Sentiment Analysis of Social Media Data using Hybrid Approach

Pradeep Kumar Jaswal
Guru Nanak Dev Engineering College, Ludhiana, Punjab, India.
pradeepjaswal16@gmail.com

Jatinder Singh Tathgir
G.H.G. Khalsa College, Gurusar Sadhar, Ludhiana, Punjab, India
tathgir_js@yahoo.com

ABSTRACT

With the explosive growth of social media, the way of interaction with individuals has changed and it has also made interaction faster and more effective. Sentiment Analysis is the most fascinating and an alluring field of the research which involves linguistic processing, computational semantics or interpretation of content in order to identify the polarity of the text. The main aim of this method is to proficiently recognize the feelings conveyed in reviews of different domains like news articles, stock prices, e-learning field and product reviews etc. With the enormous expansion of web content, complexity of extracting sentiments from text is increasing. In this paper, sentiment analysis using k-Mean clustering and Back Propagation Neural Network was proposed to study the classification efficiency of product reviews. The performance results were evaluated in terms of false acceptance rate, false rejection rate and accuracy. The integration of clustering and classification gives much better results than the existing state of the art methods. For understanding the sentiments of customer reviews this approach proves to be an efficient method.

KEYWORDS

Sentiment analysis, Feature Extraction, k-Mean cluster, Back Propagation Neural Network

I. INTRODUCTION

In this modern digital world, social media is imparted as the most utilized media for exchanging information. The fast increment in web applications [2] [4] [8] has an immense volume of data accessible on the web today. It provides broad range of online resources that enable clients to join web groups, manage and exchange content or contribute information. Online networking sites appear to have a basic impact on individual's lives. Today, a huge range of web-based social networking sites [1] [10] has been created that facilitate the advance of sharing data and information in an online manner. There are various web-based social networks that encourage these services, for example, Twitter, Wikipedia, YouTube and Facebook [12]. The individual's way of living has been changed by using these social means of interaction. Moreover, people are interested to use these online services to share content and get social help by connecting with other users. Business utilizes online networks to upgrade an efficiency of system in order to achieve business goals.

Over the years, various ecommerce [5] websites were introduced including: Amazon, Snapdeal, Jabong and Ebay which helps customers to make buying choices. Customers are now able to give their point of view regarding any product or services on the web. The content created by individuals can be useful to different associations or systems. Discovering methods, to mine [8] such data or information is important in this web period. One such method for mining customer sentiment is referred as Sentiment Analysis, also called Opinion Mining. It is a way of extracting the sentiments or viewpoints from reviews in order to determine the polarity [4] of the text into positive, negative or neutral classes. Social media mining is a sub-discipline of Data Mining [5] [14] and is referred as the method of identifying or discovering new patterns from raw information. Opinion mining is basically used by various organizations, especially in marketing, web based communities, business analytics [5], political associations for satisfying the business goals.

The different social websites or blogs get thousands of reviews and it is difficult to read every single review, hence it is challenging to extricate the user's correct sentiments [9]. The analysis can be done by mining the opinions of different individuals. To perform analysis task, text needs to be in proper structure depicting correct sentiments. As the text fields for giving reviews are generally short having specific character limit for different sites, so in order to give their viewpoints individuals give reviews in unstructured form making it difficult for the analysis purpose. Different approaches [4] for the sentiment analysis can be used i.e. machine learning based approach and lexicon based approach. The machine learning technique utilizes various learning methods to identify the opinion via implementing algorithms on a labeled dataset while the dictionary-based approach includes computing linguistic semantic [4] or valence detection of the reviews utilizing the predefined list of words. Sentiment classification of text can be done at the various levels i.e. document level, sentence level and aspect level [4] [10] [12]:

- **Document Level:** In this procedure, based on the entire viewpoint of the reviewer, the sentiment is identified or categorized from the whole text.
- **Sentence Level:** In this method, the overall sentiment is extracted or evaluated from each comment or sentence in a text.
- **Feature Level:** This level provides brief sentiments of the aspects of a particular object in a text or comment. This categorization is used for recognizing and evaluating item aspects in a detailed manner from the source document.

The reviews or ratings can be from various domains including product brands, music, e-learning, commercials, restaurant services, news articles and stock prices [1] etc. Sentiment analysis helps in e-commerce [5] and business applications for understanding the customer's perspectives and plays an important part in the domain of various text mining fields. For business analysts, this analysis process identifies the user's attitude toward different brands or product. Thus, it is an effective method for making business strategies as most of the business analysts rely on the user's feedback to make better customer services.

II. RELATED WORK

Over the last decade, opinion mining research has developed considerably because of the accessibility of rich content assets and different aspect of sentiment analysis has been explored by various researchers.

Chatterjee and Perizzo [1] discussed the priority of investors and their impact on the stocks in the business. They present the method to recognize the right stock symbol to trade based on the twitter data of stock investors. The Azure ML analyzer was utilized to compute the scores of the tweets having stock symbol in their text and depending on the score of the tweets they are classified into positive, neutral or negative classes.

Hamzah and Widyastuti [2] proposed a model to identify opinion orientation of Indonesian reviews on academic facilities by implementing two different algorithms i.e. K-mean clustering and maximum entropy which shows that the k-mean algorithm perform better than the maximum entropy classifier with high precision and categorizes data into classes in less time.

Povoda et al., [3] proposed a method which depends on Support Vector Machine classifier for classifying text polarity. Their work was assessed with various dialects– English, German, Czech and Spanish. Around 11% exactness was accomplished with the Big Data approach and the best exactness accomplished was 95% for the detection of positive and negative content valence.

Khatri and Srivastava [6] defined that opinion mining can help in anticipating the feelings of individuals which influences the stock costs and the reviews were classified into four classes which were joyful, up, down and refused. It was supposed that the high closing cost of market data will favour more investors but the system suggest of investing in market with high sentiment score.

The distinctive methodologies and techniques utilized for the opinion mining were outlined by D'Andrea et al., [7] for the classification of text with their merits or demerits. Along with this they proposed various mechanisms regarding the distinctive procedures which can be implemented on diverse application fields like marketing, fraud detection and financial data analysis.

Ma et al., [13] presents the strategy to classify the domain oriented dataset of the product reviews based on cross language corpus i.e. English and Chinese. The support vector machine was implemented for the categorization of text after the identification of the feature word by using statistic calculations.

Gautam and Yadav [15] explore the opinion identification to understand the sentiments of the users where data was exceedingly unstructured in the form of tweets. The three supervised learning strategies were implemented for categorization which was Bayesian classification, Support Vector Machine and Maximum entropy. Among these Naive Bayes outperforms the other two strategies with maximum accuracy.

Fong et al., [16] described that the problem starts from the fact that similar sentences or phrases have distinctive sense. Their work concentrates on news articles, which tend to utilize a more impartial vocabulary rather than the sentiment loaded text, for example, publications, surveys, and websites. They compute calculations by training different classifiers and utilized MALLET (Machine Learning for Language Toolkit) and run trials for comparative study.

To give new directions in sentiment analysis field, lots of research has been conducted to facilitate future work. There is problem of unstructured text with high dimensional features which needs to be tackled by finding different feature reduction methods. Different approaches of sentiment analysis needs to be incorporated and combination of different algorithms can be implemented by comparing with the outcomes of existing classifiers to improve the efficiency of the sentiment classification. So, there is a need of evolution of conventional methodologies to resolve existing problems in sentiment classification.

III. RESEARCH METHODOLOGY

In this paper, the work design of the proposed approach is shown in Figure 1.

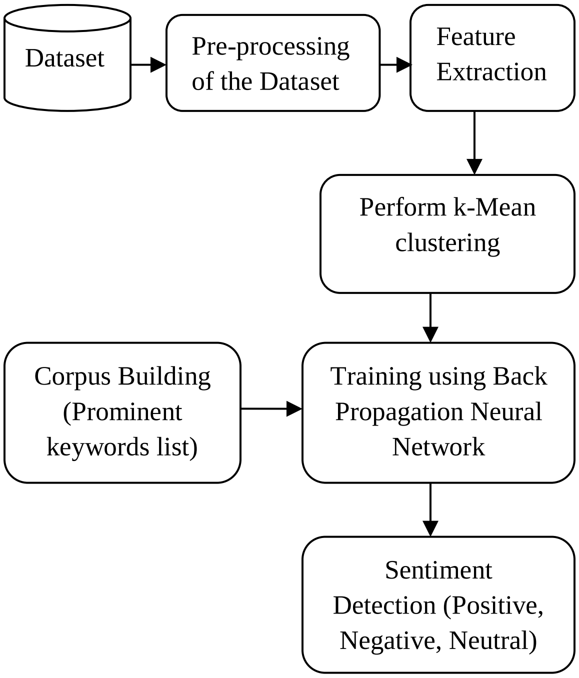
Dataset
Pre-processing of the Dataset
Feature Extraction
Perform k-Mean clustering
Corpus Building (Prominent keywords list)
Training using Back Propagation Neural Network
Sentiment Detection (Positive, Negative, Neutral)

Figure 1 Design Work Flow

At the beginning, the dataset is collected from UCI Machine Learning Repository site. The dataset contain reviews of products, movies and restaurants from three different sites. The system collected 500 positive and 500 negative reviews on the product category which were labeled with the valence value of either 1 (positive) or 0 (negative). These product reviews were gathered from the Amazon site having text in the form of sentences [11]. The reviews clearly depicting the polarity of the sentences were considered for the sentiment analysis task.

A. Pre-processing

Preprocessing eliminates the part which does not contribute significantly to the polarity detection. To perform analysis task, text needs to be in proper structure depicting correct sentiments. The splitting of words was done to read each single word separately to get more meaningful information of the text. As the machine easily perceive the numeric data, the textual data was converted into numeric format forming an array of integers. For building knowledge base, prominent keywords of both positive and negative categories were created manually. Some of these words are given in Table 1.

Table 1 Prominent Keywords

Category	Keywords
Positive	Admire, Adorable, Appealing, Beautiful, Benefit, Best, Brilliant, Charming, Comfortable, Cool, Elegant, Energetic, Excellent, Fine, Good, Great, Helpful, Ideal, Impressed, Happy, Love, Marvelous
Negative	Angry, Arguing, Awful, Bad, Deny, Depressed, Dirty, Disappointed, Evil, Harmful, Hate, Horrible, Misleading, Offensive, Poor, Reject, Rude, Sad, Stressful, Trouble, Ugly, Unpleasant, Worst, Worthless

B. Feature Extraction

For feature extraction, PCA (Principal Component Analysis) was implemented which identify the hidden information and was used to reduce the dimensionality of the feature space. PCA is an algorithm which utilizes the linear orthogonal transformation for finding patterns in the given data. This step was performed to build a low space vector reducing the computational time of the system. The most important step was to calculate the Eigen values and Eigenvectors. The Eigenvector with the highest Eigen value is the first principal component. Further, more vectors were selected retaining most significant information with maximum variance and rest values were neglected.

C. K-Mean clustering

This algorithm was employed to categorize data into two groups (k = 2) namely, positive and negative classes. After feature extraction, the extricated features were provided to the clustering algorithm. The similar data points were placed in one group and rest similar data points into another group. The clusters were formed on the basis of Euclidean distance function. This procedure helps to provide more efficient results and also reduces the processing time of the classifier. Moreover, this method is the simplest procedure which separates or makes distinction between the data points in the space.

D. Back Propagation Neural Network

For the classification purpose back propagation neural network (BPNN) was implemented. This network works same as the human biological neural system. The network consists of three layers: input layer, output layer and the intermediate layer i.e. the hidden layer. It consists of an extensive number of interconnected processing units called neurons to process the input values. Each neuron was assigned with weights and was linked with other neurons to provide desired output. The role of hidden layer is to update the weights on the connections based on the input signal and error signal. At each hidden neuron, activation function was used to perform the calculations. This algorithm repeats the data to every possible path of the network in order to minimize the cost function. Moreover, the implementation of this algorithm is faster and efficient depending upon the amount of input-output data available in the layers. The simulate model was used to train and test the dataset. The testing data was given to the network for the prediction of the output. Then the similarity based matching was done to give the desired output.

IV. EXPERIMENTAL RESULTS

The proposed method of analysis was experimented using MATLAB tool. The classification metrics considered for sentiment analysis is Accuracy as it is a common measure for computing classification performance. The integration of clustering and classification was utilized to improve the performance of the system. The knowledge base and simulate system matches the features of the provided data and gives the result based on similarity factor. For the performance evaluation, False Acknowledgement Rate (FAR) and False Rejection Rate (FRR) were computed to calculate the accuracy of the system. FRR and FAR parameters needs to have low values in order to improve the accuracy. These performance measures check the effectiveness and exactness of the system based on the provided input.

Table 2 Results of Performance Measures

Parameters	Result Values
False Acceptance Rate	0.0064
False Rejection Rate	0.0230
Accuracy (%)	94.5%

The above table shows the result values of performance parameters i.e. False Rejection Rate and False Acceptance Rate. The Accuracy parameter computed of proposed work is 94.5%.

Table 3 Comparison with the Existing Work

Study	Techniques	Accuracy (%)
Ma et al. [13]	SVM	75.5%
Gautam and Yadav [15]	Naive Bayes	88.2%
Proposed Approach	k-Mean + BPNN	94.5%

Table 3 depicts the difference of accuracy between proposed and existing work. The proposed work accuracy factor is high as compared to the existing work. The accuracy rate of the proposed system is 94.5% while the accuracy of existing classifiers Naïve Bayes and SVM is comparatively less than the proposed approach.

V. CONCLUSION AND FUTURE SCOPE

Results of sentiment analysis are necessary for the proficiency and achievement of progressive work. In this paper, sentiment analysis using hybrid of clustering and classification was proposed to study the classification efficiency of the reviews. The system was designed to evaluate the study of user's viewpoint expressed in piece of text in an effective manner. Principal Component Analysis plays an important part in this work to extricate most relevant features as this procedure reduces the computational time of the system by reducing the dimensions and removing irrelevant values. The similarity factor was used to match the features and this factor produces appropriate results in terms of accuracy. The hybrid of clustering and classification gives much better results than the conventional classifiers.

This work can be further explored with the real time dataset for particular time period to attain more optimised results. For the analysis task, features can also be extracted from the videos and images rather than just the textual data.

REFERENCES

[1] A. Chatterjee and W. Perrizo, "Investor classification and sentiment analysis," *Advances in Social Networks Analysis and Mining (ASONAM), International Conference on IEEE, San Francisco, USA*, pp. 1177-1180, Aug 18-21, 2016.

[2] A. Hamzah and N. Widyastuti, "Opinion classification using Maximum Entropy and K-Means Clustering," *Information & Communication Technology and Systems (ICTS), International Conference on IEEE, Surabaya, Indonesia*, pp. 162-166, Oct 12-21, 2016.

[3] L. Povoda, R. Burget and M. K. Dutta, "Sentiment Analysis Based on Support Vector Machine and Big Data," *Telecommunications and Signal Processing (TSP), 39th International Conference on IEEE, Vienna, Austria*, pp. 543-545, June 27-29, 2016.

[4] M. D. Devika, C. Sunitha and A. Ganesh, "Sentiment Analysis: A Comparative Study on Different Approaches," *Procedia Computer Science*, vol. 87, pp. 44-49, 2016.

[5] R. Mehta and S. Jain, "Sentiment Mining and Related Classifiers: A Review," *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 18, no. 1, pp. 50-54, 2016.

[6] S. K. Khatri and A. Srivastava, "Using sentimental analysis in prediction of stock market investment," *Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 5th International Conference on IEEE, Noida, India*, pp. 566-569, Sept 7-9, 2016.

[7] A. D’Andrea, F. Ferri, P. Grifoni and T. Guzzo, "Approaches, tools and Applications for Sentiment Analysis Implementation, "*International Journal of Computer Applications*, vol. 125, no. 3, pp. 26-33, 2015.

[8] A. Mukwazvure and K. P. Supreethi, "A hybrid approach to sentiment analysis of news comments," *Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), 4th International Conference on IEEE, Noida, India*, pp. 1-6, Sept 2-4, 2015.

[9] A. P. Chauhan and K. M. Patel, "Sentiment Analysis Using Hybrid Approach: A Survey," *Int. Journal of Engineering Research and Applications*, vol. 5, no. 1, pp. 73-77, 2015.

[10] A. Tripathy, A. Agarwal and S. K. Rath, "Classification of Sentimental Reviews Using Machine Learning Techniques," *Procedia Computer Science*, vol. 57, pp. 821-829, 2015.

[11] D. Kotzias, Denil, M. Denil, N. Freitas and P. Smyth, "From group to individual labels using deep features," *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia*, vol. 9, no. 4, pp. 597-606, Aug 10-13, 2015.

[12] D. Kulkarni and S. F. Rodd, "A Survey on Opinion Mining problem and levels of Analysis," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 4, no. 12, pp. 12390-12394, 2015.

[13] H. Ma, Y. Zhang and Z. Du, "Cross-language Sentiment Classification based on Support Vector Machine," *Natural Computation (ICNC), 11th International Conference on IEEE, Zhangjiajie, China*, pp. 507-513, Aug 15-17, 2015.

[14] T. Zatari, "Data mining in social media," *International Journal of Scientific & Engineering Research*, vol. 6, no. 7, pp. 152-154, 2015.

[15] G. Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," *Contemporary computing (IC3), Seventh International Conference on IEEE, Noida, India*, pp. 437-442, Aug 7-9, 2014.

[16] S. Fong, Y. Zhuang, J. Li and R. Khoury, "Sentiment analysis of Online News using MALLET," *Computational and Business Intelligence (ISCBI), International Symposium on IEEE, New Delhi, India*, pp. 301-304, Aug 24-26, 2013.