ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com
Journal DOI: 10.21474/IJAR01

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

An Approach On Two - Fold Sentiment Analysis.

***Anju Murali J[1] and Varghese S Chooralil[2].**

1. Department of Computer Science and Engineering, Rajagiri School of Engineering and Technology Kochi, India.
2. Department of Computer Science and Engineering, Rajagiri School of Engineering and Technology Kochi, India.

Manuscript Info

Manuscript History:

Received: 11 May 2016
Final Accepted: 22 June 2016
Published Online: July 2016

Key words:
Sentiment Analysis,
Dual Training,
Dual Prediction,
Naive Bayes,
Probability-based classifier

****Corresponding Author***

Anju Murali J.
Varghese S. Chooralil

Abstract

Sentiment classification is a fundamental task in sentiment analysis, with its aim to classify the sentiment (e.g., positive or negative) of a given text. Polarity shift is a kind of linguistic phenomenon which can reverse the sentiment polarity of the text. To model text in statistical machine learning approaches in sentiment analysis, Bag-of-words (BOW) is used. But the performance of BOW sometimes reduced due to some fundamental deficiencies in dealing the polarity shift problem. A Two-fold (dual) sentiment analysis model is used to address this problem for sentiment classification. A data expansion technique, which involves Text reversion and Label reversion, is used for creating a sentiment-reversed review for each training and test review. DSA framework consists of two parts such as Dual Training (DT) and Dual Prediction (DP). For learning a sentiment classifier by means of original and reversed training reviews in pairs, a dual training algorithm is used and along with, for classifying the test reviews by considering two sides of one review, a dual prediction algorithm is used. Feature selection and extraction is done with classifier training. A novel method of introducing a Probability based classifier in addition to the Naive Bayes classifier.

Copy Right, IJAR, 2016,. All rights reserved.

Introduction:-

With the rapid development of internet, demand of online data analysis becomes key role in all areas. Sentiment analysis and Opinion mining involves the study of opinions and its related concepts such as sentiments, evaluations, attitudes and emotions. It is widely used in Data mining, Web mining, Text mining and Natural Language Processing (B. Liu, 2012). Data mining is the analysis step of the "knowledge discovery in databases" process (KDD). During the decision making process, "what people think" has always been an important piece of information. In the past, when an individual needed to make a decision he typically asked for opinions from friends and family. But now it depends on online reviews. Sentiment Analysis (SA) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. It is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service. SA is the text mining task for subjective attitude. Sentiment classification is a basic task in sentiment analysis, with its aim to classify the sentiment (e.g., positive or negative) of a given text. The general practice in sentiment classification follows the techniques in traditional topic-based text classification, where the bag-of-words (BOW) model is typically used for text representation (Xia et al., 2015). In the BOW model, a review text is represented by a vector of independent words. The statistical machine learning algorithms (such as Naive Bayes, Maximum Entropy classifier, and Support Vector Machines) are then employed to train a sentiment classifier (Wang et al., 2013). The Polarity classification is the most classical sentiment analysis task which aims at classifying reviews into either positive or negative. Polarity shift is a kind of linguistic phenomenon which can reverse the sentiment polarity of the text. Negation is the most important type of polarity shift. For example, by adding a negation word "don't" to a positive text "I like this book" in front of the word "like", the sentiment of the text will be reversed from positive to negative.

The scope of this paper is to create reversed reviews that are sentiment-opposite to the original reviews, and make use of the original and reversed reviews in pairs to train a sentiment classifier and make predictions. This model is very efficient for polarity classification and it significantly outperforms several alternating methods of considering polarity shift. The novel approach of basic probability-based classifier is used in addition to feature extraction and selection. This paper includes the related works on two-fold or dual sentiment analysis and their related techniques, the proposed approach, architecture & methodology, results and discussions, conclusion and references.

Related Works and Background Knowledge:-

This section reviews some of the previous work in this field. The web contains a wealth of opinions about products, politics, newsgroup posts, review sites, and elsewhere. According to the levels of granularity, tasks in sentiment analysis can be classified into four categories such as document-level, sentence-level, phrase-level, and aspect-level sentiment analysis. Sentiment Analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic or product. A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level - whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral.

Xia et al (2015, 2013) proposed a simple yet efficient model, called dual sentiment analysis (DSA) is proposed to address the polarity shift problem in sentiment classification. By using the property that sentiment classification has two opposite class labels (i.e., positive and negative), initially a data expansion technique is proposed by creating sentiment-reversed reviews. Harihara et al (2013) proposes a novel approach to contextual analysis that differentiates between single words and phrases. The semantics of a single word in context from that of a phrase are fundamentally different. Since one word will have multiple contexts and is heavily influenced by the surrounding words, more consideration is given to adjacent words. Maximum Entropy classifier is used for training for each set.

S. Li et al (2010) introduced a feature selection method to automatically generate a large scale polarity shifting training data for polarity shifting detection of sentences. By using the obtained binary classifier, each document in the original polarity classification training data is split into two partitions, polarity-shifted and polarity- unshifted, which are used to train two base classifiers respectively for further classifier combination. D. Ikeda et al (2008) introduced a machine learning based method of sentiment classification of sentences using word-level polarity. The proposed method models the polarity-shifters and it can be trained in two different ways: word-wise and sentence-wise learning. In sentence-wise learning, the model can be trained so that the prediction of sentence polarities should be accurate.

D. Kotziasl et al (2015) developed a new approach is introduced to the problem of using group-level labels to learn instance-level classification models. The resulting classifiers can be used to transfer information from the group-level to the instance level when group labels are available, in addition to making predictions about new instances and groups. The approach is evaluated using three large review data sets from IMDB, Yelp, and Amazon. P. D. Turney (2002) presented a simple unsupervised learning algorithm for classifying a review as recommended or not recommended. PMI-IR uses Point wise Mutual Information (PMI) and Information Retrieval (IR) to measure the similarity of pairs of words or phrases. B. Pang et al. (2002) examined the effectiveness of applying machine learning techniques such as Support Vector Machine (SVM), Naive Bayes and Maximum Entropy to sentiment classification problem. A challenging aspect of this problem that seems to distinguish it from traditional topic-based classification is that while topics are often identifiable by keywords alone, sentiment can be expressed in a more subtle manner.

The Proposed Approach:-

The goal of this paper is to effectively address polarity shift problem. Polarity shift is a kind of linguistic phenomenon which can reverse the sentiment polarity of the text. Negation is the most important type of polarity shift (Xia et al., 2015). The motivation is by generating artificial samples that are polarity-opposite to the original ones. The original and opposite training samples are used together for both training the sentiment classifier and for prediction. The novelty of this work is own basic probability-based classifier, which efficiently address polarity shift problems and also feature extraction and feature selection on the reviews are performed and these are taken as input for both training and prediction. The two-fold sentiment analysis model is proposed to address the problem of polarity shift by using basic probability based classifier rather other machine learning techniques. A Dual Training (DT) algorithm and a Dual Prediction (DP) algorithm are introduced respectively, to make use of the original and reversed samples in pairs for training a statistical classifier and make predictions. In DT, the classifier is learned by maximizing a combination of likelihoods of the original and reversed training data set. In DP, predictions are made by considering two sides of one review. That is, we measure not only how positive/negative the original review is, but also how negative/positive the reversed review is. The proposed design is shown in Fig. 1.

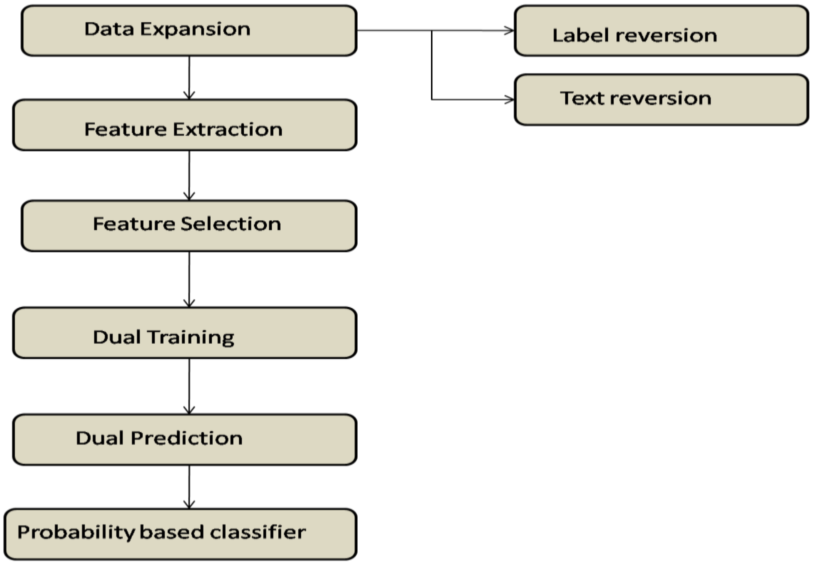

Fig. 1:- Proposed design

The comparison between Naive Bayes and Probability-based classifier are analyzed with the dataset. Feature extraction is done using Tf-idf weight, which stands for term frequency-inverse document frequency. Feature selection is done using 'chi-square distribution' and by 'select k-best'. The select k-best means to select features according to the k highest scores. The chi-square test is used in statistics, among other things, to test the independence of two events. More specifically, in feature selection we use it to test whether the occurrence of a specific term and the occurrence of a specific class are independent.

Architecture & Methodology:-

The overall system architecture will clearly give an idea of each module. The basic work flow is shown in Fig. 2.

Data Expansion:-

The data expansion technique used here is the first work to bring data expansion in sentiment analysis (Wang et al, 2013). The original and reversed reviews are created as one-to-one correspondence by data expansion technique. The data set is expanded not only in the training stage, but also in the testing stage. The original and reversed test review samples are used as pairs for sentiment prediction. A joint prediction is determined based on both original and reversed review. The reversed review is created according to certain criterion such as:

- Label reversion: The class label is reversed to its opposite for each of the training review.
- Text reversion: If there is negation, first identify the scope of negation and eliminate negation words (e.g., "no", "not", etc.)

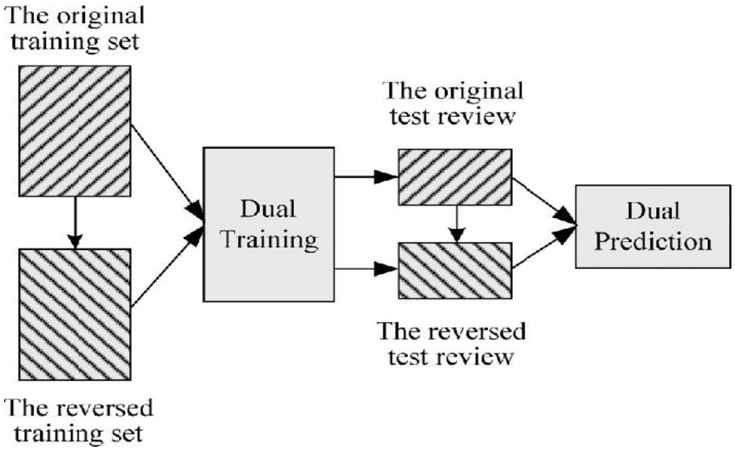

Fig. 2:- Basic workflow

Data Preprocessing:-

IMDB movie review sentiment dataset is taken. The data set is divided into half for training and the remaining half for testing. Each of the labeled reviews has a binary sentiment label, either positive or negative (D. Kotziasl, 2015). In the preprocessing stage, the dataset is converted to an excel format for processing. The data is tokenized, stemmed, part-of-speech (POS) tagging is performed and stop-words are removed.

Feature Extraction and Selection:-

Feature Extraction is done using Tf-idf weight. Tf-idf stands for term frequency-inverse document frequency, and it is a weight often used in information retrieval and text mining (Singhal, 2001). In a collection, this weight is a statistical measure to evaluate how important a word is to a document. In other words, $\mathrm{tf}-\mathrm{idf}_{t,d}$ assigns to term't' a weight in document'd' that is

- Highest when *'t'* occurs many times within a small number of documents (thus lending high discriminating power to those documents);
- Lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal);
- Lowest when the term occurs in virtually all documents.

Feature selection is done using 'chi-square distribution' and by 'select k-best'. In text classification, the feature selection is the process of selecting a specific subset of the terms of the training set and using only them in the classification algorithm. The feature selection process takes place before the training of the classifier. The main advantages for using feature selection algorithms are the facts that it reduces the dimension of our data, it makes the training faster and it can improve accuracy by removing noisy features (Manning, 2008). As a consequence feature selection can help us to avoid over fitting. The chi-square test is used in statistics, among other things, to test the independence of two events. More specifically in feature selection we use it to test whether the occurrence of a specific term and the occurrence of a specific class are independent. High scores on chi-square indicate that the null hypothesis (H0) of independence should be rejected and thus that the occurrence of the term and class are dependent. If they are dependent then we select the feature for the text classification.

Dual Training:-

In the dual training process, Dual Training (DT) algorithm is used. All of the original training samples are reversed to their opposites and called as "original training set" and "reversed training set" respectively (Xia et al, 2015). There is a one-to-one correspondence between the original and reversed reviews in the selective data expansion technique. By maximizing a combination of the likelihoods of both the original and reversed training samples, the classifier is trained. This process is called dual training. The learning errors caused by negation can be partly compensated in the dual training process. From the entire set, half set is given to training and rest half is given for testing. The features with labels are passed to classifier as input. Naive Bayes classifier is used in dual training which uses combined likelihood of the training parameters.

Dual Prediction:-

In dual predicting stage also, the original and reversed samples are used for testing. So the process is known as dual training (Xia et al, 2015).

- To measure not only how positive a test review x is, but also how negative the reversed test review is;
- To measure the probability of x being negative with considering the probability of x being positive.

A weighted combination of two joint predictions is taken. During dual prediction, only features are given to classifier, not any labels.

Probability-based classifier:-

This is the overview of how a probabilistic classifier's working, its training and classification. The weighted probability function is used to calculate the probability that a feature is associated with a given label. The probability that a set of features matches a label is calculated by simply multiply together all the probabilities of the individual features. The final step is to weight the probabilities of the individual features to the overall probability that a document has a given label. To classify a set of features by using probability-based classifier, calculate the probability for each label and then return them sorted so the best match is first.

Requirement Specifications:-

Hardware Requirements:-

- Processor: Intel Pentium III (or higher)
- RAM: 512 MB (or higher)
- Hard Disk: 20 GB (or higher)
- Motherboard: Intel D845 GVSR (or higher)

Software Requirements:-

- Operating System: Microsoft Windows 7 Professional 32/64 bit
- Front-End: Python 2.7.10
- Back-End: MySQL Workbench 6.3.4

Results and Discussions:-

The reverse review is created by using Label reversion and Text reversion. TFIDF method is used for feature extraction. Select k-best and chi-square distribution are used for feature selection method. The reviews are randomly split; a half set is given for training and rest half is given to testing. Each review and its opposite are passed to the training model. So that it can predict well by learning the classifier and thereby avoiding polarity shift problem. We can further observe that with the selected training reviews for data expansion, two-fold sentiment analysis can achieve comparative or even better performance than that of existing methods. Probability-based classifier shows better performance to avoid polarity shift problem over naive bayes classifier. Comparison of precision and recall between Naive Bayes and Probability-based classifier is shown in Fig. 3 and Fig. 4 respectively.

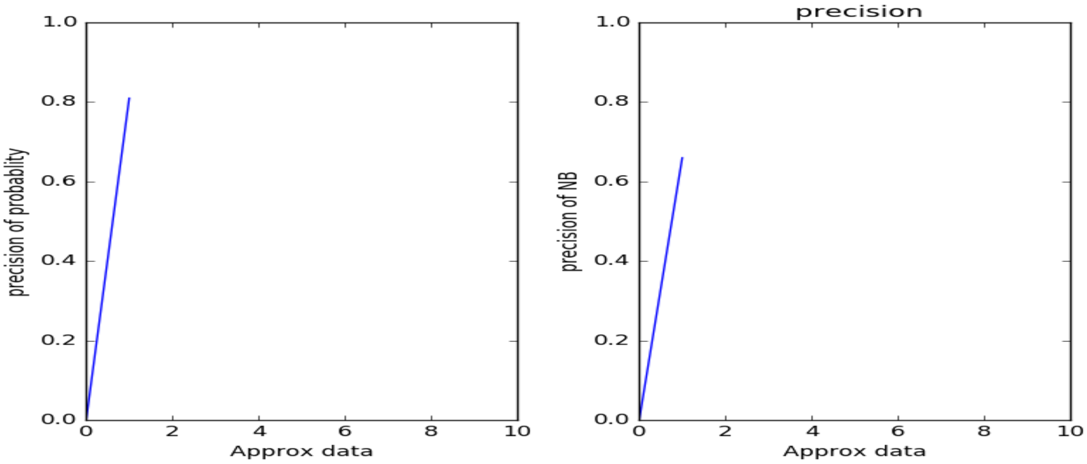

Fig. 3:- Comparison of precision between Naive Bayes and Probability-based classifier

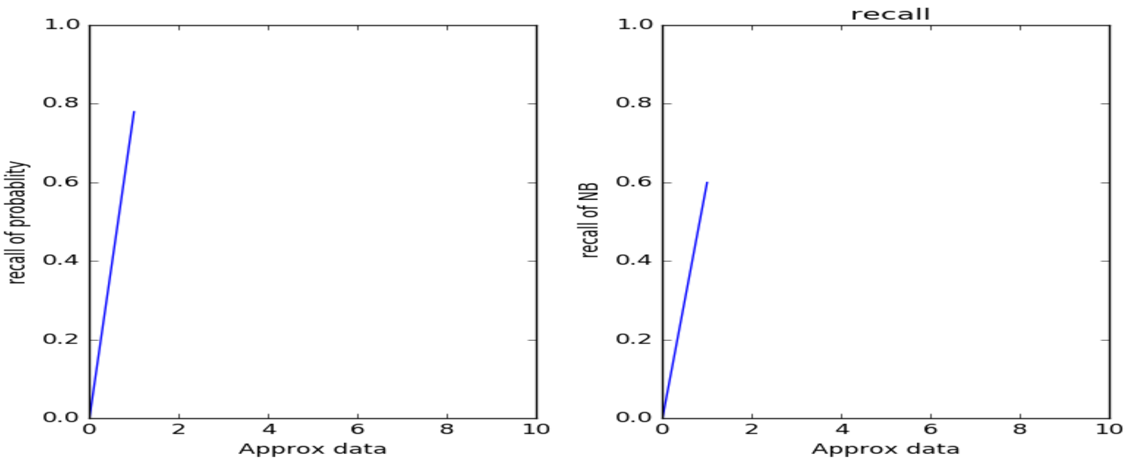

Fig. 4:- Comparison of recall between Naive Bayes and Probability-based classifier

Conclusion:-

The Two-fold sentiment analysis model is proposed to avoid polarity shift problems due to which, most of the standard machine learning algorithms may fail. The Polarity shift is a kind of linguistic phenomenon which can reverse the sentiment polarity of the text. Negation is the most important type of polarity shift (Xia et al., 2015). The two methods used are dual training (DT) and a dual prediction (DP) respectively, to make use of the original and reversed samples in pairs for training a classifier and make predictions. Feature extraction and feature selection are carried out in the reviews and the output is given to training model. With the application of basic probability-based classifier, performance can be improved rather than naive bayes. In the future, the current work can be extended to several applications of sentiment analysis tasks and to other dataset samples, and also in terms of intermediary, subjunctive sentences in constructing reversed reviews to solve the complex polarity shift patterns.

References:-

1. **R. Xia, F. Xu, C. Zong, Q. Li, Y. Qi, and T. Li (Aug. 2015).** Dual Sentiment Analysis: Considering Two Sides of One Review, IEEE Trans. Knowledge Data Eng., vol. 27, no. 8.
2. **D. Kotzias1 M. Denil, N. D. Freitas, and P. Smyth (Aug. 2015).** From Group to Individual Labels using Deep Features, Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
3. **G. Harihara, E. Yang, and N. Chambers (2013).** USNA: A Dual-Classifier Approach to Contextual Sentiment Analysis, United States Naval Academy Annapolis, MD 21401, USA.
4. **R. Xia, T. Wang, X. Hu, S. Li, and C. Zong (2013).** Dual training and dual prediction for polarity classification, in Proc. Annu. Meeting Assoc. Comput. Linguistics, pp. 521-525.
5. **S. Li, S. Lee, Y. Chen, C. Huang and G. Zhou (2010).** Sentiment classification and polarity shifting, in Proc. Int. Conf. Comput. Linguistics, pp. 635- 643.
6. **D. Ikeda, H. Takamura, L. Ratinov, and M. Okumura (2008).** Learning to shift the polarity of words for sentiment classification, in Proc. Int. Joint Conf. Natural Language Process.
7. **B. Pang, L. Lee, and S. Vaithyanathan (2002).** Thumbs up?: Sentiment classification using machine learning techniques, in Proc. Conf. Empirical Methods Natural Language Process., pp. 79-86.
8. **P. D. Turney (2002).** Thumbsup or thumbsdown?: Semantic orientation applied to unsupervised classification of reviews, in Proc. 40th Annu. Meet. Assoc. Comput. Linguist., pp. 417-424.
9. **C. D. Manning, P. Raghavan and H. Schutze (2008).** Introduction to Information Retrieval," Cambridge University Press.
10. **A. Singhal (2001). Modern Information Retrieval:** A Brief Overview, IEEE Computer Society Technical Committee on Data Engineering.
11. **B. Liu (2012).** Sentiment analysis and opinion mining, in Synthesis Lectures on Human Language Technologies, vol. 5, no. 1. San Rafael, CA, USA: Morgan & Claypool, pp. 1-165.