

"@Geet is Noww Hardworkingg" converts to "Geet now hardworking".

Feature Extraction

The improved dataset after pre-processing has a lot of distinctive properties. The feature extraction method, extracts the aspect (adjec-tive) from the dataset. Later this adjective is used to show the posi-tive and negative polarity in a sentence which is useful for determin-ing the opinion of the individuals using unigram model [15]. Uni-gram model extracts the adjective and segregates it. It discards the preceding and successive word occurring with the adjective in the sentences.

For above example, i.e. "painting Beautiful" through unigram model, only Beautiful is extracted from the sentence.

Tokenization: It is the process of separating a sequence of strings into individuals such as words, keywords, phrases, symbols and other elements known as tokens. Tokens can be

Individual words, phrases or even whole sentences. In the process of tokenization, some characters like punctuation marks are discarded. The tokens work as the input for different process like parsing and text mining.

Removing Stop Words: Stop words are those objects in a sentence which are not necessary in any sector in text mining. So we generally ignore these words to enhance the accuracy of the analysis. In different format there are different stop words depending on the country, language etc. In English format there are several stop words.

POS tagging: The process of --assigning one of the parts of speech to the given word is called Parts of Speech tagging. It is generally referred to as POS tagging. Parts of speech generally contain nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories. Parts of Speech tagger or POS tagger is a program that does this job.

Bag of Words: Bag of word is a process of extracting features by representing simplified text or data, used in natural language processing and information retrieval. In this model, a text or a document is represented as the bag (multiple set) of its words.

So, simply bag of words in sentiment analysis is creating a list of useful words. We have used bag of words approach to extract our feature sets. After pre-processed dataset we used pos tagging to separate different parts of speech and from that we select nouns and adjectives and use those to create a bag of words.

Then we run it through a supervised learning and find our results and also the top used words from the review dataset.

TF-IDF is an information retrieval technique which weighs a term's frequency (TF) and also inverse document frequency (IDF). Each word or term has its own TF and IDF score. The TF and IDF product scores of a term is referred to the TF*IDF weight of that term. Simply we can state that the higher the TF*IDF score (weight) the rarer the term and vice versa. TF of a word is the frequency of a word.

IDF of a word is the measure of how significant that term is throughout the corpus. When words do have high TF*IDF weight in content, content will always be amongst the top search results, so anyone can:

1. Stop worrying about using the stop-words,
2. Successfully find words

Readability Features

Automated Readability Index (ARI) and Coleman-Liau Index (CLI) to evaluate content quality and helpfulness of online reviews [20]. It is highly significant that consumers can understand the con-tent of the review. The greater the readability of reviews' content, the better a customer feels. What's more, ARI and CLI indicate one's level of education, the lower the education, the higher and the readability of the reviews' content. Next, ARI and CLI are defined as follows.

Automated Readability Index (ARI). Automated

Readability Index is a good indicator of the readability of an

English text. In order to calculate the ARI for a given review, we first calculated the total number of characters (excluding standard syntax such as hyphens and semicolons) and the total number of words. Then we need to calculate the review length. The following formula presents the Automated Readability Index:

$$ARI = 4.71 \times \left(\frac{characters}{words}\right) + 0.5 \times \left(\frac{words}{sentence}\right) - 21.43.$$

Coleman-Liau Index (CLI). The Coleman-Liau Index is similar to the Automated Readability Index, the only difference is that the second formula considers a more careful selection of the textual characteristics of the piece of text assessed. The following formula describes the Coleman– Liau index:

$$CLI = 5.89 \times \left(\frac{characters}{words}\right) + 0.5 \times \left(\frac{sentences}{word}\right) - 15.8.$$

We define AC by the following formula:

AC=ARI/CLI

We first attempted to use ARI and CLI as two features to identify fake reviews. The performance is well in some sense.

Experiments show that AC also have ability to make a contribution to the result of classification, and the performance is better than ARI and CLI. Then AC is used as a new feature. Thus we consider the above three indicators as topic features.

We use RF (Readability Features) to represent a feature set, including three features of ALI, CLI and AC.

Behavior Features

This part studies some behavioral features of fakers. For deeper analysis, we separate reviewers in our data into two groups, one is spammers (authors of fake reviews) group, and the other is non-spammers (authors of non-fake reviews) group. The spammers post fake reviews, while the non-spammers product truthful reviews. We analyze the reviewers' profile with the following behavioral dimensions.

Percentage of Positive Reviews (PR). According to the research in [17], we use 4-5 star reviews as positive reviews.

As per [13] the CDF of percentage of positive reviews in all reviews for spammers and non-spammers in, only 5% of the spammers have less than 80% of their reviews as positive, in other words, a majority (95%) of spammers rated higher than 90% of their reviews as positive. Compared with spammers, non-spammers show a rather evenly distributed trend where each kind of reviewers basically have different percentage of positive reviews. This is reasonable as in real-life, real reviewers generally have different rating levels.

Review Length (RL). Writing fake reviews requires some experience, so there is probably not much to write or a spammer may not want to spend too much time in writing. The CDF of the average number of words per review for all reviewers is calculated in [13] A majority of spammers are bounded by 135 words in average review length which is a bit short as compared to non-spammers.

Training and classification

Supervised learning is an important technique for solving classification problems. In this work too, we applied various supervised techniques to get the desired result for sentiment analysis. In next few paragraphs we have briefly discussed about the three supervised techniques i.e. naïve Bayes, maximum entropy and support vector machine followed by the semantic analysis which was used along with all three techniques to compute the similarity.

Naive Bayes

It has been used because of its simplicity in both during training and classifying stage. It is a probabilistic classifier and can learn the pat-tern of examining a set of documents that has been categorized. It compare the contents with the list of words to classify the documents to their right category [16].

C* = argmaccPNB(c|d)

Class c* is assigned to tweet d, where, f represents a feature and ni(d) represents the count of feature fi found in tweet d.

There are a total of m features. Parameters P(c) and P (f|c) are obtained through maximum likelihood estimates which are incremented by one for smoothing.

Pre-processed data along with extracted feature is provided as input for training the classifier using naïve bayes. Once the training is complete, during classification it provides the polarity of the sentiments. For example for the review comment "I am happy' it provide Positive polarity as result."

Support vector machine

Support vector machine analyses the data, define the decision boundaries and uses the kernels for computation which are per-formed in input space. The input data are two sets of vectors of size m each. Then every data represented as a vector is classified in a particular class. Now the task is to find a margin between two classes that is far from any document. The distance defines the margin of the classifier, maximizing the margin reduces indecisive decisions. SVM also supports classification and regression which are useful for statistical learning theory and it helps recognizing the factors precisely, that needs to be taken into account, to understand it success-fully [18].

Semantic Analysis

After the training and classification we used semantic analysis. Se-mantic analysis is derived from the WordNet database where each term is associated with each other. If two words are close to each other, they are semantically similar.

More specifically, we are able to determine synonym like similarity. We map terms and examine their relationship in the ontology. The key task is to use the stored documents that contain terms and then check the similarity with the words that the user uses in their sentences.

Apply Model

The Apply Model operator applies a model on an example set. In this experiment, a new real time data set of 0-1000 tweets was extracted, pre-processed, and feed into the model to predict the polarity of each tweet. This example set is processed to be compatible with the model with the same attributes that were used to generate the model. Similar way amazon review is analyzed using product code, and spam review is analyzed using user input text data.

IV. IMPLEMENTATION AND RESULTS

Now we evaluate the effectiveness of the proposed model. We conduct experiments on a dataset of Twitter, Amazon, Yelp , User text input and report our findings.

Data

The initial dataset of yelp is comprised of 66887 reviews created by 35102 reviewers. To facilitate experiments, we sample this dataset to acquire a smaller and more easily evaluated dataset. In addition, to obtain a higher accuracy, we ensure the number of reviewers instead of reviews is basicly balanced by employing the technique of under sampling. If we ensure that the number of reviews is balanced, the number of reviewers will be extremely uneven. Besides, we also manually removed some reviews considering the feature value should be meaningful. Our final dataset is consist of 33681 reviews and 10463 reviewers. For Amazon, Twitter and user text spam detection, we have used real time data.

Software Specification

R Language and RStudio is used to develop the application and make statistical evaluations on twitter, amazon data sets. Python is used for analysis on yelp data. Rapidminer is a data-driven data mining tool that discovers patterns in large collections of text [21]. It is an analytic platform that integrates machine learning and predictive model deployment used by data scientists. It contains rich libraries of data science and machine learning algorithms [22]. The software used in this experiment is the latest version - RapidMiner Studio Educational version 8.0.001.

A PHP-MySql based web application is developed and used as an e-commerce web app for recording and storing user patterns and users behavioural features.

Hardware Specification

The experiment was developed using the following hardware specifications:
Processor: Intel Core i5 8th Gen@3.4GHz
RAM: 4 GB (3.98 GB usable)
System Type: 64-bit Operating System

Twitter sentiment results

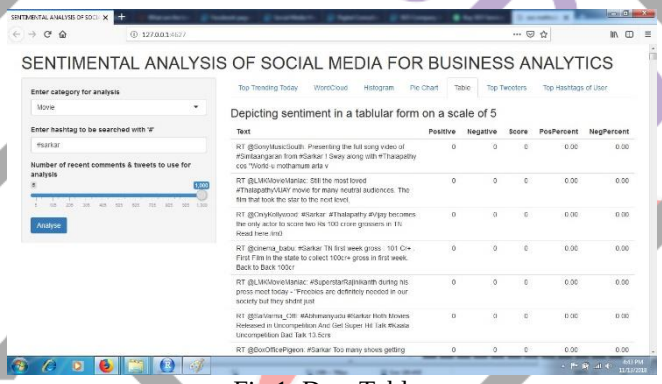

Fig 1, Data Tables

Figure 1 shows the tweets extracted using twitter API and represented using Rshiny interface.

Word cloud

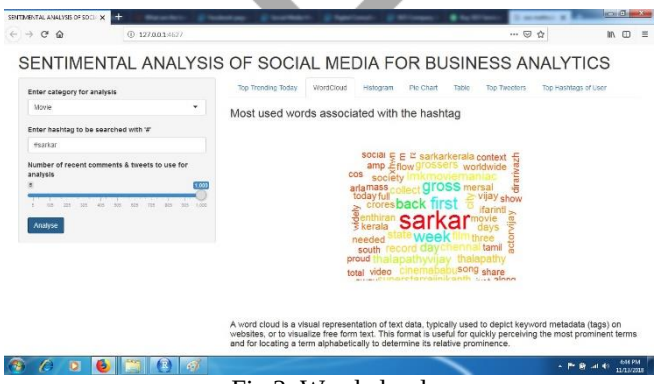

Fig 2, Word cloud

Figure 2 shows the word cloud of the tweets extracted using twitter API and represented using Rshiny interface.

Pie chart:

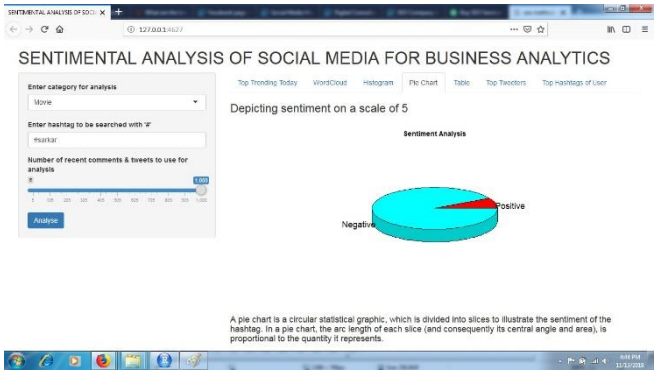

Fig 3, Pie Chart

Figure 3 shows the pie chart of polarity, the positive tweets and negative tweets are represented through Rshiny interface.

Amazon review results

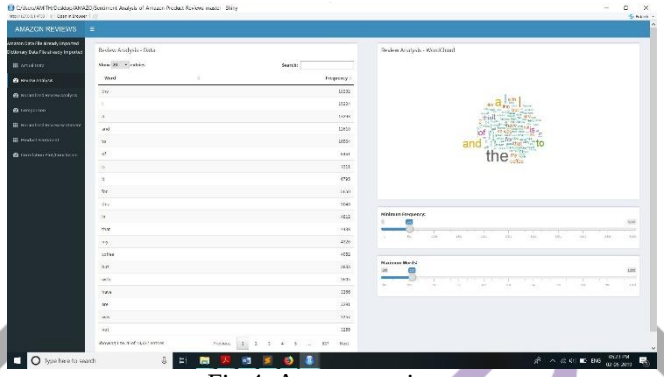

Fig 4, Amazon reviews

Fig 4 shows Amazon reviews collected and sentiment analysis representation and polarity of sentiments.

Yelp Spam review results

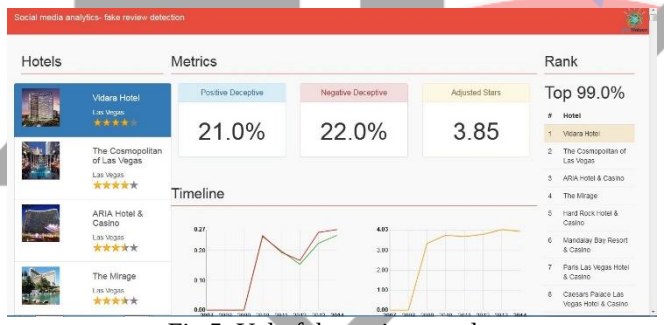

Fig 5, Yelp fake review results

Figure 5 shows the yelp fake reviews detected using our application.

V. CONCLUSION AND FUTURE WORK

Social media sentiment analysis is developed to help businesses to analyze customers' perspectives toward their products or their chances to win the market. The program is using a machine-based learning approach which is more accurate for analyzing a sentiment; together with natural language processing techniques that clearly depicts the attitude of public. By analyzing a text of customer feedback and reviews allows effective quality management. With sentiment analysis, companies can now strategically reposition their businesses according to customers' sentiments.

This paper provided an introduction and rationale behind the value of text analytics of Twitter, Amazon, Yelp data to businesses in gaining customer views on products and services, and brand. This paper also focus on detecting fake reviews as a solution for being misled by wrong information which may harm the inference made by business analyst. The paper demonstrated a practical application of text classification and clustering of Social media or online data, and revealed ways on how to analyze these to gain business insights. The accuracy rate for this experiment is good and acceptable in this application domain. But still it is suggested that future work needs to increase the accuracy of the model by improving using other algorithms.

The work is actually a real time analytics of Twitter data stream and amazon reviews. Hence sentiment is calculated on real time data. For fake review analysis, a readily available yelp dataset is used. Hence that open up as space for performing fake review analysis on large scale real time data.

REFERENCES

[1] B. Liu, "Sentiment Analysis and Opinion Mining," Synth. Lect. Hum.Lang. Technol., vol. 5, no. 1, pp. 1–167, 2012.

[2] marketsandmarkets.com, "Text Analytics Market by Component (Software, Services), Application (Customer Experience Management,Marketing Management, Governance, Risk and Compliance Management), Deployment Model, Organization Size, Industry Vertical,Region - Global Forecast to 20," 2017.

[3] F. N. Ribeiro, M. Araújo, P. Gonçalves, M. André Gonçalves, and F.Benevenuto, "SentiBench - a benchmark comparison of state-of-the practice sentiment analysis methods," EPJ Data Sci., vol. 5, no. 1, 2016

[4] A. Moreno and T. Redondo, "Text Analytics: the convergence of Big Data and Artificial Intelligence," Int. J. Interact. Multimed. Artif. Intell.,vol. 3, no. 6, p. 57, 2016.

[5] V. A. Kharde and S. S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques," Int. J. Comput. Appl., vol. 139, no. 11, pp.975–8887, 2016.

[6] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Eng. J., vol. 5, no. 4, pp. 1093–1113, 2014.

[7] L. Ziora, "The sentiment analysis as a tool of business analytics in contemporary organizations," Stud. Ekon., pp. 234–241, 2016.

[8] S. Yaram, "Machine learning algorithms for document clustering and fraud detection," in Proceedings of the 2016 International Conference on Data Science and Engineering, ICDSE 2016, 2017.

[9] N. Yussupova, M. Boyko, and D. Bogdanova, "A Decision Support Approach based on Sentiment Analysis Combined with Data Mining for Customer Satisfaction Research," Int. J. Adv. Intell. Syst., vol. 1&2, 2015.

[10] S. K. Markham, M. Kowolenko, and T. L. Michaelis, "Unstructured Text Analytics to Support New Product Development Decisions," Res.Technol. Manag., vol. 58, no. 2, pp. 30–39, 2015.

[11] N. Jindal and B. Liu, "Analyzing and detecting review spam,"International Conference on Web Search and Data Mining, 2007, pp.547-552.

[12] N. Jindal and B. Liu, "Opinion spam and analysis," International Conference on Web Search and Data Mining, 2008, pp. 219-230.

[13] S. Xie, G. Wang, S. Lin, and P. S. Yu, "Review spam detection via temporal pattern discovery," ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, 2012, pp. 823-831.

[14] A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellano, and R. Ghosh, "Spotting opinion spammers using behavioral footprints," ACM sigkdd International Conference on Knowledge Discovery and Data Mining, 2013, pp. 632-640.

[15] E. P. Lim, V. A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," ACM International Conference on Information and Knowledge Management, 2010, pp. 939-948.

[16] B.Agarwal,V.K.Sharma,andN.Mittal,"Sentiment Classification of Review Documents using Phrase Patterns," International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1577-1580, . 2013.

[17] J.Zhu, H.Wang, M.Zhu, B.K.Tsou, and M.Ma,," Aspect-Based Opinion Polling from Customer Reviews," T. Affective Computing2(1):pp. 37-49, 2011.

[18] M.Karamibekr,A.A.Ghorbani,"Verb Oriented Sentiment Classification," Processed of the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol (1): pp.327-331, 2012.

[19] A. Neviarouskaya, H.Prendinger, and M.Ishizuka," SentiFul: A Lexicon for Sentiment Analysis," T. Affective Computing 2(1), pp.22-36, 2011.

[20] L.Liu, X.Nie,and H.Wang," Toward a Fuzzy Domain Sentiment Ontology Tree for Sentiment Analysis," Processed of the 5th Image International Congress on Signal Processing (CISP), pp. 1620 – 1624,2012