ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Sentiment TFIDF Feature Selection Approach for Sentiment Analysis

Ganesh Shinde[1], Sachin N. Deshmukh[2]

M. Tech Student, Dept. of Computer Science and Information Technology, Dr B. A. M. University Aurangabad, Maharashtra, India.

Professor, Dept. of Computer Science and Information Technology, Dr B. A. M. University, Aurangabad, Maharashtra, India.

ABSTRACT: Sentiment classification is a task of classifying whether the sentiments of text are positive or negative. Different Machine Learning and Lexicon approaches are used for sentiment analysis. Statistical Techniques for sentiment analysis are more popular. These techniques are based on Term Presence and Term Frequency. Approach given in this paper concentrates on characterizing a term as positive or negative based on its proportional frequency count distribution and proportional presence count distribution across positively tagged documents in comparison with negatively tagged documents. Proposed methodology depends on term weighting techniques that are used for information retrieval and sentiment classification. It contrasts fundamentally from these customary techniques because of proposed model of logarithmic differential term frequency and term presence distribution for sentiment classification. Terms with about equivalent dispersion in positively tagged documents and negatively tagged documents were classified as a neutral word. We evaluated the model for sentiment classification using the movie review dataset**.**

KEYWORDS: Sentiment Classification; Term weighting; Term Frequency; TermPresence

I. INTRODUCTION

Sentiment Analysis involves extracting, classifying and presenting the opinions expressed by the users. Sentiment Classification generally involves the polarity classification of a piece of text [1]. Polarity of a term, sentence, paragraph or document is classified as positive or negative. Sentiment analysis can be done using two approaches Machine Learning Approach and Lexicon Based Approach [2]. Lexicon Based Approach uses existing lexical resources (like WordNet, MPQA, and Senti WordNet 3.0) [3]. These unsupervised learning techniques assigned a generalized polarity and weight to analyses the polarity of texts. Machine Learning Approach constructed sentiment model trained with the help of labelled data. This technique includes the use of supervised machine learning algorithms such as Support Vector Machine (SVM), Naïve-Bayes (NB),Maximum-Entropy (MaxEnt) andForm these SVM is more popular and gives better results as compared to other algorithms [4].

Proposed approach his based on traditional techniques of Information Retrieval. Term Presence and Term Frequency are two prominent techniques for Information Retrieval when representing documents as vectors. InTerm Presence technique an element can take a binary value. This element is set to zero if the term is not present in document and set to one if the term is present in document. In Term Frequency technique an element take integer value that isset to count of the given term in a document. Training data we generate for sentiment classification contain reviews labelled as positive and negative. All reviews with positive labelled are called positively tagged documents where as all reviews with negative labelled are called negatively tagged documents. For further processing we generate the vector that contains terms that occurred in training set documents. Each element of vector has two counts associated with it, first count is for occurrence of the tem in positive tagged documents and second for occurrence of the tem in negative tagged documents. Proposed methodology depends on term weighting functions. This term weighting function based on Term Frequency Inverse Document Frequency (TF-IDF) in which vector of terms are processed to identify the index terms. This method is combination of overall frequency count of term and the presence count distribution. [5,6]. Accordingly we have endeavoured to adapt the model for sentiment classification in which a term was classified as

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

positive if its TFIDF in positively tagged documents was more than negatively tagged documents and vice versa. The TFIDF of the term is calculated using the document vector and its succeeding term vector. The every term t in the term vector contains two counts, first is positivity of term calculated from the positively tagged document given in training data and second is negativity of term calculated from the negatively tagged document given in training data. Proposed approach is different from the traditional approaches. The conventional methodologies, for example, delta TFIDF and other term weighting techniques rely on combination of overall frequency count of term and proportional presence count distribution whereas we focus on proportional frequency count distribution and proportional presence count distribution.

II. RELATED WORK

Earlier work done on sentiment analysis was probably based on the two approaches Machine Learning Approach and Lexicon Based Approach. Pang, Lee and Vaithyanathan uses supervised machine learning techniques for Sentiment Classification. The uses movie review dataset. They use movie review dataset and algorithms like Naive Bayes, maximum entropy classification, and support vector machines applied on unigrams and bigrams features [7]. They concluded that sentiment analysis problem needs to be handled in a more sophisticated way as compared to traditional text categorization techniques. Lexicon based approach includes performing the sentiment analysis at document and sentence level by searching polarity of word from predefined word list.[8,9,10,11] determine polarity of sentence using predefined dictionary. Examples of such a Lexicon dictionaries are MPQA [12] and SentiWordNet 3.0 [13].

TFIDF is a popular statistical technique to index the term. It is based on documents and term vectors that represent term frequency as well as term presence [14, 15]. TFIDF of term is calculating using the term frequency. Larger value of a Term Frequency indicates its prominence in a given document. Terms present in too many documents were suppressed as these tend to be stop words. TFIDF which used single term presence, instead of that Martineau and Finin constructed vectors to classify a term based on term frequency vector as well as term presence vectors. In this technique two vectors were separately constructed for presence in positively tagged documents and negatively tagged documents [5]. There equation is based on logarithmic function of positive and negative tagged documents with term and count of term in documents that gives a negative value if a term occurred in more number of positively tagged documents as compared to negatively tagged documents and vice-versa. If a term is present in equal number of positive and negative document then this equation returned zero. These terms were classified as stop words. Delta TFIDF returned a negative value if the term was classified as positive and vice-versa. The Delta TFIDF considered overall count of terms in all documents ignoring the frequency distribution of terms across positively and negatively tagged documents.

III. PROPOSED TECHNIQUE

The proposed model has slightly difference as compared to the previously TFIDF models. This Model works on the principle logarithmic proportion of TFIDF of a term across positively tagged documents and negatively tagged documents. If the TFIDF of a term in positively tagged documents is greater than its TFIDF in negatively tagged documents the term is assigned positive polarity and vice-versa.

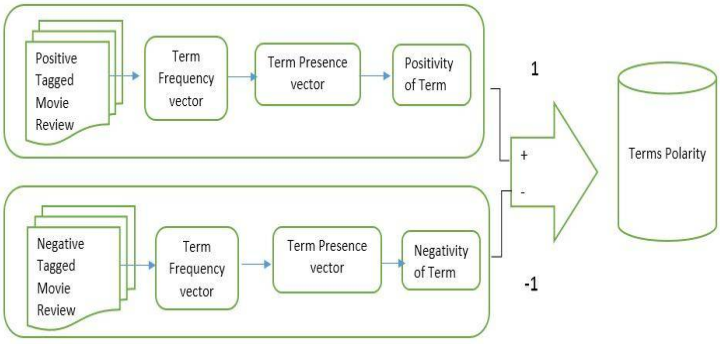

Fig.1.The Proposed Sentiment TFIDF approach

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Figure 1 shows the flow of Sentiment TFIDF. It can be divided into three parts. In first part the positivity of a term is calculated. Similarly negativity of a term is calculated in second part. Then the classification of the term into the Positive, Negative or Neutral classes based on its positivity and negativity is the third step. In the first part term frequency vector and term presence vector are constructed for positively tagged documents using two concepts Term Frequency matrix (TFM) and Term Presence Matrix (TPM). The Term Frequency matrix (TFM) gives the total number of times the term occur in positive documents. And Term Presence Matrix (TPM) relates to document that shows whether the documents contains the term or not. It set to '1' if the term is present in document otherwise it set to '0'. Using these two calculations we create the term frequency vectors for the total terms that are occurred in positive tagged documents and term presence vector for all positive tagged documents. Next we calculate the positivity of each term occur in positively tagged documents. This positivity of term is calculated using the term frequency vector and term presence vector. We use logarithmic function which results in positivity of term.

Similarly in second step we create the vectors of term frequency and term presence for negatively tagged documents. And next to it we calculate the negativity of term. In the third step we find the polarity of term. This polarity is calculated using the logarithmic differential TFIDF (i.e. Terms positivity and negativity). This LDT gives positive output if positivity of term is larger than its negative and vice versa. And if LDT values give output '0' we classify the term as neutral

$$Polt = \begin{cases} 1 & \text{if } >0 \\ 0 & \text{if } =0 \\ -1 & \text{if } <0 \end{cases} \quad (1)$$

Where,
Polt = Polarity of term t.
LDTt = Logarithmic differential TFIDF.
The Polarity of term is calculated using its LDT value given in "(1)".We add small value 0.001 to both Post and Negt. This is because when negativity of a term is zero the model would have been affected by divide by zero error. As a result, the term was classified as positive.

IV. RESULTS

Dataset: For performing sentiment analysis we considered Pang and Lee's movie dataset that contains 1000 Positive and 1000 Negative Documents. As shows in section 3, the term frequency vector and term presence vector for both positive and negative document is generated. Then Positivity and Negativity of term is calculated. The Sentiment TDIDF (I.e. Logarithmic TFIDF) is calculated for the terms that occurred in both positive and negative documents. Initially we perform the 2 experiments using each 500 reviews (i.e. 500 positive and 500 negative). For these experiments we use Support Vector Machine (SVM) classifier

Experiment 1: Initially we perform the Sentiment analysis on 500 reviews each. And then check the accuracy of sentiment analysis model using the 10 fold Cross Validation. We calculate the accuracy at each of 10 folds. Figure 2 shows the accuracy of model. At every fold this 10% dataset was used for testing and remaining 90% dataset was used for training the classifier

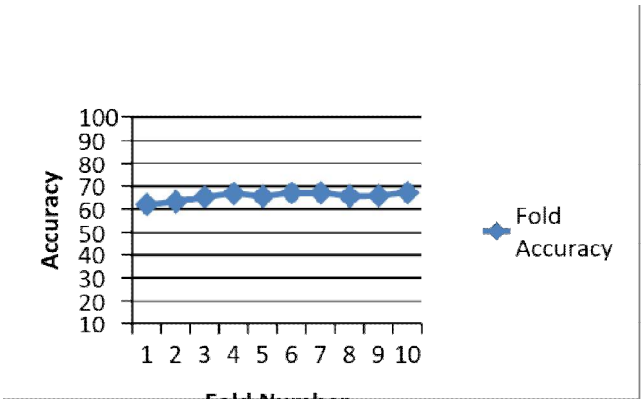

Fig.2. 10 Fold Cross validation accuracy

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Experiment 2:Consider second experiment, accuracy is calculated at different size of training data like first accuracy is calculated using 10% dataset as training data and remaining 90% as testing data. Then the training data is increased by 10% and remaining was used as test data. In further iterations till 90% data was used for training and 10% for testing. Accuracy was calculated at every repetition. Figure 3 shows the results for incrementing training size. The graph shows as increasing size of training data increases accuracy.

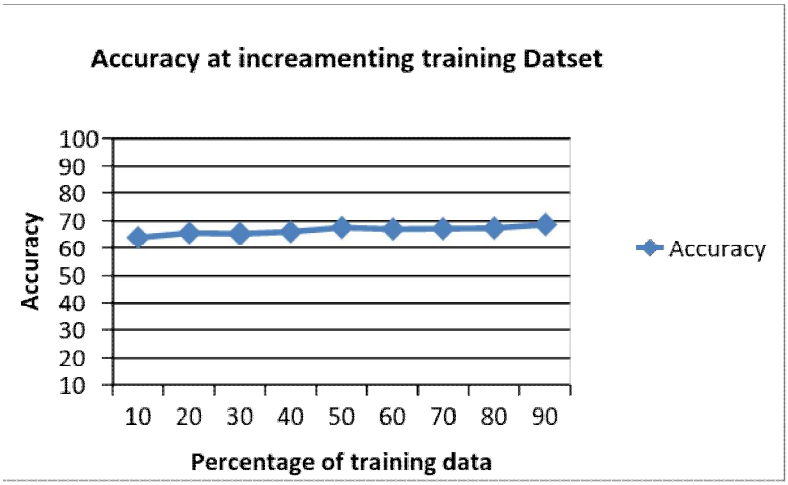

Fig.3. Accuracy by incrementing training size

Experiment 3: In third experiment we calculate the accuracy of overall dataset (i.e. 1000 positive and 1000 negative). We calculate this using 90% of dataset as training data and remaining 10% as testing data. We use SVM Classifier for calculate cross validation accuracy using 2 folds. After performing this experiment we got 75.76% Accuracy. We calculated accuracy for delta tfidf but we got less accuracy for delta tfidf as shown in figure 4. Following figure 4 shows results of accuracy of both approaches.

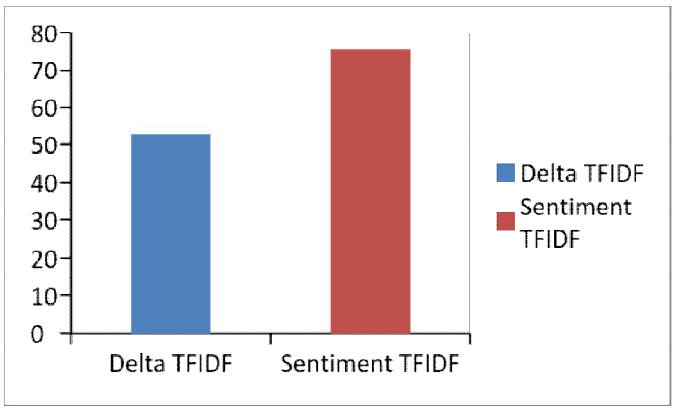

Fig.4. Accuracy

IV. CONCLUSION

As we conduct different experiments on the movie review dataset, from the results of the experiments we observed that the accuracy of sentiment classification for large documents is more than any other sentiment classification technique. Unlike the technique Delta TFIDF, proposed approach efficiently handles the term getting wrongly

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

classified as well as eliminating divide by zero error while doing logarithmic differential term frequency and term presence distribution for sentiment classification. The important difference between Delta TFIDF and proposed approach is that proposed approach considers the Frequency and presence distribution of a term across positive and negative tagged document as compared to Delta TFIDF which considers frequency of a term in all documents and distribution of presence even for Sentiment Classification. The accuracy of this approach on overall movie review dataset is good as compared to surveyed techniques that werType equation here.e tested using movie review dataset. Although these accuracies cannot be directly compared as the experimental parameters may vary. This method using TFIDF performs better. Proposed classifier is based on term frequency and presence distribution. In future we aim to experiment the effect of other distributional count associated with terms and analyse the results.

REFERENCES

1. B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, vol. 2, nos. 1–2, pp.1–135, 2008.
2. E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New Avenues in Opinion Mining and Sentiment Analysis, "Intelligent Systems, IEEE, vol.28, no.2, pp. 15-21, 2013.
3. S. Baccianella, A. Esuli, and F. Sebastiani," SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," In Proc. of 7th Int'l Conf. on Language Resources and Evaluation, pp 2200-2204, 2010.
4. K. Ghag and K. Shah, "Comparative analysis of the techniques for Sentiment Analysis," In Proc. of Int'l Conf. on Advances in Technology and Engineering, pp. 1-7, 2013.
5. J. Martineau and T. Finin, "Delta TFIDF: An Improved Feature Space for Sentiment Analysis," In Proc. of 3rd Int'l AAAI Conf. on Weblogs and Social Media, pp.258-261, 2009.
6. G. Paltoglou and M. Thelwall, "A study of Information Retrieval weighting schemes for sentiment analysis," In Proc. of 48th Annual Meeting of the Association for Computational Linguistics, pp. 1386-1395, 2010.
7. B. Pang, L. Lee, and S. Vaithyanathan,"Thumbs up? Sentiment classification using machine learning techniques," In Proc. Of Conf. on Empirical Methods in Natural Language Processing, pp 79-86, 2002.
8. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. 2011. Lexicon- based methods for sentiment analysis. Comput. Linguist. 37, (2): 267--307.
9. Hu, M., & Liu, B. 2004. Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04). ACM, New York, NY, USA. pp. 168—177.
10. Kim, S., & Hovy, E. 2004. Determining the sentiment of opinions. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING '04). Association for Computational Linguistics, Stroudsburg, PA, USA
11. Ding, X., Liu, B., & Yu, P.S. 2008. A holistic lexicon-based approach to opinion mining. In: Proceedings of the International Conference on Web Search and Web Data Mining (WSDM '08). ACM, New York, NY, USA. pp. 231-240.
12. Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.
13. MPQA Online Lexicon "http://www.cs.pitt.edu/mpqa/subj_lexicon.html".
14. G. Salton and M. McGill, "Introduction to modern Information Retrieval," McGraw-Hill, pp. 105-107 & 205 1983.
15. J. Han and M. Kamber, "Data Mining Concepts and Techniques, "Morgan Kaufmann Publishers, 02nd edition, pp. 364-365, 2006.