SENTIMENT ORIENTATION SYSTEM OF AUTOMOTIVE REVIEWS USING MULTINOMIAL NAIVE BAYES CLASSIFIER AT DOCUMENT LEVEL

[1]THIDA WIN, [2]PHYO THU ZAR TUN

[1,2]Computer Engineering and Information Technology Department MTU Myanmar
Email: [1]thidawin199237@gmail.com, [2]pinquinphyo@gmail.com

Abstract - The abundance of opinions available on the World Wide Web represents an information repository of enormous intellectual and economic value. Web is used in every field. Almost of all people use web for an ordinary purpose like online shopping sites, blogs, social network sites, forums etc. The large numbers of reviews are given by the users that reflect whether the product is good or bad. People's opinions and experience are very valuable information in decision making process. Opinion Mining or Sentiment Analysis is a natural language processing task that concerns with finding orientation of opinion in a piece of text with respect to a topic. This paper focuses on document level opinion mining and proposes clustering documents according to their polarity scores using K-means algorithm and classifying the particular documents based on the clusters using Naïve Bayes classifier. In the proposed system, TF-IDF approach is applied to denote how important a word is to a document and SentiWordNet supports the system to determine the scores of opinion words. The experimental work was done on automotive reviews. The proposed method achieved total accuracy of 93.7% on the test set.

Indexterms - K-means, Naïve Bayes classifier, Opinion mining, Sentiment analysis, SentiWordNet, TF-IDF approach.

I. INTRODUCTION

Nowadays, a huge number of product reviews or suggestions are springing up on the Web. Reviews may contain the reviews on products, user or critic reviews on movies, suggestions for hotel services etc. From these reviews, product information and direct supervision of customers can be obtained. With sentiment analysis or opinion mining, the task of assigning a sentiment polarity is referred to documents to determine whether the reviewers expressed a positive, negative or neutral thinking about a subject or topic [6]. Opinion mining is a technique which is used to detect and extract subjective information in text documents. In general, sentiment analysis tries to determine the sentiment of a writer about some aspect and also the overall contextual polarity of a document [3]. It can be useful in several ways, i.e., in marketing it helps in judging the success of a new product launch, determine which version of a product are popular and even differentiate which demographics like or dislike particular ones [2].Opinion mining is a Natural Language Processing (NLP) and Information Extraction (IE) task that aims to obtain feelings of the writer expressed in positive or negative comments by analyzing a large number of documents. The main concern is how to automatically identify opinion about an entity from huge volume of unstructured text. The main task of sentiment analysis is to classify the documents and determine its polarity. Sentiment analysis can be performed as three levels: document level, sentence level and aspect and feature level. In document level, the overall opinion about the document is finding out and classifies them as positive or negative. In sentence level, each sentence in the document is analyzed for finding the fine grained opinions about different topics in a document. Entity or aspect level is the featured based approach. Major data mining techniques used to extract the knowledge and information are: generalization, classification, clustering, association rule mining, data visualization, neural networks, fuzzy logic, Bayesian networks and genetic algorithm, decision tree. The techniques of Opinion mining are machine learning approach, a lexicon-based approach and a hybrid approach [7]. Opinion mining represents a large problem space. There are different names with slightly different tasks, e.g., sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc. However, they are now all under the area of sentiment analysis or opinion mining. The term opinion mining is commonly used in academic and sentiment analysis in industry. Many researchers have been focusing on the use of traditional classifiers like Naïve Bayes, Maximum Entropy and Support Vector Machine (SVM) to classify opinions [12].This paper focuses on document-level sentiment classification on automotive reviews. Specifically, the document-level sentiment analysis is to identify either a positive or negative opinion in a given opinionated review. First step is pre-processing including stop word removal and stemming to get clean data. Then extracting opinion words is done by using SentiWordNet. After extracting opinion words, TF-IDF weighting is used to calculate polarity score of each document. K-

means algorithm is applied to cluster documents according to polarity score and to obtain a training dataset. Clustering can provide supplementary constraints that help to classify new data. The particular documents are used to classify as a testing data based on training data by applying Naïve Bayes Classifier. This work proposes a supervised learning approach for document based opinion mining which uses score of opinion words, SentiWordNet.

II. RELATED WORKS

Nowadays, the topic of opinion mining from customers' reviews is frequently discussed. Web opinion mining generally refers to the process of extracting product features and opinions from review documents and summarizing them using graphical representation. There are several techniques to perform opinion mining tasks. In this section, others' related works for opinion words extraction are discussed [1].K. Bun et al.[5] carried out work on topic extraction from news archive using TF-IDF algorithm. There exist prominent topics that were discussed frequently in many documents from many newswire sources. TF-IDF algorithm assigns heavy term weight to the kinds of terms and thus reveals the main topics.

Harb et al.[8] used two sets of seed words with positive and negative semantic orientations to perform blog classification. Google's search engine is used to create association rules. Total number of positive and negative adjectives is counted in a document to classify the documents. They achieved 0.717 F1 score identifying positive documents and 0.622 F1 score identifying negative documents.

Taboada et al.[13] used lexicon-based method to perform sentiment classification. Zagibalov et al.[14] performed unsupervised sentiment classification of product reviews in Chinese using automatic seed word selection method.

K. Bun and M. Ishizuka[15] worked on topic extraction from news archive and carried out using TF-IDF weighting. There exist prominent topics that were discussed frequently in many documents from many newswire sources. TF-IDF recognizes the terms that try to explain the main prominent topics. These would be the terms that appear frequently in many documents from many newswire sources concurrently. It assigns heavy term weight to these kinds of terms and thus reveals the main topics.

Valarmathi and Palanisamy [4] analyzed a method to create exclusive lists from a document's extracted words. Corpus of words created after exclude list was based on Singular Value Decomposition (SVD) scores. Classification and Regression Trees (CART) and Bayes Net with 10 fold cross validation determined classification accuracy. Output of 76% and 78.667% respectively was obtained.

Chinsha T C[6] described a new different syntactic approach to aspect level opinion mining which use aspect dictionary, SentiWordNet, Dependency parsing, adverb adjective, adverb verb combinations, adjectives and adverbs together for opinion mining process with automatic acquisition of aspects. In the proposed method, aspect matched opinion words are extracted using dependency parsing. Polarity of opinions i.e. positivity or negativity of an aspect is found out using SentiWordNet, adjective, adverb adjective and adverb verb combinations using this aspect based visual summary can be produced which shows positive and negative of each aspect from total reviews. The total accuracy of 78.04%, precision of 79.56% and recall of 83.05% were obtained on manually annotated test dataset.

III. DESIGN OF PROPOSED SYSTEM

The proposed design for opinion mining is given in the Figure 1. The following section describes the proposed opinion mining model with each step in detail.

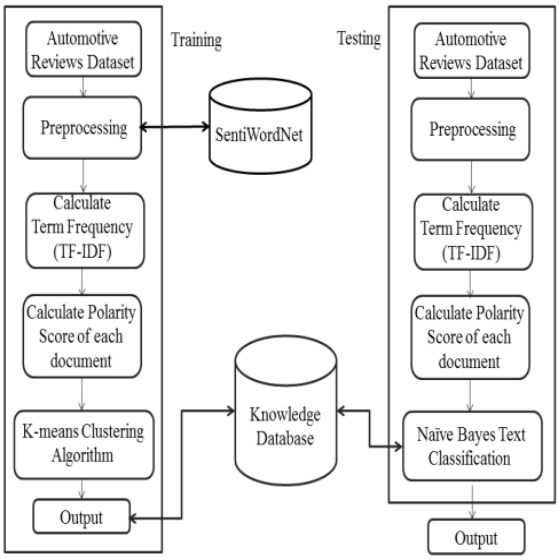

Figure 1. Proposed Design System

3.1. Review Collection

Input to the opinion mining process is the set of reviews. Reviews contain the user and critic reviews. There are various websites available on the web which contains movie reviews, automotive reviews, book reviews, product reviews, hotel reviews, etc. Large numbers of automotive reviews are collected from a website (www.amazon.com). The automotive review dataset contains the contents such as reviewID, asin, reviewerName, helpful, reviewText, overall, summary. The "reviewID" is ID of the reviewer, "asin" is ID of the product, "reviewerName" is the name of reviewer, "helpful"

means helpfulness rating of the review, "reviewText"is the text of the review,"overall" means rating of the product and "summary" is summary of the review.

3.2. Pre-processing

Before determining the polarity of the collected reviews, pre-processing step should be done to improve accuracy and also avoid the unnecessary processing overhead of the opinion mining process. The pre-processing step includes stop words removal and stemming. After stemming, opinion words of each document should be extracted to obtain their polarity scores using SentiWordNet. The all unnecessary characters should be eliminated beforedoing opinion mining steps. In this way, filtering or pre-processing of review documents are to be done.

3.3. Term Frequency and Inverse Document Frequency Approach (IF-IDF)

Frequent words are extracted from reviews using tf (Term Frequency), idf (Inverse Document Frequency) weighting, which denote how important a word is to a document.Term Frequency is the simplest approach to assign the weight to be equal to the number of occurrences of term t_i in document d_j.

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j}, \ldots, f_{|v|j}\}} \qquad (1)$$

where,

tf_{ij} = term frequency of term t_i in document d_j

f_{ij} = the raw frequency count of term t_j in document d_j

Inverse frequency is measure of whether the term is common or rare across all documents. The inverse document frequency of a term can be defined as follows:

$$idf_i = \log \frac{N}{df_i} \qquad (2)$$

where,

idf_i = inverse document frequency of term t_i

N = the total number of documents in the system or the collection

df_i = the number of documents in which term t_i appears at least once

For term frequency and inverse document frequency weighting scheme,

$$w_{ij} = tf_{ij} \times idf_i \qquad (3)$$

where, w_{ij} = weight of term t_i in document dj

3.4. Polarity Score

Assigning a positive or negative score to a word out of context, i.e., a word's prior polarity, is a challenging task for sentiment analysis. Document classification task that quantifies how much a given document (review) appreciates the target object using not binary polarity (good or bad) but a continuous measure called polarity score. SentiWordNet is one of the resources and has been widely adopted since it provides a broad-coverage lexicon-built in a semi-automatic manner for English.

Sentiment words are employed to compute the tendency of a sentence. Scores of words are important to indicate their strengths. Several formulae to compute prior polarities starting from posterior polarity scores have been used. However, their performance varies significantly depending on the adopted variant. Therefore, polarity score calculation is needed to determine whether the word is positive score or negative score and finally the score of each document is obtained by summing the opinion word score in each document.

Polarity of opinion word = (Positive Score –Negative Score)× ($TF - IDF$ Weighting) (4)

3.5. K-means Algorithm

Clustering is a standard procedure in multivariate data analysis. It is designated to explore an inherent structure of the data objects, where objects in the same cluster are as similar as possible and objects in different clusters are as dissimilar as possible.A widely used algorithm for clustering is K-means which is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. It is easy to implement and perform well not only for small datasets but also can be applied even on large datasets. It is simple algorithm that has been successfully adapted to many applications like computer vision, data mining and many others. The goal of k-means is to identify similarities between items in a dataset and group them together based on a similarity function [15].The procedure follows a simple and easy way to classify a given dataset through a certain number of clusters. The main idea is to define k-centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given dataset and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point, k-new centroids are needed to recalculate as barycenter of the clusters resulting from the previous step. After having these k new centroids, a new binding has to be done between the same dataset points and the nearest new center. A loop has been generated. As a result of this loop, the k centers change their location step by step until no more changes are done or in other words centers do not move any more. K-means clustering algorithm is as follows:

Input: N = {n_1, n_2, n_3…… n_n}

Step1: Give n objects, Initialize k clusters.

Step2: Calculate similarity distance between each data to each centroid.

Euclidean Distance Equation:

$$d = \sqrt{\sum_{i=1}^{n} (x_i \quad y_i)^2} \qquad (5)$$

Step3: Assign each data to closest centroid's center.

Step4: Repeat Step 2 and 3 until no change in each cluster.

3.5. Multinomial Naïve Bayes Classification

Naïve Bayes text classification model assumes that all attributes are independent of each other given the context of the class.Multinomial Naïve Bayes is a supervised, probabilistic learning method, which cares about the number of occurrences of each word in the document [10]. The probability of a document d being in class c is computed as

$$P(c \setminus d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k \setminus c) \qquad (6)$$

where $P(t_k \setminus c)$ is the conditional probability of term t_k occurring in the document of class c. P(c) is the prior probability of a document occurring in class c. $<t_1, t_2 \ldots\ldots t_n>$ are the tokens in d that are part of the vocabulary we use for classification and n_d is the number of such tokens in d. For the priors, this estimate is:

$$\hat{P}(c) = \frac{N_c}{N} \qquad (7)$$

Where N_c is the number of documents in class c and N is the total number of documents the conditional probability P(t\c) as the relative frequency of term t in documents belonging to class c:

$$\hat{P}(t \setminus c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}} \qquad (8)$$

To eliminate zeros, we use add-one or Laplace smoothing, which simply adds one to each count:

$$\hat{P}(t \setminus c) = \frac{T_{ct} + 1}{\sum_{t' \in V} T_{ct'} + B'} \qquad (9).$$

IV. RESULTS AND DISCUSSION

For experiment scientifically more stable, we are going to use automotive review dataset. We applied above methods on automotive review dataset that contains 1000 various reviews retrieved from Amazon. These 500 reviews are clustered at first using K-means algorithm for training and the rest 500 reviews for testing are classified with Naïve Bayes classifier according to training database.Documents are clustered into at least 2 to 7 clusters. If polarity score is greater than or equal to 4.5, cluster is "strong positive". If score is greater than or equal 3.5 and less than 4.5, then the cluster name is "positive". The polarity score between 2.5 and 3.5 is "weak positive" and if score is equal to 2.5, this is neutral class. If polarity score is less than 2.5 and greater than or equal to 1.5, then the cluster is "weak negative". The score between 1.5 and 0.5 is "negative" cluster, and if the score is less than or equal to 0.5 then it is in "strong negative" cluster.The evaluation result is observed by varying the number of clusters. When the number of cluster is large and number of testing document is large, it surprisingly gives good accuracy for automotive dataset. Due to the characteristic of dataset and K-means, it is very unlikely that we obtain good clustering using iterations associated with K-means and Naïve Bayes algorithms.

4.1. Performance Evaluation

The performance measures used for the evaluation purpose were Accuracy, Precision and Recall. Formulas are given below:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$(10)\text{Precision} = \frac{TP}{TP + FP}$$

$$(11)\text{Recall} = \frac{TP}{TP + FN}$$

(12)

In the above formula, TP is the true positive instances. TN is the true negative instances. FP is the false positive instances and FN is the false negative instances.

In this system, the performance of the proposed system, in terms of accuracy, is calculated by using 300 training dataset and 150 test dataset. The accuracy is the percentage of number of correct classifications by the number of total test data sets.When the 150 test data are classified by using the proposed system, it can classify 133 true positive documents, 7 false positive documents, 1 true negative document and 2 false negative documents.

Document level accuracy precision and recall of the proposed system on the dataset are calculated given in the Table 1 and Figure 2.

Table 1. Accuracy, Precision and Recall of Proposed System

Total Documents	50	100	150
Accuracy	82.22%	90.43%	93.71%
Precision	0.92	0.98	0.95
Recall	0.88	0.92	0.99

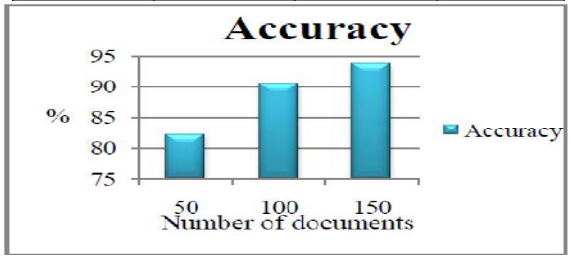

Figure 2. Accuracy, Precision and Recall of Proposed System

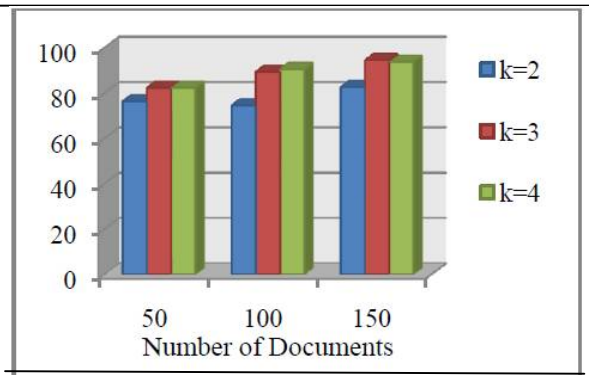

Figure 3. Accuracy based on number of k

The accuracy level based on the number of cluster k is shown in Figure 3.In the experimental result, when k is 2 and 50 test documents are used, weak negative cluster and strong positive cluster can be categorized and the accuracy of 71.11% is obtained. The accuracy level of 82.22% is the same at k = 3 and k= 4 using the 50 test documents. When 100 testing documents are used, the accuracy level is decreasing slightly at k = 2 but it increases at k = 3 and 4. Lastly, the accuracy increases significantly up to 93.7% when 150 documents are used. It is better than the previous test documents 50 and 100. To sum up, the accuracy is significantly better when the number of clusters k is 3 and 4 and the result is more accurate.

CONCLUSIONS

This paper proposes a sentiment orientation system of automotive reviews using Naïve Bayes Classifier associated with K-means algorithm and SentiWordNet for opinion mining process. In the proposed system, polarity of opinions i.e. positivity or negativity is find out using SentiWordNet. K-means algorithm is used to cluster the documents because training dataset is needed in the proposed system and Naïve Bayes Classifier is applied to classify the test documents based on training data. Finally, the performance of the proposed method is evaluated by using particular document reviews. Evaluation is done by measuring accuracy, precision and recall. The total accuracy of 93.7% was obtained on the test dataset. We show that combination of a supervised model and a cluster ensemble based on different data partitions can improve classification accuracy.The main objective of this paper is to determine the polarity of the automotive reviews at document level. The results generated by the system are summarized and helpful for the user in decision making. Experimental results indicate that the "Document based Sentiment Orientation System Using Naïve Bayes Classifier" perform well in this domain. Nowadays, opinion mining is very important from the common man to a businessman and everyone is dependent on the web. The opinion expressed on the web helps the businessman to determine what the customers think about their product, brand or movie etc. So it is necessary to mine this large number of reviews and to classify them, so it is helpful for them to read and take decisions. Sentence-level and aspect-level sentiment analysis can also be performed by using dictionary based approach, hybrid approach.In the future, more and more researches can be studied in sentiment analysis such as Bipolar assessment of opinion mining, Machine learning algorithms, Cross-platform opinion mining, Automatic irony detection, etc.

REFERENCES

[1] J. Prichystal, "Mobile application for customers' reviews opinion mining", 19th International Conference Enterprise and Competitive Environment 2016.
[2] A.S. Ali, "Opinion Mining Techniques", International Journal of Innovative Science, Engineering & Technology, Vol.2, Issue 6, June 2015.
[3] P.Kalarani, Dr. S. Brunda, "An overview on research challenges in opinion mining and sentiment analysis", International Journal of Innovative Research in Computer and Communication Engineering", Vol.3, Issue 10, October 2015.
[4] Vaalarmathi B., Palanisamy V., "Opinion mining classification using keyword summarization based singular value decomposition", International Journal on Computer Science and Engineering, Vol.3, No.5, pp.212-215,2011.
[5] K. Bun, M. Ishizuka, "Topic extraction from news archive using TF-IDF algorithm", Proceedings of Third International Conference on Web Information System Engineering.
[6] Herusgegeben von, "Practice and theory of opinion mining and sentiment analysis", Journal for Language Technology and Computational Linguistic, Vol. 29, No.1, 2014.
[7] N. R. Sharma, Prof. V.D Chitre, "Mining, identifying and summarizing features from web opinion sources in customer reviews", International Journal of Innovation and Advancement in Computer Science, Vol.3, Issue 7, September 2014.
[8] R. Sharma, S.Nigam and R. Jain, "Opinion mining of Movie Reviews at Document Level", International Journal on Information Theory (IJIT), Vol. 3, No.3, July 2014.
[9] B. Liu, "Sentiment Analysis and Opinion Mining", May 2012.
[10] "Text classification and Naïve Bayes", Cambridge University, April 1, 2009.
[11] S.A. Fahad, "A Modified K-means Algorithm for Big Data Clustering", April 2015.
[12] Luiz F. S. Coletta, Nadia F. F. da Silva, Eduardo R. Hruschka Estevam R. Hruschka Jr, " Combining classification and clustering for tweet sentiment analysis", Brazilian Conference on Intelligent Systems, 2014.
[13] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M.Stede, "Lexicon based methods for sentiment analysis", Vol.37, 2011.
[14] "Text Classification and Naïve Bayes", 2009.
[15] K. Bun and M. Ishizuka, "Topic extraction from news archive using TF-IDF algorithm", In Proceeding of Third International Conference on Web Information System Engineering.

★★★