# THREE-STAGE SENTIMENT ANALYSIS BY POLARITY SHIFT DETECTION, ELIMINATION AND ENSEMBLE

**[1]POOJA K.MANNA, [2]SONALI BODKHE**

[1]M.Tech CSE Department, Nagpur University, India,
[2]CSE Department, Nagpur University, India
E-mail: [1]poojamanna16@gmail.com, [2]sonali.bodkhe@raisoni.net

**Abstract-** The volume of user-generated text on the Web in the form of reviews, blogs, and social networks has grown dramatically in recent years. This was mirrored by an increasing interest, from both the academic and the business world, in the field of sentiment analysis, which aims to automatically extract sentiment from natural language text and can be broadly categorized into knowledge-based or statistics-based. Bag-of-words (BOW) is the popular way to model text in statistical machine learning methods in sentiment analysis. However, the achievement of BOW sometimes remains bounded due to some primitive dearth in handling the polarity shift problem. The greater findings were that out of the classification algorithms assessed it was that the Random forest classifier provide the much more high classification accurateness for this domain. From the assessing of this study it can be concluded that the proposed machine learning and natural language processing techniques are an impressive and real time approach for sentiment analysis The polarity shift problem is a major factor that affects classification performance of machine-learning-based sentiment analysis systems .Proposing a three-stage cascade model to address the polarity shift problem in the context of document-level sentiment classification.
We first split each document into a set of sub sentences and build a hybrid model that employs rules and statistical methods to detect explicit and implicit polarity shifts, respectively. Secondly, proposing a polarity shift elimination method, to remove polarity shift in negations. Finally, we train base classifiers on training subsets divided by different types of polarity shifts, and use a weighted combination of the component classifiers for sentiment classification. On this basis, also proposing a dual training algorithm to make use of original and reversed training reviews in duality for learning a sentiment classifier, and a dual prediction algorithm to classify the test results by keeping in mind both of two phases of one result. An extended framework from polarity (positive-negative) classification to 3-class (positive-negative-neutral) classification has to be done, by taking the neutral reviews into consideration. Finally, a corpus-based approach is constructed for pseudo-antonym dictionary, which elimination of Dual Sentential Approch's dependency on an external antonym dictionary for review reversion.

**Keywords**- DSA, Polarity Shift, Random Forest, Knowledge-Based, Bag-Of-Words (BOW), PSDEE.

## I. INTRODUCTION

In late years, with the increasing of online reviews gettable on the internet , sentiment analysis and opinion mining, as a special text mining task for studying or interpret the subjective attitude (i.e., sentiment) assert by the text, is becoming a difficulty in the field of data mining and natural language processing . Sentiment classification is a primary task in sentiment analysis, with its focus to classify the sentiment (e.g., positive or negative) of a given text. The extensive practice in sentiment classification chase the techniques which is widely used in topic-based text classification, where the bag-of-words (BOW) model as a rule used for text representation. In the BOW model, a review text is represented by a vector of autonomous words The statistical machine learning algorithms (such as na€ıve Bayes, maximum entropy classifier, and support vector machines) are then applied to train a sentiment classifier. Although the BOW model is very elementary and quite efficient in topic-based text classification, it is actually not very suitable. However, the BOW model disrupts word order, breaks the syntactic structures and discards some semantic information of the text. Therefore, it brings about some fundamental deficiencies including the polarity shift problem. Polarity shift refers to a linguistic phenomenon in which the polarity of sentiment can be reversed (i.e., positive to negative or vice versa) by some special linguistic structures called polarity shifters, e.g., negation ("*I don't like this movie*") and contrast ("*Fairly good, but not my style*"). Obviously, in the BOW model, it is hard to capture the sentiment reversion caused by polarity shifters, because two sentiment-opposite texts (e.g., "*I don't like this movie*" and "*I like this movie*") are regarded to be very similar in the BOW representation.

However, most of them focused on either modeling polarity shift in phrase/subsentence-level sentiment classification, or encoding polarity shift in rule-based term-counting methods. Even there were few of them dealing with polarity shift by using machine learning methods for document-level sentiment classification, their performances were not satisfactory, e.g., the improvements were less than 2% after considering polarity shift. In this work, we propose a three-stage model, namely Polarity Shift Detection, Elimination and Ensemble (PSDEE), to address polarity shift for document-level sentiment classification. Firstly, we propose a hybrid polarity shift detection approach, which employs a rule-based method to detect some polarity shifts such as explicit negations and contrasts, and a statistical method to detect some implicit polarity shifts such as sentiment

inconsistencies. Secondly, we propose a novel polarity shift elimination algorithm to eliminate polarity shifts in negations. For example, the review "*this movie is not interesting*" is reversed to "*this movie is boring*". It can make the BOW representation more feasible due to the elimination of negations. Finally, we separate the training and test data into four component subsets, i.e., negation subset, contrast subset, sentiment-inconsistency set as well as polarity-unshifted subset, and train the base classifiers based on each of the component subset. A weighted ensemble of four component predictions are finally used in testing, with the motivation to distinguish texts with different types of polarity shifts such that the polarity-unshifted part will have a higher weight, while the polarity-shifted part will have a lower weight in sentiment prediction. We systematically evaluate our PSDEE model by conducting experiments on four sentiment datasets, three kinds of classification algorithms and two types of features.

## II. NECESSITY AND OBJECTIVES

Proposing a three-stage model, namely Polarity Shift Detection, Elimination and Ensemble (PSDEE), to address polarity shift for document-level sentiment classification. Firstly, we propose a hybrid polarity shift detection approach, which applies a rule-based method to recognize some polarity shifts such as explicit negations and contrasts, and a statistical method to recognize some implicit polarity shifts such as sentiment difference. Secondly, we propose a novel polarity shift elimination algorithm to eliminate polarity shifts in negations. For example, the review "*this movie is not interesting*" is reversed to "*this movie is boring*". It can make the BOW representation more feasible due to the elimination of negations. Finally, we separate the training and test data into four component subsets, i.e., negation subset, contrast subset, sentiment-inconsistency set as well as polarity-unshifted subset, and train the base classifiers based on each of the component subset and also proposing a data expansion method by organizing sentiment-reversed reviews.

The original and reversed reviews are established in correspondence with one to one mapping. That is, measuring not only how positive/negative the original review is, but also how negative/positive the reversed review is. Further exploring our framework from polarity (positive-negative) classification to 3-class (positive-negative-neutral) sentiment classification, by taking the neutral reviews into consideration in both dual training and dual prediction. A typical advent to sentiment analysis is to start with a lexicon of positive and negative words and phrases. In these lexicons, entries are tagged with
their a priori *prior polarity*: out of context, does the word seem to invoke something positive or something negative. For example, *beautiful* has a positive prior polarity, and *horrid* has a negative prior polarity. However, the *contextual polarity* of the phrase in which a word appears may be different from the word's prior polarity.

## III. LITERATURE REVIEW

Focusing on the phrase/subsentence- and aspect-level sentiment analysis, Wilson et al. [15] discussed effects of complex polarity shift. They began with a lexicon of words with established prior polarities, and identify the "contextual polarity" of phrases, based on some refined annotations. Choi and Cardie [4] further combined different kinds of negators with lexical polarity items though various compositional semantic models, both heuristic and machine learned, to improved subsentential sentiment analysis. Nakagawa et al. [14] developed a semi-supervised model for subsentential sentiment analysis that predicts polarity based on the interactions between nodes in dependency graphs, which potentially can induce the scope of negation. In aspect-level sentiment analysis, the polarity shift problem was considered in both corpus- and lexicon based methods [8], [9], [10], [13].

A frequent threat in the sentiment analysis of a text is to analyze, on those form of the text which are in some way representative of the accent of the entire text. In earlier work these has been done in the area of characterizing words and phrases according to their emotional accent (Turney and Littman, 2003; Turney, 2002; Kamps et al., 2002; Hatzivassiloglou and Wiebe, 2000; Hatzivassiloglou and McKeown, 2002; Wiebe, 2000), but in many domains of text, the values of individual phrases may handle little relation to the overall sentiment expressed by the text. Pang et al. (2002)'s treatment of the task as analogous to topical classification caption and the difference between the two tasks. Sources of ambiguous phrases include what Pang et al. (2002) refer to as "thwarted expectations" narrative, where emotive effect is attained by emphasizing the comparison between what the reviewer expected and the for real experience

## IV. METHODOLOGY

[1] Logistic regression is a classification algorithm that can be trained as long as expect the features to be approximately linear and the trouble to be linearly separable. It can do some feature engineering to turn wide range of non-linear features into linear much efficiently. It is also very robust to noise and can avoid over fitting and even do feature selection by using l2 or l1 regularization. Logistic regression can also be used in Big Data scenarios since its performance is efficient and can be distributed using, for example, ADMM . A final advantage of LR is that

the output can be evaluated as a probability. This is something that comes as a nice side effect since you can use it, for example, for ranking instead of classification.

Even in a case where you would not expect Logistic Regression to work 100%, run a simple l2-regularized LR to come up with a baseline before you go into using "fancier" approaches.

[2] Support Vector Machines (SVMs) uses a various loss function from Logistic Regration. They are also identified differently (maximum-margin). However, in traditional way, an SVM with a linear kernel is not so much different from a Logistic Regression . The main moderation is to use an SVM instead of a Logistic Regression is because the problem might not be linearly separable. In that case ,will have to use an SVM with a non linear kernel (e.g. RBF). The veracity is that a Logistic Regression can also be used with a different kernel, but at that point you might be better off going for SVMs for functional reasons. Another related reason to use SVMs is if you are in a highly dimensional space. The term *semantic orientation* (SO) (Hatzivassilo glou and McKeown, 2002) refers to a real number measurement of the positive or negative sentiment expressed by a word or phrase. In the present work, the approach taken by Turney (2002) is used to derive such values for selected phrases in the text. This approach is simple and surprisingly effective. Moreover, is not restricted to words of a accurate part of speech, nor even restricted to single words, but can be used with multiple word phrases. Usually, two word phrases conforming to particular part-of-speech templates representing possible descriptive composition are used. Unfortunately, the better downside of SVMs is that they can be painfully inefficient to train. So, would not commended them for any its problem.

[3] The third category of algorithms used in decision making is the Tree Ensembles. This basically comprises of two distinct algorithms: Random Forests and Gradient Boosted Trees. Tree Ensembles have different advantages over LR. One of the main benefit is that they do not expect linear features or even features that interact linearly. Something it is not defined that LR is can hardly handle categorical (binary) features. Tree Ensembles, because they are nothing more than a bundle of Decision Trees combined, can hold this very well. The other main advantage is that, because of how they are build by (using bagging or boosting) these algorithms ,they handle very well high dimensional spaces as well as large number of training examples.

As for the difference between Random Forests and Gradient Boosted Decision Trees, GBDTs will usually perform better, but they are harder to get right. More concretely, GBDTs have more hyper-parameters to tune and are also more prone to over fitting. RFs can almost work beyond the limit and that is one reason why they are very popular.

## V. DATA FLOW DIAGRAM

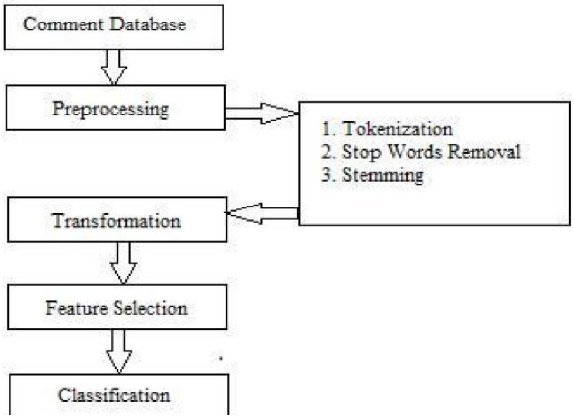


**Fig1:Steps and techniques used in sentiment classification.**

Given input is an Social network messages dataset and particular of a social network site to be searched as an input. Sometimes the dataset are noisy. For that pre processing of data is done by preparing and cleaning the data of dataset for classification.. Apply Random forest algorithm and extract features from dataset . The dataset is trained by the presented algorithm and matching of object is also done by it.

## CONCLUSION AND FUTURE WORK

In this paper review of object identification, feature extraction and tracking of object has been studied. The future work is to analyzing and extracting of feature in less time with high performance and more type of sentiments are analyzed. In existing system there is more occlusion, less feature extraction but in proposed system there will be occlusion reduction, and more feature extraction in less time. Random forest classifier is an ensemble of decision trees. Each decision tree can be constructed independently making this an embarrassingly parallelizable problem. Consequently, we were able to very easily incorporate where each process was responsible for creating a decision tree. When classifying a piece of text, the random forest passes the item to each decision tree and the output is the majority vote.

## REFERENCES

[1] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," Comput. Intell., vol. 22,pp. 110–125, 2006.

[2] D. Ikeda, H. Takamura, L. Ratinov, and M. Okumura,"Learning to shift the polarity of words for sentiment classification," in Proc.Int. Joint Conf. Natural Language Process., 2008.

[3] C. Lin, Y. He, R. Everson, and S. Ruger, "Weakly supervised joint sentiment-topic detection from text," IEEE Trans. Knowl.DataEng., vol. 24, no. 6, pp. 1134–1145, Jun. 2012.

[4] B. Liu, "Sentiment analysis and opinion mining," in Synthesis Lectures on Human Language Technologies, vol. 5, no. 1. San Rafael, CA,USA: Morgan & Claypool, 2012, pp. 1–165.R. Morante and W. Daelemans, "A metalearning approach to processing the scope of negation," in Proc. 30th Conf. Comput. Natural Language Learning, 2009, pp. 21–29.

[5] Abbasi, S. France, Z. Zhang, and H. Chen, "Selecting attributes for sentiment classification using feature relation networks," IEEE Trans. Knowl. Data Eng., vol. 23, no. 3, pp. 447–462, Mar. 2011.

[6] E. Agirre, and D. Martinez, "Exploring automatic word sense disambiguation with decision lists and the web," in Proc. COLING Workshop Semantic Annotation Intell. Content, 2000, pp. 11–19.

[7] J. Cano, J. Perez-Cortes, J. Arlandis, and R. Llobet, "Training set expansion in handwritten character recognition," in Proc. Struct., Syntactic, Statistical Pattern Recognit., 2002, pp. 548–556.

[8] Y. Choi and C. Cardie, "Learning with compositional semantics as structural inference for subsentential sentiment analysis," in Proc. Conf. EmpiricalMethodsNatural Language Process., 2008, pp. 793–801.

[9] Councill, R. MaDonald, and L. Velikovich, "What's great andwhat's not: Learning to classify the scope of negation forimproved sentiment analysis," in Proc. Workshop Negation SpeculationNatural Lang. Process., 2010, pp. 51–59.

[10] S. Das and M. Chen, "Yahoo! for Amazon: Extracting market sentimentfrom stock message boards," in Proc. Asia Pacific FinanceAssoc. Annu.Conf., 2001.

[11] K. Dave, S. Lawrence and D. Pen-nock, "Mining the peanut gallery:Opinion extraction and semantic classification of productreviews," in Proc. Int. World Wide Web Conf., 2003, pp. 519–528.

[12] X. Ding and B. Liu, "The utility of linguistic rules in opinion mining,"in Proc. 30th ACM SIGIR Conf. Res. Development Inf. Retrieval,2007, pp. 811–812. X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach toopinion mining," in Proc. Int. Conf. Web Search Data Mining, 2008,pp. 231–240.Comput. Linguistics, vol. 35, no. 3, pp. 399–433, 2009

★★★