ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940
www.ijsem.org
Vol 15, Issuse.2 April 2021

A Comparative Study on Improving Sentiment Analysis using Machine Learning Classifiers

Chandu Delhipolice , Malleswari, D Purnima

Abstract:
Sentiment analysis, which seeks to ascertain the general public's emotional response to a given topic, is a cornerstone of NLP. Several criteria, including feature selection, model selection, and hyperparameter tuning, greatly affect the performance of ML classifiers, even if they have shown promising outcomes in sentiment analysis applications. Improving sentiment analysis using machine learning classifiers is the goal of this study, which does a thorough comparative analysis. We investigate several feature extraction approaches, test many classifiers, and tweak hyperparameters for optimal performance. Our findings give important insight into the most effective approaches to sentiment analysis and provide practical suggestions for improving current systems.

Keywords: Topics covered include machine learning, sentiment analysis, NLP, and the OneR method.

I. Introduction

Finding the overarching tone or emotional undercurrent in text data retrieved from sources like news stories, social media posts, and product reviews is the goal of sentiment analysis, also known as opinion mining, a subfield of natural language processing (NLP). This endeavor has taken on more significance in recent years because to the growing volume of user-generated content online and the need for companies and corporations to comprehend public opinion. Machine learning classifiers, which can automatically identify positive, negative, or neutral text, are now the most prevalent approach to sentiment analysis. There are a lot of factors that could significantly impact how well these classifiers work, including the classifier algorithm, the features used, and the hyperparameter settings. It is crucial to optimize these components in order to get trustworthy sentiment analysis results. Imparting subjective information from textual data, sentiment analysis is a subfield of machine learning and natural language processing (NLP). Opinion mining is another name for it. Numerous applications depend on it, including monitoring social media, evaluating customer comments, assessing items, and maintaining brand reputation. Sentiment analysis aims to detect emotional tones, sometimes called sentiments, which might be positive, negative, or neutral. Companies, groups, and individuals looking for statistics on public opinion and decision-making may find this information useful.

Associate Professor[1] , Assistant Professor[2,3]
(Department of Computer Science Engineering)
priyadarshini institute of technology

Recent advances and emerging trends in sentiment analysis include domain adaptation, multi-modal sentiment analysis (which integrates text with supplementary data sources like images and audio), and the incorporation of sentiment analysis into real-time applications. After the poll is over, we will go over the key issues and discuss potential future directions for sentiment analysis, particularly in the fields of machine learning and natural language processing.

1.1 Levels of Sentiment analysis:

1. When it comes to sentiment analysis, we will compare and contrast the effectiveness of several machine learning classifiers, such as Support Vector Machines (SVM), Random Forests, Naive Bayes, and deep learning models. The goal of this performance comparison is to find the best classifiers for various sentiment analysis tasks.
2. Methods for Extracting Features: Two cornerstones of sentiment analysis are the selection and extraction of features. In order to find the most effective feature extraction methods, we will investigate many approaches, including bag-of-words, word embeddings, and representations based on deep learning.

Third, tune the hyperparameters: To get the most out of machine learning classifiers, you have to fine-tune them. Various hyperparameter optimization algorithms will be used, including grid search and random search, to determine the optimal parameter configurations for each classifier.

Practical Implications: Our work will provide practical recommendations for constructing and improving sentiment analysis systems in addition to focusing on theoretical issues. Our goal is to compile a set of guidelines and suggestions for use by academics and industry professionals.

Through the pursuit of these goals, our study aims to enhance sentiment analysis and provide light on how to improve machine learning classifiers for this crucial natural language processing (NLP) problem. Because of the critical importance of understanding sentiment for decision-making and strategy creation, we think that our results will be advantageous for sectors like social media monitoring, marketing, and customer feedback analysis.

1.2 Machine Leering techniques for sentiment analysis:

The social media platforms make their data easily and openly available online. Due to the abundance of publicly available data, sentiment analysis has recently attracted the attention of a new generation of academics. On the social media message boards, people share their thoughts and feelings. Companies use researchers to find out hidden information about their goods and services. The primary focus of multinational corporations is the automated and spontaneous determination of feelings from reviews. These days, automated data assessment is faster and more accurate thanks to machine learning algorithms. In this study, we looked at four different machine learning approaches that may be useful for sentiment analysis. Below, we will quickly go over the four methods' models.

1.3 Naïve Bayes used for sentiment classification

In most cases, the author's attitude toward his statement—whether positive or negative—determines the duality of mood. The Naïve Bayes classifier facilitates the expression of positive, negative, and neutral emotions in online text and is a widely used supervised classifier. A Naïve Bayes classifier sorts words into their correct categories by using conditional probability. Making use of Naïve Bayes for text classification has the advantage of requiring a minimal dataset for training. Before being converted into a collection of words, raw web data goes through preprocessing, which includes removing numeric data, foreign words, HTML elements, and special symbols. Words are tagged with positive, negative, and neutral labels by hand

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940
www.ijsem.org
Vol 15, Issuse.2 April 2021

by human specialists. For the training set, this preprocessing creates word-category pairings. Here we have a word 'y' from the test set, which is the unlabeled word set, and a document window containing n-words (x1, x2, …… xn). Given a data point 'y,' the conditional probability that it belongs to the training set's n-word category is:

$$P(y/x_1, x_2, \ldots\ldots x_n) = P(y) \times \prod_{i=1}^{n} \frac{P(x_i/y)}{P(x_1, x_2, \ldots\ldots x_n)}$$

Consider an example of a movie review for movie "Exposed". The experimentation with Naïve Bayes yields the following results.

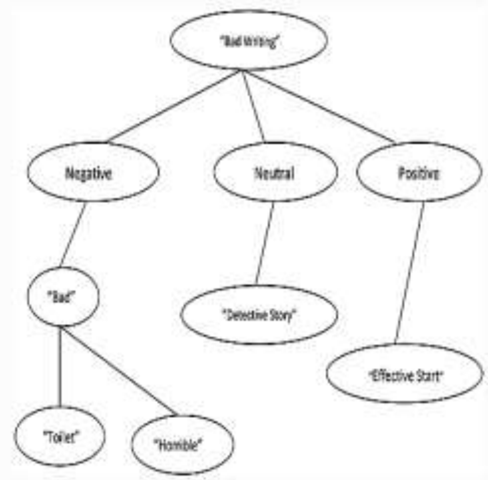
1.4. J48 algorithm used for sentiment prediction

The hierarchical mechanism divides feature space into distinct regions followed by the categorization of sample into category labels. J48 is a decision tree based classifier used to generate rules for the prediction of target terms. It has an ability to deal with larger training datasets than other classifiers. The word features for sentences of corpus taken from labeled arff file of training set are represented in the leaf nodes of decision tree. In the test set every time when a near feature qualifies the label condition of internal feature node, its level is lifted up in the same branch of decision tree. The assignment of labels to the word features of test set gradually generates different two branches of decision tree. J48 algorithm uses entropy function for testing the classification of terms from the test set.

$$Entropy\,(Term) = -\sum_{j=1}^{n} \frac{|Termj|}{|Term|} \log_2 \frac{|Termj|}{|Term|}$$

where (Term) can be unigram, bigram and trigram. In this study we have considered unigrams and bigrams. The example in the Table 2 contains bigrams like "Horrible acting", "Bad writing" and "Very misleading" are labeled with negative sentiment whereas the term "More enjoyable" reflects positive sentiment towards the movie. The decision tree of J48 algorithm for obtaining sentiment form text is represented in the Fig. 1 below.

Fig. 1



2. Literature Survey

To get a feel for where things stand, it's a good idea to peruse the literature on improving sentiment analysis using machine learning classifiers. The goal of sentiment analysis, often called opinion mining, is to extract the underlying emotional tone or sentiment from a given text using natural language processing techniques. The following is a general overview of a literature review on the subject:

Emotional evaluation: Provide a high-level overview of sentiment analysis, including its purpose, potential uses, and obstacles. An explanation of sentiment analysis's essential ideas and terminology is required.

Classifiers for sentiment analysis using machine learning: Outline the several machine learning classifiers that are often used in sentiment analysis. Ensemble techniques, Neural Networks (e.g., LSTM, CNN), Decision Trees, Random Forest, and Naive Bayes are all possible examples.

Discover the best practices for improving sentiment analysis with the use of machine learning classifiers by investigating a variety of approaches. The methodologies of text representation, data preparation, and feature engineering should be discussed. Emphasize methods for dealing with datasets that are unbalanced. Lay down the process of selecting a model and tweaking its hyperparameters.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN2454-9940
www.ijsem.org
Vol 15, Issuse.2 April 2021

3. Problem Statement

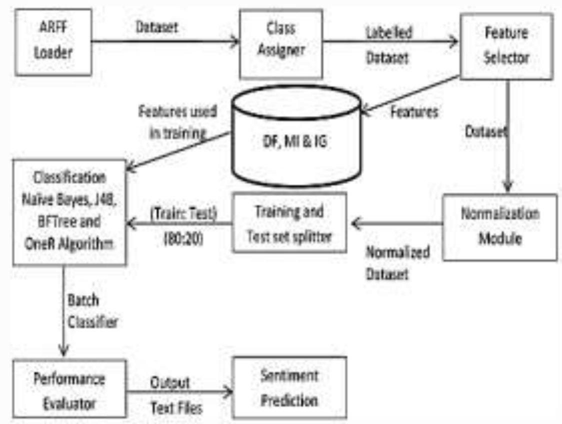
OneR algorithm is a classification approach which restricts decision tree to level one thereby generating one rule. One rule makes prediction on word feature terms with minimal error rate due to repetitive assessment of word occurrences. The classification of most frequent terms of a particular sentence is made on the basis of class of featured terms from training set. The demonstration of OneR algorithm for sentiment prediction with smallest error of classification is given below:

Step 1: Select a featured term from training set.
Step 2: Train a model using step 3 and step 4.
Step 3: For each prediction term.
 For each value of that predictor.
 Count frequency of each value of target term. Find most frequent class.
 Make a rule and assign that class to predictor.
Step 4: Calculate total error of rules of each predictor.
Step 5: Choose predictor with smallest error.

4. Proposed System

The preprocessing of raw text from web is done in python 3.5 using NLTK and bs4 libraries. Each review in the first dataset is parsed with NLTK's parser and title of the review is considered as a feature. We have obtained 15 features from first dataset and 42 features from each of second and third dataset. The CSV files generated from Python are converted to ARFF files for WEKA 3.8. Only two sentiment labels namely Pos for positive and Neg for negative are used for assigning sentences. The working methodology of proposed work for optimization of sentiment prediction is given below in Fig. 2.

Fig. 2



After loading files with ARFF loader, the class assigner picks up appropriate class labels from dataset and performs feature selection on the basis of frequently used headings and most frequent titles. The feature selector module is implemented using three feature selection methods namely Document Frequency (DF), Mutual Information (MI) and Information Gain (IG). The mathematical modeling of these feature selection methods requires some probability distributions and statistical notations described below:

P(w): Probability that a document 'd' contains term 'w'.

P(c'): Probability that document 'd' does not belongs to category 'c'.

P(w, c): Joint probability that document 'd' contains word term 'w' of category 'c'.

P(c/w): Conditional probability that a document 'd' belongs to category 'c' under the condition that 'd' contains word term 'w'.

Similarly other notations like P(w'), P(w/c), P(w/c'), P(c/w') and P(c'/w) are taken and {c} is the set of categories.

N_1: Number of documents that exhibit category 'c' and contain term 'w'.

N_2: Number of documents that do not belong to category 'c' but contains term 'w'.

N_3: Number of documents that belong to category 'c' and do not contain term 'w'.

N_4: Number of documents that neither belong to category 'c' nor contain term 'w'.

N: Total number of document reviews.

DF method qualifies only those documents in which a higher frequency terms are considered.

ISSN2454-9940
www.ijsem.org
Vol 15, Issuse.2 April 2021

$$DF = \sum_{i=1}^{m} N_{1i}$$

The MI method measures features of text by computing similarity of word terms ‘w’ and category ‘c’.

$$Sim_{Info}(w,c) = \log \frac{P(w/c)}{P(w)}$$

$$MI = \log \frac{N_1 \times N}{(N_1+N_3)(N_1+N_2)}$$

The IG-construct measures similarity information for category by exploiting probabilities of absence or presence of terms in a document review.

$$IG(w) = -\sum P(c) \cdot \log P(c) + P(w)\left[\sum P(c/w) \cdot \log P(c/w)\right] + P\left(w'\right)\left[\sum P\left(c/w'\right) \cdot \log P\left(c/w'\right)\right]$$

The normalization module converts all letters into lowercase, removal of punctuation marks and special symbols, conversion of numbers into words, expansion of abbreviation and limiting the average length of twenty words in a sentence. Each sentence is delimited by a newline character. The Python’s NLTK and bs4 libraries are used for this purpose. Data splitter take the ratio of (80:20) of (Train: Test) subsets. We have used manual splitting of dataset at the time of retrieval of data from web. The four classifiers are trained with training subsets followed by performance evaluation. The evaluation metrics taken in the experiment are precision, recall, accuracy and F-measure.

5. Result

The experiment is carried out by using freeware WEKA software tool for classification of sentiments in the text. Standard implementations of Naïve Bayes, J48, BFTree and OneR algorithms are exploited from WEKA version 3.8. The classification accuracy of first dataset shows 100% classification accuracy with Naïve Bayes in some of the epochs because of small size of dataset. The average of 29 epochs for all four classifiers on second and third datasets is presented in Table 4 below. Naïve Bayes shows faster learning among four classifiers whereas J48 found to be slower. OneR classifier is leading from other three classifiers in percentage of correctly classified instances. The accuracy of J48 algorithm is promising in true positive and false positive rates.

5. Conclusion

Using three datasets that have been annotated by hand, this article employs four ML classifiers to analyze sentiment. Table 4 reveals that OneR is more accurate in terms of percentage of properly identified cases, based on the mean of 29 experimental epochs. In contrast, Naïve Bayes shows a quicker learning rate, and J48 shows that both the true positive and false positive rates are adequate. Based on the results shown in Table 5, it is clear that J48 and OneR perform better with smaller datasets of Woodland's wallet evaluations. Extraction of foreign words, emoticons, and extended words together with their corresponding feelings is the extent of the preprocessing in the suggested technique. Future work on sentiment analysis might expand this research using convolutional neural networks and enhance preprocessing using word embeddings using deep neural networks.

References

1) A probabilistic generative model for mining cybercriminal networks from online social media: a review, by Parvathy G and Bindhu JS (2016). International Journal of Computer Applications 134(14):1-4. doi:10.5120/ijca2016908121.
2) Two, a survey of studies on natural language processing titled "Jumping NLP curves" was published in 2014 by Cambria and White. Publication: IEEE Computer Magazine, Volume 9, Issue 2, Pages 48–57. DOI: 10.1109/mci.2014.2307227.
3) Thirdly, in 2002, (Pang, Lee, and Vaithyanathan) Good job? Includes citation information from the ENNLP '02 conference proceedings (doi:10.3115/1118693.1118704).
4)
5) Fourthly, a study of the design framework, applications, and potential future scopes of sentiment analysis Link to the publication:

ISSN2454-9940
www.ijsem.org
Vol 15, Issuse.2 April 2021

Monali Bordoloi and Saroj Kumar Biswas, with the DOI: 10.1007/s10462-023-10442-2.

6)

7) 5) A comprehensive overview of ensemble-based methods for sentiment analysis in social networks. (Dimple Tiwari, Bharti Nagpal, Bhoopesh Singh Bhati, Ashutosh Mishra, and Manoj Kumar), April 20, 2023, 13407–13461, springer/10.1007/s10462-023-1047.