A Review of Feature Selection Techniques for Opinion Mining Systems

S. Jha¹*, S. Gupta², Y. Sharma³

¹,² IT Dept. BVCOE, GGSIPU, New Delhi, India
³ Corresponding Author: jhashankar0077@gmail.com, Tel.: +91-7210338609

Available online at: www.isroset.org

Abstract—Opinion Mining or Sentiment analysis is a subfield of Natural Language Processing that is used to categorize documents, reviews, social media posts etc into different categories of opinion based on sentiment expressed in the text. Large datasets often collected from the worldwide web using different sources like social media, blogs, articles, news reports are used by various organizations for sentiment analysis, Thus better-performing models in terms of accuracy and speed are required which can be achieved through feature reduction. For the research, we have used supervised machine learning algorithms, a learning model that uses mutual information, chi-square and genetic algorithm for feature selection techniques and ensemble model for classification. We have done a comparative study on the feature selection techniques like mutual information, chi-square and union of mutual information gain and chi-square. The comparative study has been done using different feature selection techniques, accuracy and results using these feature selection techniques with different supervised learning algorithms are then compared. We have then used genetic algorithms for feature selection and classification and compared its result with above-mentioned feature selection techniques.

Keywords—Opinion Mining, Feature Selection, Sentiment Analysis

I. INTRODUCTION

Opinion Mining allows us to understand the sentiments behind various documents and sentences and classify them accordingly. Many websites allow people to write their reviews about products and services received and sentiment analysis allows companies to understand those reviews and improves their products. Sentiment analysis is also important for social media monitoring to understand public opinion. Many corporations, political parties, celebrities and individuals remain concerned about the reaction of people to their action and words. Many corporations decide their strategy and future course of action based on the reaction of people about their product in the market. Opinion Mining is an application of text analytics to process and extracts people's reaction toward a topic by examining the text material [1-3]. With the volume of data generated each day opinion mining is being automated. Supervised and unsupervised algorithms such as SVM(support vector machine), maximum entropy, Naive Bayes, Random forest, decision tree algorithms, K-nearest neighbour, are used in combination with feature selection techniques to understand the polarity of the user opinion, emotion, and review [4–11] about a product or a subject. Opinion Mining takes as input some input text(sentence, document or review) and produces a score which helps us to classify the input text as positive or negative.

Many times, large data-sets are used for sentiment analysis that contains a lot of features. Many features out of these can be irrelevant that increase the model’s learning time and decrease accuracy. For such cases, feature selection plays an important role in selecting the relevant features and rejecting the ones that are irrelevant.

Feature selection is necessary while working with large data-sets with many features as it reduces the number of features by rejecting the irrelevant features. Feature selection helps in increasing the model’s accuracy by rejecting the misleading data and reduces the model’s learning time. It also reduces overfitting by removing extra noise in data. A feature vector with n features can have 2n possible feature subsets and a brute-force approach to try every possible subset would take too long in many situations so instead, we use feature selection techniques. We have compared different feature selection techniques for classifying movie reviews. We have compared the results of mutual information, chi-square, genetic algorithm and union of mutual Information gain and chi-square for various classification algorithms.

The remaining sections of the paper are as follows, an overview of the work done in the opinion mining domain is presented in section II. In section III, the tools and methods used are discussed. In section IV, the outcomes of the research are discussed. Further section V provides a summary of the results and a perspective of further development of this research.

II. RELATED WORK

Opinion mining has become one of the most important parts of the industry. Those industries which rely on
people's perception or say opinion to grow are required to know about the people's opinion about their product and company. Opinion mining is also used in other industries like entertainment to know about the people's opinion about a movie or music. Many attempts have been made to streamline the opinion classification. Ketan Sarvakar and Urvashi K Kuchara [1] used feature based classification of sentiment on movie review dataset [12] to determine the opinion about the movie. Pang et al. [2] used different machine learning algorithms along with feature selection techniques on movie review data-set [12]. They used the presence of words to create the feature vector and didn't use the frequency of words as later researchers used. They reported that best performance was reported by SVM with unigram features. The best accuracy achieved on the movie review dataset is 82.9%.

Wiebe et al. [6] focussed their research on an amalgamation of datasets such as banks, automobiles, travel destinations and movies review dataset. The authors used an approach that classified the individual words into two categories i.e positive words and negative words and then counted the occurrence of words in a sentence. The occurrence of each word in the sentence was used as a scoring parameter for the polarity of the sentence. If a sentence contains more negative words than the positive words then that sentence is classified as having negative polarity. This approach increased the efficiency of the classification but was unable to provide insight about the individual features that users like in a product.

Zhang et al. [7] use decision learning algorithms on customer feedback and product review dataset. Decision trees use a step based approach in classification. Different variables from the feature set are used at different levels as a deciding factor to move to the next level of the tree till the leaf of the decision tree is reached. The leaf of the decision tree acts as a classifier.

Ye et al. [10] used Naive Bayes, SVM, and other character based N-gram models for opinion mining of the reviews. According to them SVM and N-gram approach performed better than Naive Bayes.

III. METHODOLOGY

For the purpose of this research IMDB movie review dataset [12] is used that has fifty thousand reviews by different users. The dataset contains twenty-five thousand reviews for training and twenty-five thousand reviews for testing. We have combined them and used 90% of those to train and test and 10% for validation.

A. DATA PROCESSING

Data Set is prepared using:

- **Tokenization**: Through tokenization, the document was split into tokens like numbers, words, special characters etc. A sentence was split into individual words. This is also known as segmentation.
- **Normalization**: Through normalization, the word tokens in a document was converted into lowercase. Another approach could also have been followed to convert the words into uppercase. Doing this is important since the letter doesn’t affect the polarity of opinion.
- **Stop words removal**: Stop words rarely contribute to the overall classification. These are commonly occurring words in a sentence for example is, are, I etc. They are removed for the enhancement of the feature selection technique’s performance.
- **Stemming**: Through stemming the words are converted into their root forms for example good, better, best have same root word i.e good.

After cleaning the data, tokens are used to form a TF-IDF [13] vector which is then utilised for feature selection and then further for classification.

B. FEATURE SELECTION

1. **Information Gain**

Information Gain is used to measure the reduction in entropy or surprise which can be found by splitting the dataset and taking the value of a given random variable as reference. Events that have less probability contain more information as compared to events with less information [15].

Information gain can be used for feature selection by calculating the information gain of each variable in context to the output or target variable and this is known as mutual information.

Mutual information is used to determine the statistical dependence between two features. Mutual information quantifies the information that can be obtained about one variable through another random variable.

2. **Chi-square**

Pearson’s test or Chi-Square test can be used for testing the independence of two events. Chi-square is calculated between each variable(feature) and the target feature. Thus, we can select the features which are independent of each other i.e having highest chi-square value [16].

The formula of chi-square is as follows:

\[
X = \frac{(\text{OF} - \text{EF})}{\text{OF}}
\]

where,

\[
\text{OF} = \text{observed frequency in class}
\]

\[
\text{EF} = \text{expected frequency with no relations among features and target}
\]

3. **GENETIC ALGORITHM**

Genetic algorithms are heuristic function-based search algorithms that mimic the behaviour of the theory of natural selection of Charles Darwin. Genetic algorithms allow us to find the best solution from previous solutions. So initially we consider some variable in the feature set as the best variable to find the polarity of opinion and then iterate to find the best variables in the feature set to predict the polarity of the opinion.

So, genetic algorithms can be utilized to find the features that are important to determine the polarity of opinion from the original features.
The genetic algorithm works on an initial population [17] for producing individuals that perform better as compared to the previous population. For every generation, a new population is created by selecting individuals on the basis of their level of fitness. The mutation may also take place in the offspring. This process results in individuals that can perform better in the environment as compared to the original population i.e classify the polarity of opinion better as compared to initial variables selected for classification of opinion.

For feature selection with genetic algorithms first, we have an initial population of features and then we create a population with potential solutions which are originally initialized randomly and then during each generation the best solutions (in our case the solutions with best classification accuracy) are taken to the next generation. They reproduce and mutate to form a population with an overall better classification accuracy on average as compared to the previous population and this process carries on. Thus, the fittest solutions or features move to the next generation which then produces a better set of features or population. Eventually, we get a feature subset that performs the best on the data.

d. CHRONOLOGY OF SETUP

In our model we have used TF-IDF model, unigram, bigram and trigram features, binary and count indicators for URLs & hashtags, POS(Parts of Speech) Tagging, Chi-Square and Genetic Algorithm and ensemble classifier for opinion mining.

1. First, we extract features from the data-set and generate unigram, bigram and trigram features.
2. In our generated features we embed extra features for URLs, hashtags, count of words and characters.
3. Then we perform POS Tagging.
4. For feature selection, we use Chi-Square and mutual information gain
5. Then for classification, different techniques such as Logistic Regression, Naive Bayes and Random Forest have been used.

IV. RESULTS AND DISCUSSION

For our comparative study of feature selection techniques, we have used various techniques used in classification algorithms including Logistic Regression, Multinomial Naive Bayes, Naive Bayes and Random Forest. The results of our comparative study are:

<table>
<thead>
<tr>
<th>Feature Selection Technique</th>
<th>Logistic Regression</th>
<th>Multinomial Naive Bayes</th>
<th>Naive Bayes</th>
<th>Random Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>88</td>
<td>84</td>
<td>81.5</td>
<td>84.5</td>
</tr>
<tr>
<td>Information Gain</td>
<td>87</td>
<td>84.5</td>
<td>81.5</td>
<td>85.5</td>
</tr>
<tr>
<td>Chi-Square</td>
<td>88.5</td>
<td>84</td>
<td>82</td>
<td>86</td>
</tr>
<tr>
<td>Genetic Algorithm</td>
<td>88</td>
<td>83.5</td>
<td>81.5</td>
<td>85</td>
</tr>
</tbody>
</table>

This table depicts the accuracy results that we got during our research for the various feature selection techniques and classification algorithms. From our findings on research models trained on various techniques, we found many variations in our findings. The point to be noted is that the best accuracy is achieved as 88.5% with Chi-Square feature selection technique with logistic regression.

Logistic regression has the second-highest accuracy of 88% with genetic algorithm. While Naive Bayes performed worst in all the cases of feature selection techniques. While feature selection using genetic algorithms doesn’t give the highest accuracy but that was with a limited population. Due to computational bottlenecks, we could not test it with a higher population number. Logistic Regression along with Chi-square feature selection technique can be used in industrial purposes as they are having high accuracy, but they can’t be used in systems in which there is an issue of having a surety of algorithm because the algorithm is not as reliable as humans and will need a human verification.

Figure 1. Confusion matrix for Logistic Regression classifier with chi-square feature selection technique

We detected the negative and positive movie reviews with very high accuracy.

Figure 2. ROC Curve for Logistic Regression classifier with chi-square feature selection technique.
Ideally, we have an area of one unit under the ROC curve which represents 100% accuracy for the model. Both points having true negative and false positive play an important role in this. More the points like these in our graph less will be the area under the curve and less accuracy. As in our model, the area under the curve is calculated as 0.95 which means it is a little deviated from the ideal area of one unit. This is due to the false positive and true negative data points present in our dataset. In this way, the ROC curve is analysed and interpreted for our model.

V. CONCLUSION AND FUTURE SCOPE

As proposed, we have used different feature selection techniques to create the input vector for the model to perform opinion classification such as chi-square, mutual information gain, genetic algorithm and union of any two or more of them. These feature selection techniques are used to select the most relevant feature and then classifiers such as Logistic Regression, Random Forest, Naïve Bayes and Multinominal Naïve Bayes. The best accuracy was for chi-square feature selection technique but the genetic algorithm performance was satisfactory with a very small population and generation parameter. Further feature selection also reduced the time required to train the classifier and inference the result while testing the classifier. Better embedding vectors like word2vec can be used in further studies. While input vectors prepared using feature selection techniques given satisfactory results, the same cannot be deployed in industrial production for real-time classification of opinion. These feature vectors deal with the word in isolation i.e. a single word might dominate the classification of opinion without taking into account the context in which the word was used. To improve the accuracy of the opinion mining system the context of words can be used to classify the opinion. This can be achieved by the use of deep learning techniques like LSTM, GRU etc. LSTM and GRU take the context of the entire sentence into account and then classify opinion based on the complete sentence.

REFERENCES


AUTHORS PROFILE

Mr. Shankar Jha is currently pursuing B.Tech in Information technology from Bharati Vidyapeeth’s college of engineering. He is in his final year of graduation. He has worked in the field of artificial intelligence. He works extensively in the field of computer science discipline like cloud computing, Networking and communication, Architecture and organization. He is currently working as a software developer engineer and has a good knowledge of development tools and technologies.
Mr. Sarthak Gupta is pursuing B.Tech in Information Technology from Bharati Vidyapeeth’s College of Engineering. He is in the final year of his graduation and has worked on various research projects besides his academic projects focusing mainly on Artificial Intelligence and Machine Learning. He is currently working as a Software Development Engineer. He has experience of working with different tools and technologies used in the industry and for research purposes as well.

Mr. Yagya Sharma is currently pursuing B.Tech in Information Technology from Bharati Vidyapeeth’s College of Engineering. He is in the final year of his graduation. He is currently working as a Data Analyst. His research work mainly focuses on Artificial Intelligence, Machine Learning and Big Data Analytics. He has experience of working with different tools and technologies used in the industry and for research purposes as well.