Principal Component Analysis Based Opinion Classification for Sentiment Analysis

Vikram Kumar. N[1] Dr. Loganathan R[2]
[1,2]Department of Computer Science & Engineering
[1,2]HKBK College of Engineering, Bangalore, India

Abstract— Sentiments express perspectives or opinions of users, and reviews gives information about how a product is seen. Online reviews are famously utilized for judging quality of product or service and impact decision making of the users while selecting a product or service. Sentiments are progressively accessible in type of reviews and feedback at websites, blogs, and micro blogs, which impacts future customers. As it is not doable to physically handle the colossal measure of sentiments created online, Sentiment Analysis utilizes automatic processes for extracting reviews and separate significant information with sentiment orientation. In this paper, it is proposed to extract the feature set from movie reviews. Reverse document frequency is computed and the feature set is reduced utilizing Principal Component Analysis. The effectiveness of the pre-processing is evaluated utilizing Naive Bayes and Linear Vector Quantization.

Key words: Opinion mining; IMDB; Inverse document frequency (IDF); Principal Component Analysis (PCA); Naïve Bayes; Linear Vector Quantization

I. INTRODUCTION

In the recent years Sentiment Analysis (additionally called Opinion Mining) has turned into the key research device to identify significant responses for the much complex inquiries, for example, "What my customer needs/think?" The whole service industry spins around the above inquiry, different review methods are utilized for customer feedback. With the quick development of the Internet, feedbacks and review of product has moved to online forums from word to mouth. Electronic communities (like face book, mouthshut.com and numerous other online consumer forums) give an abundance of information. The reviews created advantage both the users and the business concerns and "who" says "what" and "how" they say it, is important [1].

Posting so as to microblogging services permit users to share content frequent, short text overhauls. Of these services, Twitter has been by a wide margin the most mainstream – growing quickly from 94K users in April 2007 [15] to more than 200 million unique users by August 2011, with more than 200 million posts or "tweets" produced every day [2]. Users can track the content produced by different users based on non-corresponding "follower" relations. Recently, an assortment of researchers has considered Twitter as a target for applying sentiment analysis and opinion mining systems. Pak and Paroubek [5] gathered Twitter data for this reason and trained a Naive Bayes classifier on both n-grams and part-of-speech tags to identify positive and negative tweets. Davidov et al., [6] performed sentiment classification utilizing distinctive sorts of features, including punctuation, words, and n-grams. Noisy labels for training were selected based on a little number of pre-indicated Twitter hash tags and smileys.

The objective of Sentiment analysis is to ensure the ranking of the supportive votes on the reviews, are enlightening and fitting into the time frame. There are numerous research headings [3], e.g., sentiment classification where opinions are classified positive or negative; subjectivity classification manages the subjective or objective of a sentence and its related opinion; feature/theme based sentiment analysis assigns positive or negative sentiments to points or product features. The sentiment analysis concentrates on passing on polarity or strength to opinion expressions for determining the objectivity-subjectivity orientation of a document [4] or the polarity of an opinion sentence in a document [7]. So far the reviews either utilize the volume of reviews or link structures to predict the pattern of product deals [9, 10] without contemplating the impact of the sentiments in the reviews. In spite of the fact that there is a solid correlation between the volume of reviews and deals, utilizing the volume or the link structures alone don't give satisfactory prediction performance [9, 10].

As opposed to previous work, in this paper we portray a system that computes the converse document frequency (IDF) of words in the movie review and select features utilizing Principal Component Analysis (PCA). The effectiveness of the features in this way selected is evaluated utilizing LVQ classifier.

II. RELATED WORK

A. Domain-Driven Data Mining (D3M)

In the previous couple of years, domain-driven data mining has risen as an imperative new worldview for knowledge discovery [11, 12]. Inspired by the significant crevice between the scholastic objectives of numerous current KDD methods and the real-life business objectives, D3 advocates the movement from data focused hidden pattern mining to domain-driven Actionable Knowledge Discovery (AKD).

B. Review Mining

With the quick development of online reviews, review mining has pulled in a lot of consideration. Early work around there was essentially centered around determining the semantic orientation with the fast development of online reviews; review mining has pulled in a lot of consideration. Early work around there was fundamentally centered around determining the semantic orientation of reviews. Among them, a portion of the studies endeavor to learn a positive/negative classifier at the document level. Jeevanandam Jotheeswaran et al.. [13] labeled the polarity of IMDB movie reviews utilizing three machine learning approaches (Naive Bayes, Maximum Entropy, and Support Vector Machine).

A few studies classify the documents at better level utilizing words for classification. The words are classified into "good" and "bad," group and after that the general

"goodness" or "badness" score for the documents are estimated utilizing certain functions. G. Vinodhini [14] evaluated the semantic distance from a word to good/bad with WordNet. Broadening previous work on plain two-class classification, reviews were determined utilizing diverse rating scales, for example, number of stars [14, 15]. Liu, et al., [12] proposed framework for looking at opinions of contending products based on different feature dimensions. Representation of the strength and weakness of the product was done utilizing "Opinion Observer".

C. Assessing the Review Helpfulness

Contrasted with sentiment mining, identifying the quality of online reviews has gotten relatively less consideration. A couple of recent studies along this course endeavor to detect the spam or low-quality posts that exist in online reviews. J Liu [12] presented a categorization of review spams, and propose some novel methodologies to detect diverse sorts of spams. Jeevanandam Jotheeswaran [15] proposed a classification-based way to deal with separate the low quality reviews from others, with the expectation that such a filtering technique can be fused to upgrade the undertaking of opinion summarization.

D. Recommender Systems

Recommender systems have risen as an imperative answer for the information overload issue where individuals think that its more hard to identify the valuable information adequately. Thinks about around there can mostly be isolated into three bearings: content-based filtering, Collaborative Filtering (CF), and hybrid systems. Content based recommenders depend on rich content portrayals of behavioral user data to construe their interests, which raises significant engineering challenges as the required domain knowledge may not be promptly accessible or simple to keep up.

As an option, collaborative filtering takes the rating data as input, and applies data mining or machine learning algorithms to find use patterns that constitute the user models. At the point when another user goes to the site, his/her activity will be coordinated against those patterns to discover likeminded users and things that could be of enthusiasm to the users are suggested.

III. METHODOLOGY

In this paper, online movie reviews is utilized as data because of its popularity and availability online. The movie reviews are sourced from Internet Movie Database (IMDb), an online database identified with movies, television shows, and actors. Bo Pang and Lillian Lee made a benchmark dataset of movie-review documents from the IMDb archives. The dataset is labeled with general sentiment polarity (positive or negative) or subjective rating (e.g., two stars). This dataset is utilized to evaluate the proposed method in this research paper. Amid preprocessing, generally happening words, which have no relevance to polarity of the document, is listed as stop words and words having the same root word is stemmed. A corpus of words from the document is prepared and the significance on every word as for the corpus is computed utilizing the converse document frequency. The feature set dimension is reduced utilizing Principal component analysis and Learning Vector Quantization is utilized to classify the opinion.

A. Inverse Document Frequency (IDF)

The documents in the dataset are modeled as vector v, for a given set of documents and a set of terms, in the dimensional space . This is a vector space model. At the point when a term happens in the document , the number of occurrence of the term is given by term frequency which is signified by freq(x,a). The relationship of a term regarding the given document is measured by the term-frequency matrix TF(x, a). The term frequencies are assigned values relying upon the occurrence of the terms, so TF(x, an) is assigned either zero if the document does not contain the term or a number generally. The number could be set as TF(x, a)= 1when term happens in the document or uses the relative term frequency. The relative term frequency is the term frequency versus the aggregate number of occurrences of the considerable number of terms in the document. The term frequency is by and large normalized by eqn (1):

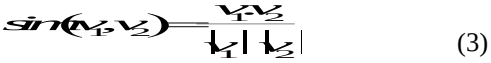

(1)

Inverse Document Frequency (IDF) represents the scaling factor. The significance of a term is scaled down if term happens frequently in numerous documents because of its reduced discriminative power. The IDF(a)is defined as follows in eqn (2):


(2)

x_a is the set of documents containing term *a*.

Similar documents have comparable relative term frequencies, which are exploited to discover comparative documents. Similarity measures are utilized to discover similarity among a set of documents or between a document and a query. Cosine measure is by and large used to discover similarity between documents; the cosine measure is got by eqn (3)


(3)

where v_1 and v_2 are two document vectors, $v_1 . v_2$ defined as

$\sum_{i=1}^{a} v_{1i} v_{2i}$ and $|v_1| = \sqrt{v_1 . v_1}$.

B. Principal Component Analysis

Principal Components Analysis (PCA) is applied to reduce the dimensions of the inputs when the dimensions of the input are extensive and the components are exceedingly corresponded. PCA determines a littler set of artificial variables that will represent the variance of a set of watched variable. The artificial variables computed are called principal components. These principal components are utilized as predictor or criterion variable in different analysis. The variables are orthogonalises by the PCA, principal components with biggest variation are picked, and components with least variation are eliminated from the dataset. The PCA is applied as takes after on a set of data.

- A dataset which has a mean of zero is framed by subtracting the mean of the data from every data dimension.

- Covariance matrix is computed.
- Eigenvectors and Eigenvalues of the covariance matrix are computed.
- Principal components of the dataset are represented by the most astounding Eigenvalues and the Eigenvalues of less hugeness are evacuated and frame a feature vector.
- A new dataset is inferred.

C. Learning Vector Quantization

Vector quantization encodes input vectors by discovering "representatives" or "code-book vectors" that is an approximation to the first input space. A set of prototype vectors characterizes the codebook. The input space is separated up into Voronoi tessellation. An input is assigned to a cluster that is its nearest prototype. The distance is generally measured by Euclidean distance. Learning Vector Quantization (LVQ) is a supervised classification algorithm that is based on self-organizing maps with input vectors and weights or Voronoi vectors. In LVQ, the input data point with the class information permits known class labels of input to find best classification label to each Voronoi cell. New inputs are classified on the premise of the Voronoi cell it falls into. The LVQ algorithm amid training moves the Voronoi cell boundary for improved classification. The input classes are checked against the Voronoi cell and move the weights appropriately as takes after:

1) At the point when input x and Voronoi vector/weight $w_{I(x)}$ have the same class label, then it is drawn nearer together by $\Delta w_{I(x)}(t) = \beta(t)(x - w_{I(x)}(t))$.
2) At the point when input x and related Voronoi vector/weight wI(x) have the distinctive class labels, then it is moved apart by Δ wI(x) (t) = $-\beta(t)(x - w_{I(x)}(t))$.
3) Voronoi vectors/weights W_J relating to other input districts continue as before. where $\beta(t)$ is a learning rate that decreases with the number of iterations / epochs of training.

IV. RESULTS AND DISCUSSIONS

For the examinations, 125 positive and 125 negative movie reviews were utilized. A corpus of 439 terms was extracted after stop words and stemming. The significance of terms was computed utilizing Inverse document frequency. Principal Component Analysis (PCA) was utilized to reduce the features. The classification accuracy got from LVQ and contrasted and Naïve Bayes classifier and Classification and Regression Tree (CART) is appeared in figure 1.

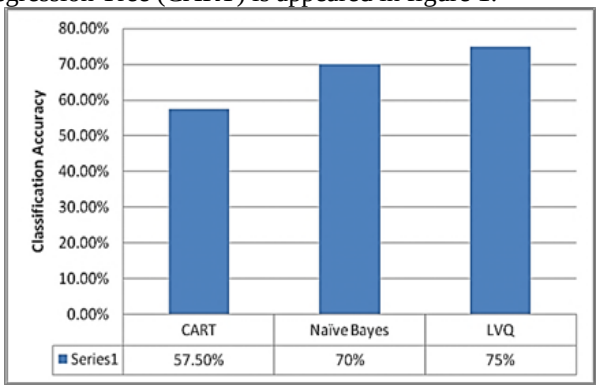


Fig. 1: Classification Accuracy Obtained

The classification accuracy acquired through LVQ is superior to anything Naïve Bayes by a factor of 5%. Figure 2 shows the Root Mean Squared Error (RMSE).

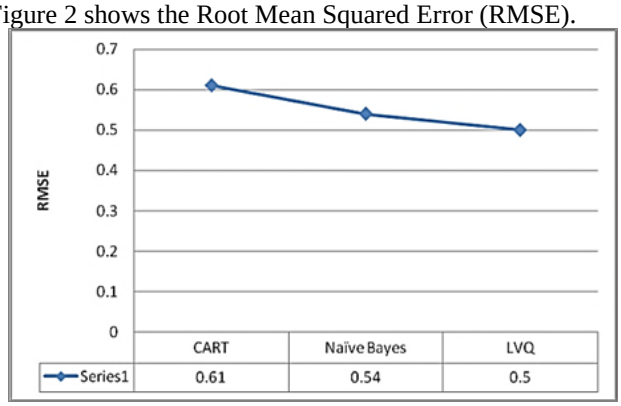


Fig. 2: The Root Mean Squared Error for Each Classifier

The precision and recall for the positive opinion and negative opinion for all the three classifiers is appeared in Figure 3.

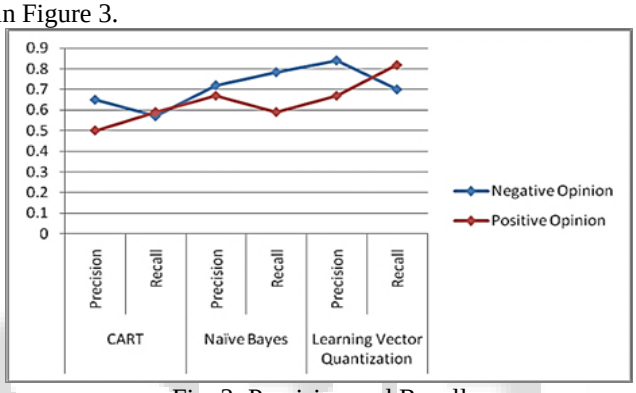


Fig. 3: Precision and Recall

From figure 3 it can be seen that the recall is low for positive opinions in Naïve Bayes and CART which reduced the classification accuracy. Correspondingly, it is seen that precision for positive opinion is very low for all the three classifiers.

V. CONCLUSION

In this paper, it is proposed to investigate the classification efficiency of Sentiment Analysis utilizing Learning Vector Quanitzation classifier. IMDb Movie review dataset was utilized and features was extracted from the review documents utilizing reverse document frequency and the significance of the word computed. Principal component analysis was utilized for feature selection based on the significance of the work as for the whole document. The classification accuracy acquired by LVQ was 75% which is 5% higher than the Naïve Bayes, however it was watched that the precision for positive opinions was entirely low. This marvel was watched on LVQ as well as different classifiers including CART and Naïve Bayes. Further work should be done to improve the classification accuracy of positive opinion.

REFERENCES

[1] Anindya Ghose and Panagiotis G. Ipeirotis on Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics in IEEE transactions on Knowledge and Data engineering, vol. 23, no. 10, October 2011

[2] http://blog.twitter.com/2011/08/your-world-more-connected.html
[3] Ramanathan Narayanan, Bing Liu and Alok Choudhary on Sentiment Analysis of Conditional Sentences in EMNLP,2009
[4] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in Proc. Joint 9th WEBKDD and 1st SNA-KDD Workshop, 2007, pp. 56–65.
[5] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in Proc. 7th Conference on International Language Resources and Evaluation (LREC'10), 2010.
[6] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," in Proceeding of the 23rd international conference on Computational Linguistics (COLING), 2010.
[7] Jeevanandam Jotheeswaran, "Sentiment Analysis using Decision Forest based Feature Selection", International Journal of Advanced Research in Computing and Information Technology, Vol. 1, No. 2, December 2014.
[8] Suke Li, "Sentiment Classification using Subjective and Objective Views", International Journal of Computer Applications (0975 – 8887), Volume 80 – No7, October 2013.
[9] L. Cao, C. Zhang, Q. Yang, D. Bell, M. Vlachos, B. Taneri, E. Keogh, P.S. Yu, N. Zhong, M.Z. Ashrafi, D. Taniar, E. Dubossarsky, and W. Graco, "Domain-Driven, Actionable Knowledge Discovery," IEEE Intelligent Systems, vol. 22, no. 4, pp. 78-88, July/Aug. 2007.
[10] L. Cao, Y. Zhao, H. Zhang, D. Luo, C. Zhang, and E.K. Park, "Flexible Frameworks for Actionable Knowledge Discovery," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 9, pp. 1299- 1312, Sept. 2009.
[11] N. Jindal and B. Liu, "Opinion Spam and Analysis," Proc. Int'l Conf. Web Search and Web Data Mining (WSDM), pp. 219-230, 2008.
[12] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou, "Low-Quality Product Review Detection in Opinion Summarization," Proc. Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP), pp. 334-342, 2007.
[13] Jeevanandam Jotheeswaran et al., "Feature Reduction using Principal Component Analysis for Opinion Mining", Feature Reduction using Principal Component Analysis for Opinion Mining, Volume 3, Issue 5, May 2012. [14] G. Vinodhini ,"Sentiment classification using principal component analysis based neural network model", International Conference on Information Communication and Embedded Systems (ICICES), DOI: 10.1109/ICICES.2014.7033961, 2014.
[14] Jeevanandam Jotheeswaran, Dr. S. Koteeswaran, "a weighted semantic feature expansion using hyponymy tree for feature integration in sentiment analysis", International conference on Green Computing and Internet of Things (ICGCIoT), DOI: 10.1109/ICGCIoT.2015.7380475 , IEEE ISBN: 978-1-4673-7909-0, 2015.

