**Sentiment Analysis using Ensemble Classifiers**

Kaushik Hande1, Prof. A. G. Phakatkar2

Department of Computer Engineering, Pune Institute of Computer Technology, Pune, Maharashtra, India.

**ABSTRACT**: The reviews and blogs obtained from social networking and online marketing sites, act as an important source for analysis and improved decision making. These reviews are mostly unstructured by nature and thus, need processing. Bag of words (BOW) model of features are used for processing in machine learning algorithms.. Reviews are classified as either positive or negative concerning a query term. This approach is useful for consumers who can use sentiment analysis to search for products, for companies that aim at monitoring the public sentiment of their brands, and for many other applications. In this work, three different machine learning algorithms such as Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machine (SVM) are considered for classification of sentiments of reviews. Classifier ensembles formed by Naive Bayes, SVM, and Logistic Regression improves the classification accuracy of sentiment analysis.

KEYWORDS: Sentiment Analysis, Natural Language Processing, Machine Learning.

1. **Introduction**

Sentiment is an attitude, thought, or judgement prompted by feeling. Sentiment analysis is also known as opinion mining, it involves studying of peoples sentiment towards certain entities. From a perspective of a user, people are able to express their views through various social media, such as forums, micro-blogs, or online social networking sites [4].With the advent of web 2.0 techniques, users started prefering to share their opinions on the Web. These user-generated and sentiment-rich data are valuable to many applications like credibility analysis of news sites on the web, recommendation system, business and government intelligence etc. At the same time, it brings urgent need for detecting overall sentiment inclinations of documents generated by users, which can be treated as a classification problem. Sentiment analysis includes several subtasks which have seen a great deal of attention in recent years :

1. To detect whether a given document is subjective or objective.

2. To identify whether given subjective document express a positive opinion or a

negative opinion.

3. To determine the sentiment strength of a document, such as strongly negative, weakly negative, neutral, weakly positive and strongly positive.

In this work we are focusing on second subtask. Besides individuals on social media marketers also need to monitor all media for information related to their brands whether it’s for public relations activities, fraud violations, or competitive intelligence. Thus, aside from individuals, sentiment analysis is also the need of companies which are anxious to understand how their products and services are perceived by the public.

The movie reviews are mostly in the text format and unstructured in nature. Thus, the stop words and other unwanted information are removed from the reviews for further analysis. These reviews goes through a process of vectorization in which, the text data are converted into matrix of numbers. These matrices are then given input to different machine learning classifiers for classification of the reviews.

Many researchers have focused on the use of traditional classifiers, like Naive Bayes, Maximum Entropy, and Support Vector Machines to solve such problems. In this work, we show that the use of ensembles of multiple base classifiers can improve the accuracy of review sentiment classification.

1. **Related work**

According to the levels of granularity, tasks in sentiment analysis can be divided into four categorizations: document- level, sentence-level, phrase-level, and aspect-level sentiment analysis.

For document and sentence-level sentiment classification, there are two main types of methods : term-counting and machine learning methods [2] [3] [4] [15]. In term-counting methods, the overall orientation of a text is obtained by summing up the orientation scores of content words in the text, based on manually-collected or external lexical resources [6] [9]. In machine learning methods, sentiment classification is regarded as a statistical classification problem, where a text is represented by a bag-of-words; then, the supervised machine learning algorithms are applied as classifier [3]. The use of ensembles of multiple base classifiers, combined with scores obtained from lexicons, can improve the accuracy of sentiment classification [15][14]. Accordingly, the way to handle polarity shift also differs in the two types of methods.

The term-counting methods [16] can be easily modified to include polarity shift. One common way is to directly reverse the sentiment of polarity-shifted words, and then sum up the sentiment score word by word [10]. Compared with term counting methods, the machine learning methods are more widely discussed in the sentiment classification literatures. However, it is relatively hard to integrate the polarity shift information into the BOW model in such methods. For example, Das and Chen[2] proposed a method by simply attaching “NOT” to words in the scope of nega-

tion, so that in the text “I don’t like this book”, the word “like” becomes a new word “like-NOT”. Yet Pang et al. [3] reported that this method only has slightly negligible effects on improving the sentiment classification accuracy.

1. **PROPOSED SOLUTION**

The reviews of tripadvisor dataset is processed to remove the stop words and unwanted information from dataset. The textual data is then transformed to a matrix of number using vectorization techniques. Further, training of the dataset is carried out using machine learning algorithm.

***Preprocessing***

The text reviews sometimes consist of absurd data, which need to be removed, before considered for classification. The usually identified absurd data are:

1. Stop words: They do not play any role in determining the sentiment.

2. Numeric and special character: In the text reviews, it is often observed that there are different numeric (1,2,...5 etc.) and special characters which do not have any effect on the analysis. But they often create confusion during conversion of text file to numeric vector.

***TF-IDF***

After the preprocessing of text reviews, reviews are represented by a table in which the columns represent the terms (or existing words) in the reviews and the values represent their frequencies. Therefore, a collection of reviews after thepreprocessing step addressed later can be represented as illustrated in Table 2, in which there are n reviews and m terms.Each review is represented as review i =(a i1 , a i2 , ..., a im ), where a ij is the frequency of term t j in the review i . This value can be calculated in various ways.

1. CountVectorizer: It converts the text reviews into a matrix of token counts. It implements both tokenization and occurrence counting. The output matrix obtained after this process is a sparse matrix.

**Classifiers**

Naive Bayes, Support vector machine and Maximum Entropy are used as classifiers

for sentiment analysis .

• Naive Bayes (NB) method: This method is used for both classification as well as training purposes. This is a probabilistic classifier method based on Bayes’ theorem. In this work, multinomial Naive Bayes classification technique is used. Multinomial model considers word frequency information in document for analysis, where a document is considered to be an ordered sequence of words obtained from vocabulary ‘V’. The probability of a wordevent is independent of word context and it’s position in the document.

• Support vector machine (SVM) method: This method analyzes data and defines decision boundaries by having hyper-planes. In binary classification problem, the hyper-plane separates the document vector in one class from other class, where the separation between hyper-planes is desired to be kept as large as possible.

• Maximum entropy (ME) method: In this method, the training data is used to set constraint on conditional distribution. Each constraint is used to express characteristics of training data. In Maximum Entropy (ME) if a word occurs frequently in a class, the weight of word-class pair becomes higher in comparison to other pairs. These highest frequency word-class pairs are considered for classification purpose. The movie reviews of acl IMDb dataset is considered for analysis, using the machine learning algorithms discussed. Then different variation of the n-gram methods i.e., unigram, bigram, trigram, unigram + bigram, unigram + trigram, and unigram + bigram + trigram are applied to obtain the result.

**Ensemble Classifiers:**

In practice, classifiers are built to classify unseen data, usually referred to as a target dataset. In a controlled experimental setting, a validation set represents the target set. Actually, in controlled experimental settings the target set is frequently referred to as either a test or a validation set. These two terms have been used interchangeably, sometimes causing confusion. In our study, we assume that the target/validation set has not been used at all in the process of building the classifier ensembles. Once the base classifiers have been trained, a classifier ensemble is formed by the average of the class probabilities obtained by each classifier or the majority voting.

1. **Implementation Details**
2. **Results**

The proposed CAMShift algorithm with Kalman filter is implemented with OpenCV. The proposed algorithm is compared with two metrics accuracy and precision on the basis of total number of a gesture actually displayed and total number of the gesture predicted by the system correctly. We considered the actual and the predicted gestures with different users to construct confusion matrix. Our results shows that,

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Measures | Web Camera | | | | | | | Leap Camera | | | | | |
| 1 | 2 | 3 | 4 | 5 | Average | 1 | | 2 | 3 | 4 | 5 | Average |
| Accuracy | 0.70 | 0.60 | 0.60 | 0.60 | 0.70 | 0.64 | 0.80 | | 0.90 | 0.90 | 0.60 | 0.80 | 0.80 |
| Precision | 0.75 | 0.67 | 0.71 | 0.67 | 0.60 | 0.68 | 0.80 | | 1.00 | 0.80 | 0.60 | 0.83 | 0.806 |

Table 1

as calculated in Table 1, Leap Motion controller’s accuracy, as well as precision, is more than Web-camera module for Human Computer Interaction system.

1. **Conclusion and Future Work**

It makes an attempt to classify movie reviews using different supervised machine learning algorithm. We also used CountVectorizer to improve the accuracy of classification. This algorithm is further applied using n-gram approach on IMDb dataset. It is observed that as the value of ’n’ in n-gram increases the classification accuracy decreases i.e., for unigram and bigram, the result obtained using the algorithm is remarkably better; but when trigram classification are carried out, the value of accuracy decreases. Also ensemble approach increases the classification accuracy.

**References**

1. R. Xia, F. Xu, C. Zong, Q. Li, Y. Qi and T. Li, “ Dual sentiment analysis : Considering two sides of one review, ” in IEEE transactions on knowledge and data engineering, vol. 27, no. 8, pp. 2120 - 2133, 2015.
2. S. Das and M. Chen, “ Yahoo ! for Amazon: Sentiment extraction from small talk on the web, ” Management science , Vol.53, issue no.9, pp. 1375-1388, 2007.
3. Pang, L. Lee, and S. Vaithyanathan, “ Thumbs up? : Sentiment classification using machine learning techniques, ”Proceedings of the ACL-02 conference on Empirical methods in natural language processing, pp. 79-86, 2002.
4. B. Pang and L. Lee, “ Opinion mining and sentiment analysis, ” Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1-135, 2008.
5. R. Xia, T. Wang, X. Hu, S. Li, and C. Zong, “ Dual Training and Dual Prediction for Polarity Classification,” Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL - 02) pp. 521-525, 2013.
6. P. Turney, “ Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, ” Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), pp. 417-424, 2002.
7. M. Li and C. Huang, “ Sentiment classification considering negation and contrast transition, ” Proceedings of the Pacific Asia Conference on Language, Information and Computation (PACLIC), pp. 307-316, 2009.
8. Li, S. Lee, Y. Chen, C. Huang and G. Zhou, “ Sentiment Classification and Polarity Shifting, ” Proceedings of the International Conference on Computational Linguistics (COLING), pp. 635-643, 2010.
9. D. Turney and Michael L. Littman, “ Un-supervised learning of semantic orientation from a hundred-billion-word corpus, ” Technical Report EGB-1094, National Research Council Canada, arXiv preprint cs/0212012, 2002.
10. Yuan Wang, Zhaohui Li, Jie Liu, Zhicheng He, Yalou Huang and Dong Li, “ Word Vector Modeling for Sentiment Analysis of Product Reviews, ” Natural Language Processing and Chinese Computing 2014, pp. 168-180, 2014.
11. A Tripathy, A Agrawal, SK Rath , “Classification of sentiment reviews using n-gram machine learning approach, ” in Expert Systems with Applications Volume 57, pp. 117-126, 2016.
12. Salvetti, Franco, Stephen Lewis, and Christoph Reichenbach. “Automatic opinion polarity classification of movie,” Colorado research in linguistics 17, no. 2004.
13. Xia, Rui and Wang, Cheng and Dai, Xinyu and Li, Tao, “ Co-training for Semi-supervised Sentiment Classification Based on Dual-view Bags-of-words Representation, ” Association for Computational Linguistics (ACL 1), pp. 1054-1063, 2015.
14. Rui Xia, Feng Xu, Jianfei Yu, Yong Qi and Erik Cambria, “ Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis, ” Information Processing & Management 52, no. 1, pp. 36 - 45, 2016
15. Rui Xia, Chengqing Zong and Shoushan Li, “ Ensemble of feature sets and classification algorithms for sentiment classification, ” Information Sciences 181, no. 6 pp. 1138-1152, 2011
16. L. Polanyi and A. Zaenen, “ Contextual lexical valence shifters, ” in Proc. AAAI Spring Symp. Exploring Attitude Affect Text, pp. 110, 2004.
17. Pang and Lee : Movie Review Dataset, Available online : http://boston.lti.cs.cmu.edu/classes/95-865-K/HW/HW3/movie-pang02.zip

**BIOGRAPHY**

**Kaushik Hande** is a student pursuing M.E. in the Computer Engineering Department, Pune Institute of Computer

Technology, Pune. His research interests are Data Mining, Data Analysis and Machine Learning.

**Prof. A. G. Phakatkar** is a Assistant Professor in the Computer Engineering Department, Pune Institute of Computer

Technology, Pune. Her research interests are Big Data and image processing.