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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Effective Sentiment Classification Using Dual Sentiment Analysis and Random Forest Classifier

Bajeela P.V[1], Swaradh P[2]

M.Tech Student, Department of Computer Science and Engineering, APJ Abdul Kalam Technological University, Kerala, India[1],

Assistant Professor, Department of Computer Science and Engineering, KMCT College of Engineering, Kozhikode, Kerala, India[2]

ABSTRACT: Online reviews available on the internet are majorly used in sentiment analysis and opinion mining to determine the sentiment expressed in the text. Sentiment classification is a basic taskin sentiment analysis, with its aim to classify the sentiment of a given text as either positive or negative. Bag-of-words (BOW) is now the most popular way to model text in statistical machine learning approaches in sentiment analysis. However, the performance of BOW sometimes remains limited due to some fundamental deficiencies in handling the polarity shift problem. The proposed model called Dual Sentiment Analysis (DSA) using random forest classifier, is to address this problem for sentiment classification. Initially, data expansion technique is to be used to create a sentiment-reversed review for each training and test review. Then a dual training is to be performed by using the original and reversed training reviews in pairs for learning a sentiment classifier. After that a dual prediction techniqueis used to classify the test reviews by considering two sides of one review.

KEYWORDS:Sentiment Analysis, Natural Language Processing, polarity shift, feature extraction.

I. Introduction

In this internet era, most people are using internet as a place for expressing their opinion, feedback and review about product, movie etc. So we must have an efficient mechanism to retrieve, organize and analyse these information from the web for effective decision making. One of the most popular method for this is the sentiment analysis. Sentiment analysis is the process of analysing the sentiment or the subjective attitude within a given text. Sentiment analysis is the sub field of Natural Language Processing. The basic task of sentiment analysis is the sentiment classification. Sentiment classification classifies a text according to the sentiment polarities of opinion it contain [7].Bag Of Words (BOW) model is the widely used method for text representation for sentiment classification. BOW model represents a text as a vector of independent words. So in BOW model, the word order, synonyms and grammatical structure of the text are discarded and as its name indicates, it just considers text as a bag of many words.

BOW model is an efficient method in text classification. But it has many fundamental deficiencies since it avoids the exact structure of the text. The main problem that BOW model suffers is the polarity shift problem. A polarity shift is the linguistic phenomenon by which the polarity of a sentence is reversed. Negation is a popular polarity shift. For example, let us consider a positive word “I like the story of that movie”.Now when we introduce a negation “don’t” before the sentiment word “like”,(i.e., “I don’t like the story of the movie”),the polarity of the entire sentence is reversed to negative. And for BOW model, since it considers text as a bundle of independent words, it considers the two polarity shifted sentences given above as almost similar. This is the polarity shift problem faced by BOW model.
Many approaches have been made for improving the performance of BOW model and for dealing with the polarity shift problem [2],[4],[5], [6], [7].However, most of these approaches needed eithercomplex linguistic knowledge or manual annotations.

This paper proposes a method to improve the performance of Dual Sentiment Analysis (DSA) [1] using the random forest classifier, which is an effective way for dealing with the polarity shift problem**.** DSA do not need such complex linguistic knowledge and extra manual annotations. The basic idea behind Dual Sentiment Analysis is considering the polarity reversed version along with the original review text.

First, there is a data expansion technique which is used to create sentiment reversed reviews. The original and reversed reviews are constructed in a one-to-one correspondence. Then a Dual Training (DT) technique is performed on this combined reviews for training the statistical classifier and Dual Prediction (DP) technique is performed for making predictions based on the training. Predictions are made by considering two sides of one review i.e., two things are measured, first, how positive/negative the original review is, and second, how negative/positive the reversed review is.

II. RELATED WORK

Sentiment analysis can be divided into four categories: document-level, sentence-level, phrase-level, and aspect-level sentiment analysis. Wilson et al. [9] focused on the phrase and aspect level sentiment analysis and studied the effects of complex polarity shift. They used a lexicon of words annotated with prior polarities, and identified the “contextual polarity” of the phrases. Nakagawa et al. [10] discussed a semi supervised model for sentiment analysis which predicts the polarity based on the dependency graph. The aspect-level sentiment analysis considers the polarity shift problem in both corpus based and lexicon-based methods [3].

Document- and sentence-level sentiment classification considers two types of methods: term-counting and machine learning methods. Term-counting methods, calculate the overall orientation of a text by summing up the orientation scores of each words in the text.It is easy to handle polarity shift in term counting method. One common way is to reverse the sentiment of the polarity shifted words and then find the sentiment score by summing up the sentiment score word by word [11].

Machine learning methods considers sentiment classification as a statistical classification problem. In machine learning method of sentiment classification, a text is represented by using bag-of words. Then the classifier is trained using supervised machine learning algorithm. Compared to term counting method, it is relatively complex to handle the polarity shift in the machine learning method. Das and Chen [2] suggested a method which adds a “NOT” after the words in the scope of negation. Example: the text “I don’t like tea”, will become “I like-NOT tea”.

Na et al. [8] proposed a method which handle negation by considering specific part-of-speech tag patterns. Ikeda et al. [4] proposed an idea based on machine learning method which uses a lexical dictionary extracted from General Inquirer to handle polarity-shifters.

To solve the problem of identifying the sense of a polysemic word based on the context of its occurrence, Word Sense Disambiguation (WSD) is used [12]. To some extent, WSD solves the problem of plain bag-of-words approach by considering thewhole sentence.

Li and Huang [6] proposed a method which first classifies each sentence in a text into a polarity-unshifted and a polarity-shifted part according to certain rules. Then they are represented as two bags-of words for sentiment classification. After that, they introduced a technique to classify the shifted and unshifted text. Classifiers are trained based on the two set. The polarity of the text is then predicted using an ensemble of two component classifiers.

III. PROPOSED METHOD

Main concept of proposed system is the Dual Sentiment Analysis (DSA)[1].Basic task of DSA is the Data Expansion Technique by which the reversed review of the original review is created. Both original and reversed reviews are used to train the statistical classifier .This process is called the Dual Training (DT). Testing is performed based on the Dual Training, which is called Dual Prediction (DP).Figure 1 shows the system architecture of the proposed system.Various Steps involved in this method are:

1. Data Preprocessing
2. Data Expansion
3. Dual Training
4. Dual Prediction

1. ***Data Pre Processing*:**

Preprocessing of text is the basic and important task of Natural Language Processing (NLP) and Information Retrieval (IR).Text pre-processing is the technique of modelling the text in to structured format which can be used for further processing. Various pre-processing steps performed here are:

a) Tokenization
b) Stop word removal
c) Repeated letters removal
d) Noise data removal
e) Stemming
f) POS tagging
g) Word Sense Disambiguation (WSD) checking.
h) Feature Vector Formation

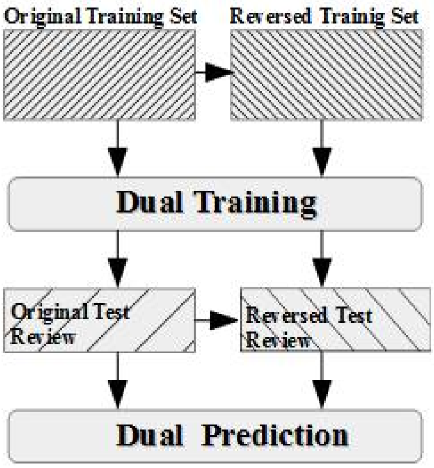

Fig.1. System Architecture

2. ***Data Expansion*:**

Data Expansion is the process of creating sentiment reversed review of the original review based on an antonym dictionary. Table 1 shows example of creating reversed review. Data expansion involves:

a) **Sentiment word reversion**

All sentiment words out of the scope of negation are reversed to their antonyms. Scope of negation is defined as the range from the negation word to the end of the sub-sentence.

b) **Handling negation:**

If there is a negation expression, first detect the scope ofnegation, and then remove the negation words (e.g., no, not, and don't). Thesentiment words in the scope of negation are not reversed.

c) **Label reversion:**

The class label is also reversed. (i.e. "Positive" to "Negative" and vice versa).

	Review Text	**Class**
Original Review	"I don't like this movie. It is boring"	Negative
Reversed Review	"I like this movie. It is interesting"	Positive

Table 1: Example for creating reversed review

3. ***Dual Training:***

In the training process, first, all the original training samples are reversed to their opposites using the data expansion technique. There is a one-to-one correspondence between the original and reversed reviews. The classifier is trained by maximizing a combination of the likelihoods of the original and reversed training samples. This process is called dual training (DT). The feature weights in DT are learnt by considering not only how likely is the review x to be positive/negative, but also how likely is, $\bar{x}$ the reversed review to be negative/positive. The classifier used here is the random forest classifier.

4. ***Dual Prediction:***

In the prediction stage, for each test sample x, we create a reversed test sample $\bar{x}$. Here, our aim was not to predict the class of $\bar{x}$. But instead, we used it to assist the prediction of x. This process is called dual prediction (DP). Let p (+|x) and p (-|$\bar{x}$) denote posterior probabilities of x and $\bar{x}$ respectively. In DP, predictions are made by considering two sides of a coin:

- When we want to measure how positive a test review x is, we not only consider how positive the original test review is (i.e., p(+|x)), but also consider how negative the reversed test review is (i.e., p(-|$\bar{x}$)).
- Conversely, when we measure how negative a test review x is, we consider the probability of x being negative (i.e., p(-|x)), as well as the probability of $\bar{x}$ being positive (i.e., p(+|$\bar{x}$)).

Then two component predictions is used as the dual prediction score, i.e.P(+|x,$\bar{x}$) and P(-|x,$\bar{x}$).

Let P_d(y |x,$\bar{x}$) denote the dual prediction of review x based on an already-trained DT model. In order to prevent DP algorithm from being damaged by low-confident predictions, instead of using all dual predictions P_d(y|x,$\bar{x}$) as the final output, we use the original prediction P_o (y|x) as an alternate. The final prediction is therefore defined as:

$$P_f(y|x) = \begin{cases} P_d(y|x,\bar{x}); & \text{if } \Delta p > 0 \\ P_o(y|x); & \text{otherwise} \end{cases}$$

Where Δp = P_d (y|x, $\bar{x}$) - P_o(y|x). That is, the prediction with a higher posterior probability will be chosen as the final prediction.The prediction process is shown in figure2.

The Lexicon-based Antonym Dictionary

In the languages where lexical resources are abundant, a straight forward way is to get the antonym dictionary directly from the well-defined lexicons, such as WordNet in English. WordNet is a lexical database which groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. Using the antonym dictionary, it is possible to obtain the words and their opposites. The WordNet antonym dictionary is simple and direct.

Feature Extraction

This paper uses two features for performing dual sentiment analysis. They are:

- **Unigrams:** This considers words one by one. For example, unigrams of thesentence “I like this movie.Best story” are “I”, “like”, “this”, “movie”, “best”and “story”.
- **Bigrams:** A group of two words is called bigrams. For example, bigrams ofthe sentence “I like this movie. Best story” are “I like”, “like this”, “thismovie”, “movie best” and “best story”

tf-idf

tf-idf stands for term frequency-inverse document frequency, and the tf-idf weight is often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus. Typically, the tf-idf weight is composed by two terms

- TF (Term Frequency): It measures how frequently a term occurs ina document. Since every document is different in length, it is possible thata term would appear much more times in long documents than shorter ones.Thus, the term frequency is often divided by the document length (i.e. the totalnumber of terms in the document) as a way of normalization:

$$\text{TF}(t, d) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document\ d}$$

- IDF (Inverse Document Frequency): It measures how important a term is.While computing TF, all terms are considered equally important. However it is known that certain terms, such as “is”, “of”, and “that”, may appear a lotof times but have little importance. Thus we need to weigh down the frequentterms while scale up the rare ones, by computing the following:

$$\text{IDF}(t, D) = \text{loge}\frac{\text{Total number of documents , D}}{\text{Number of documents with term t in it}}$$

- Then tf-idf is calculated as: tf-idf (t,d,D) = TF(t,d)* IDF(t,D)

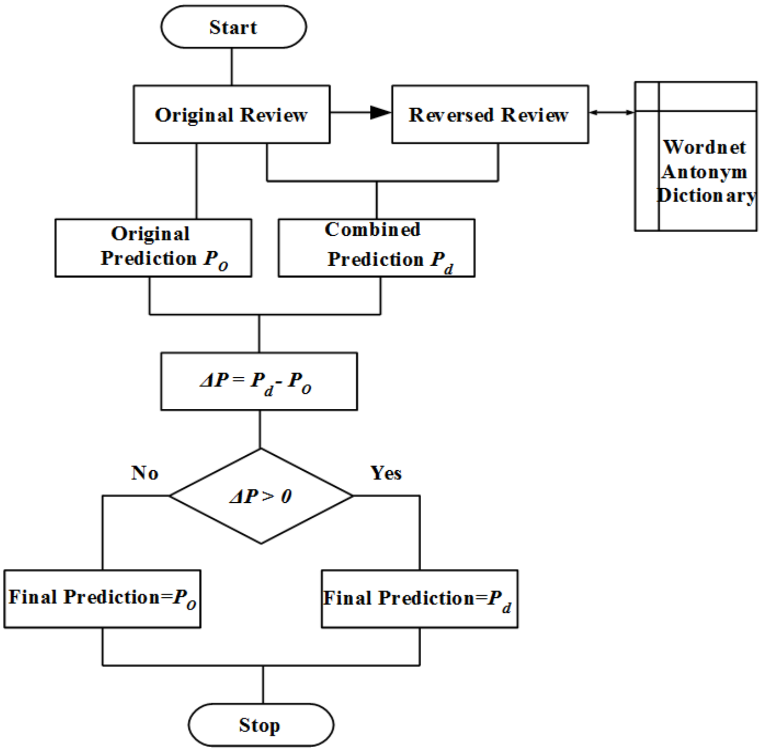

Fig. 2 Prediction process

Random Forest Classifier

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that is operated by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The random forest is an ensemble approach that can also be thought of as a form of nearest neighbour predictor.

Features of Random Forests are:

- It runs efficiently on large data bases.
- It gives estimates of what variables are important in the classification.

- It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.
- Generated forests can be saved for future use on other data.
- The capabilities of the above can be extended to unlabelled data, leading to unsupervised clustering, data views and outlier detection.
- Random forests does not over fit.

IV. EXPERIMENTAL RESULTS

Four Multi-Domain Sentiment datasets are used here. They contain product reviews taken from Amazon.com including four different domains: Apparel, DVD, Electronics and Kitchen. Each of the reviews is rated by the customers from Star-1 to Star-5. The reviews with Star-1 and Star-2 are labelled as Negative, and those with Star-4 and Star-5 are labelled as Positive. Each of the four datasets contains 1,000 positive and 1,000 negative reviews. Reviews in each category are randomly split up into 60%-40% (with 60% serving as training data and the remaining 40% serving as test data). Following four models are taken for evaluation:

- Normal Sentiment Analysis (NSA): Direct sentiment analysis that considers the original review only.
- Dual Sentiment Analysis with unigrams (DSA-UN): Here DSA is performed considering only unigrams as feature.
- Dual Sentiment Analysis with unigrams and bigrams (DSA-UB): In this case, DSA is performed considering both unigrams and bigrams as features.

From the above three systems, better performance is obtained for DSA with unigrams and bigrams (DSA-UB) compared to other two systems.

The results of the experiment performed using apparel, DVD, electronics and kitchen datasets are as shown in the table 2 below.

DATASET	NSA	DSA-UN	DSA-UB
APPAREL	82.207	83.558	84.234
DVD	84.234	84.684	85.135
ELECTRONICS	83.108	84.234	85.135
KITCHEN	84.685	86.909	88.10
AVERAGE	83.558	84.846	85.651

Table 2. Result analysis

Precision and recall for DSA-UB is shown figure 3.

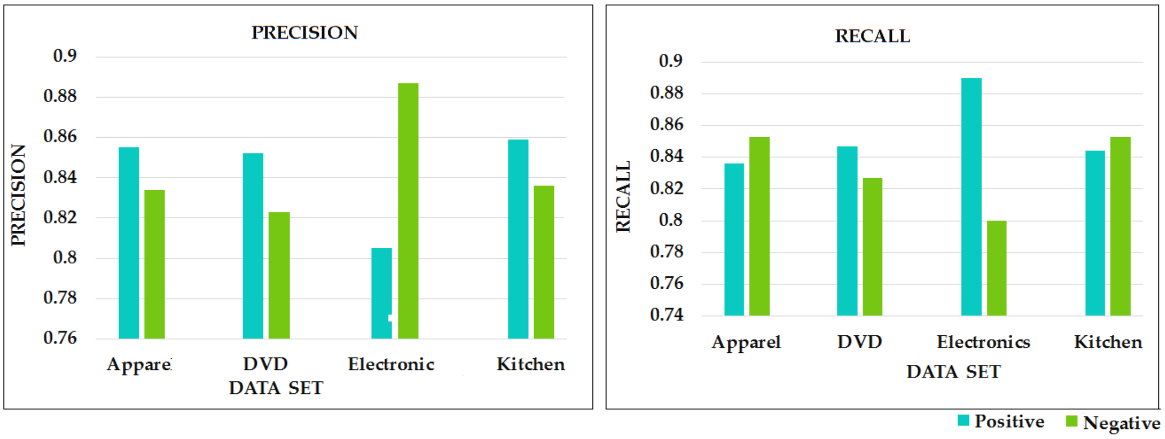

Fig 3.Precision and Recall of DSA-UB

Existing system used DVD, electronics and Kitchen datasets. Inorder to compare with the existing system, experiments are conducted using same datasets. Figure 4 shows the accuracy comparison between existing DSA [1] and proposed system and from the figure, it is clear that proposed system outperforms the existing system.

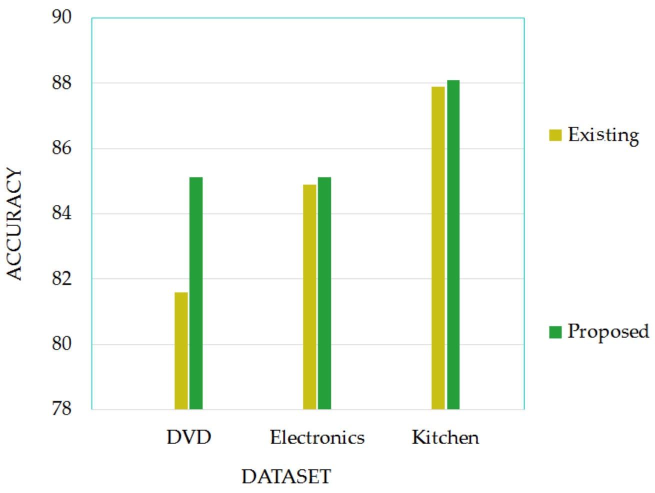

Fig 4. Accuracy Comparison-Existing Vs Proposed

V. Conclusion and Future Work

This work proposes a model for efficient sentiment classification using improved version of dual sentiment analysis (DSA) using the random forest classifier for addressing the polarity shift problem. The basic idea of DSA is to create reversed reviews that are sentiment-opposite to the original reviews, and then train the sentiment classifier using the original and reversed reviews in pairs. There is a one to one correspondence between original and reversed review in training (Dual Training) and prediction (Dual Prediction). An external antonym dictionary (WordNet antonym dictionary)is used for creating the reversed review. Word Sense Disambiguation is used to detect the correct sense of the word in the sentence. The features used in this model are unigrams and bigrams.For feature extraction, tf-idf is used. Experiments are conducted with normal sentiment analysis(NSA), dual sentiment analysis with unigrams only(DSA-UN)and dual sentiment analysis with unigrams and bigrams(DSA-UB).From the analyzed results, DSA-UB out performs the other two models. Finally when comparing the performance of existing DSA [1] and the proposed system, the proposed system shows better performance compared to the other.

REFERENCES

1. Rui Xia, Feng Xu, ChengqingZong, Qianmu Li, Yong Qi, and Tao L, Dual Sentiment Analysis: Considering Two Sides of One Review, Ieee Transactions On Knowledge And Data Engineering, Vol. 27, No. 8, August 2015.
2. S. Das and M. Chen, “Yahoo! for Amazon: Extracting market sentiment from stock message boards,” Proceedings of the Asia Pacific Finance Association Annual Conference, 2001.
3. X. Ding and B. Liu, "The utility of linguistic rules in opinion mining," Proceedings of the 30th ACM SIGIR conference on research and development in information retrieval (SIGIR), 2007.
4. D. Ikeda, H. Takamura, L. Ratinov, and M. Okumura, “Learning to Shift the Polarity of Words for Sentiment Classification,” Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), 2008.
5. A. Kennedy and D. Inkpen, “Sentiment classification of movie reviews using contextual valence shifters,” Computational Intelligence,vol. 22, pp. 110–125, 2006.
6. S. Li and C. Huang, “Sentiment classification considering negation and contrast transition,” Proceedings of the Pacific Asia Conference on Language, Information and Computation (PACLIC), 2009.
7. B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques,” Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP),pp. 79-86, 2002.
8. J. Na, H. Sui, C. Khoo, S. Chan, and Y. Zhou, “Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews,” Proceedings of the Conference of the International Society for Knowledge Organization (ISKO), 2004.
9. T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity:An exploration of features for phrase-level sentiment analysis,"Computational Linguistics, vol. 35, no. 3, pp. 399-433, 2009.
10. T. Nakagawa, K. Inui, and S. Kurohashi. “Dependency tree-basedsentiment classification using CRFs with hidden variables,” Proceedingsof the Annual Conference of the North American Chapter of the Associationfor Computational Linguistics (NAACL), pp. 786-794, 2010.
11. Y. Choi and C. Cardie, “Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis,” Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 793-801, 2008.
12. M. Stevenson and Y.Wilks,, Word-sense disambiguation, The Oxford Handbookof Comp. Linguistics, pp.249265.

BIOGRAPHY

Bajeela P.V received her B.Tech degree in Information Technology from Calicut University, in 2006 and doing M.Tech in Computer Science at KMCT College of Engineering, Calicut. Her areas of research interest include Data Mining and Information Retrieval.

Swaradh P is an Assistant Professor in the Department of Computer Science and Engineering at KMCT College of Engineering, Calicut and has an experience of 9 years in the academic field. He pursued his B.Tech Degree in Information Technology from Calicut University and ME Degree in Computer Science and Engineering from PSG College of Technology, Coimbatore. His subject expertise include Data Mining, Information Retrieval, Data Intensive Computing and Data Structures.