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 9, September 2017

Sentiment Analysis using Ensemble Classifier

Kaushik Hande

ME Student, 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), Logistic Regression (LR), and Support Vector Machine (SVM) are considered for classification of sentiment 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.

I. INTRODUCTION

Sentiment is an attitude, thought, or judgement prompted by feeling. Sentiment analysis is also known as opinion mining, it involves studying of people's 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 on-line social networking sites [4].With the advent of web 2.0 techniques, users started preferring 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, Logistic Regression 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.

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 9, September 2017

II. 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 [8].

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 [12], 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 literature [13]. 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 negation, 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.

III. PROPOSED SOLUTION

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

1) Reviews Cleaning

The text reviews sometimes consist of useless data, which needs to be removed, before considered for classification. The text data usually consists of:

1. Stop words: They play no role in classification of 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 [2] [3].

2) TF-IDF

After the pre-processing 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 the pre-processing step can be represented as illustrated in Table 1, 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. Here we used TF-IDF feature instead of frequency count [11].

For example consider 2 reviews such as

1) This movie is boring.

2) This movie is interesting.

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 9, September 2017

They are represented in matrix form as shown below:

Review #	Interesting	Boring	Is	Movie	This
1	0.0	0.6300	0.4482	0.4482	0.4482
2	0.6300	0.0	0.4482	0.4482	0.4482

Table 1: Sample Document Term Matrix with TF-IDF features

In information retrieval, **tf–idf**, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The output matrix obtained after this process is a sparse matrix such as this:

ex. 1: “This movie is boring” is represented by sparse matrix as below,

(0, 4) 0.448320873199
(0, 3) 0.448320873199
(0, 2) 0.448320873199
(0, 1) 0.630099344518

ex. 2: “This movie is interesting” is represented by sparse matrix as below,

(1, 4) 0.448320873199
(1, 3) 0.448320873199
(1, 2) 0.448320873199
(1, 0) 0.630099344518

3) Classifiers

Naive Bayes, Support vector machine and Logistic regression are used as classifiers for sentiment analysis.

- Naive Bayes (NB) method: Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features. An advantage of naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification [1] [7].
- Support vector machine (SVM) method: They are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall [1][5].
- Logistic regression: It is a regression model where the dependent variable (DV) is categorical. This article covers the case of a binary dependent variable—that is, where the output can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. Cases where the

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 9, September 2017

dependent variable has more than two outcome categories may be analysed in multinomial logistic regression, or, if the multiple categories are ordered, in ordinal logistic regression.[1] In the terminology of economics, logistic regression is an example of a qualitative response/discrete choice model.

The reviews of trip advisor dataset are 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 [1] [5].

4) Ensemble Classifier:

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 majority voting of the class obtained by each classifier.

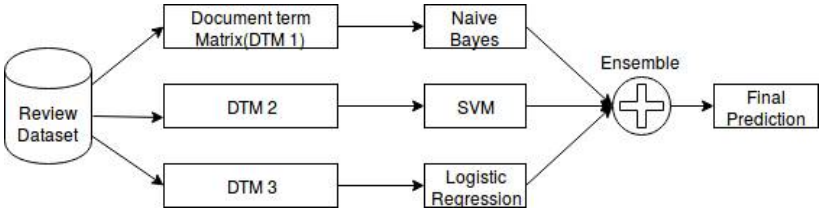

Fig: 1 Ensemble Classifier

IV. IMPLEMENTATION DETAILS

Dataset used is obtained from HackerEarth.com competition of Predict the Happiness for trip advisor dataset. [17] In this challenge, you have to predict if a customer is happy or not happy. The data consists of 38932 reviews. Of which 80% of the reviews are used for training purpose and 20% is used for test data.

The data from csv file is loaded into the python environment using import csv. Text reviews are converted into matrix form of reviews and words using countvectorizer of sklearn package. In this matrix reviews are represented by rows and word occurrence in that review is represented by columns. Each cell consists of occurrence frequency of particular word in that review. This matrix is converted into TF-IDF matrix. TF-IDF is known to represent text data more accurately. It gives importance to a word with respect to the entire corpus.

This matrix is applied to the machine learning algorithms along with labels for training purpose. The training data has 31145 reviews. The trained model obtained from training is applied on the test data. The test data has 7787 unlabelled reviews. The results obtained are compared with the original labels of the reviews and accuracy is calculated.

V. RESULTS

Below is the table of accuracy results obtained for different features such as unigram, bigram, trigram, unigram + bigram, bigram + trigram and unigram + bigram + trigram. Naïve Bayes performed poorly of all the algorithms. Logistic regression had better results than all the others. Ensemble performed better than Naïve Bayes and Support

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 9, September 2017

Vector Machine. Most effective feature to use for creating a matrix is unigram+bigram. All the algorithms performed better when unigram+bigram feature of vector was used. Highest accuracy attained is **89.34** by using unigram+bigram feature with logistic regression.

Features	Naive Bayes	Support Vector Machine	Logistic Regression	Ensemble
Unigram	79.29	87.04	88.49	87.90
Bigram	80.93	86.88	87.94	87.70
Trigram	83.55	83.95	84.90	85.24
Unigram+Bigram	79.84	88.26	89.34	88.89
Bigram+Trigram	81.73	86.21	87.68	87.26
Unigram+Bigram+Trigram	80.51	87.98	89.24	88.81

Table 1: Accuracy results of different algorithms.

VI. CONCLUSION

This work attempts to classify text reviews using three different machine learning algorithms. TF-IDF is used as features to these algorithms. Further unigram, bigram, trigram and combination of these are used as input to these algorithms. All the algorithms show best results when unigram + bigram approach is used. It is also observed that as the value of 'n' in n-gram increases the classification accuracy decreases. The ensemble algorithm which is based on majority voting gives better results as compared to other classifiers.

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 ChristophReichenbach. "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, FengXu, 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, ChengqingZong and Shoushan Li, " Ensemble of feature sets and classification algorithms for sentiment classification, " Information Sciences 181, no. 6 pp. 1138-1152, 2011

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 9, September 2017

16. L. Polanyi and A. Zaenen, "Contextual lexical valence shifters," in Proc. AAAI Spring Symp. Exploring Attitude Affect Text, pp. 110, 2004.
17. Hackerearth:Predict the Happiness:https://www.hackerearth.com/challenge/competitive/predict-the-happiness/machine-learning/predict-the-happiness/

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.