ISSN: 2581-3404 (Online) IF: 5.445 (SJIF)
IJIRTM, Volume-7, Issue-4, July-2023.

An Approach Based on Machine Learning to Analyze the Movie Reviews Using the IMDB Dataset

Shashi Ranjan[1], Chetan Agrawal[2], Darshna Rai[3]
CSE Department, Radharaman Institute of Technology and Science, Bhopal, India[1, 2, 3]
wellshashi@gmail.com[1], Chetan.agrawal12@gmail.coms[2], darshna.rai@gmail.com[3]

Abstract: *The term "sentiment analysis" refers to one of the many specialized subfields that fall under the umbrella of "opinion mining." Textual information can be structured, semi-structured, or unstructured depending on the type of research being conducted in this discipline of social psychology. The goal of this research is to determine how members of the general public feel and think about a particular subject. This thesis contains a study that utilizes sentiment analysis, and the primary emphasis of that study is the database of IMDB movie reviews. In the work that we completed for our thesis, we offer a novel approach to an improved version of the Naive Bayes algorithm. This method was developed by us. TF-IDF is utilized in order to attain this goal successfully. The comparison is done out using datasets of specific sizes, and it is carried out on the basis of certain parameters, including mean square error, accuracy, precision, recall, and F1 score. The comparison between our work and other categorization algorithms has been completed, and our results have been proved to have a higher level of accuracy.*

Keywords: Movie Review, Sentiment Analysis, TF-IDF, IMDB.

Introduction

The present era of Internet has become a huge Cyber Database which hosts gigantic amount of data which is created and consumed by the users. The database has been growing at an exponential rate giving rise to a new industry filled with it, in which users express their opinions across channels such as Facebook, Twitter, Rotten Tomatoes and Foursquare. Opinions which are being expressed in the form of reviews provide an opportunity for new explorations to find collective likes and dislikes of cyber community. One such domain of reviews is the domain of movie reviews which affects everyone from audience, film critics to the production company. The movie reviews being posted on the websites are not formal reviews but are rather very informal and are unstructured form of grammar. Opinions expressed in movie reviews give a very true reflection of the emotion that is being conveyed. The presence of such a great use of sentiment words to express the review inspired us to devise an approach to classify the polarity of the movie using these sentiment words.

Sentiment Analysis is a technology that will be very important in the next few years. With opinion mining, we can distinguish poor content from high quality content. With the technologies available we can know if a movie has more good opinions than bad opinions and find the reasons why those opinions are positive or negative. Much of the early research in this field was centered around product reviews, such as reviews on different products on Amazon.com [1], defining sentiments as positive, negative, or neutral. Most sentiment analysis studies are now focused on social media sources such as IMDB, Twitter [2] and Facebook, requiring the approaches is tailored to serve the rising demand of opinions in the form of text. Furthermore, performing the phrase-level analysis of movie reviews proves to be a challenging task.

Sentiment analysis is language processing task that uses a computational approach to identify opinionated content and categorize it as positive or negative. The unstructured textual data on the Web often carries expression of opinions of users. Sentiment analysis tries to identify the expressions of opinion and mood of writers. A simple sentiment analysis algorithm attempts to classify a document as 'positive' or 'negative', based on the opinion expressed in it. The document-level sentiment analysis problem is essentially as follows: Given a set of documents D, a sentiment analysis algorithm classifies each document d □ D into one of the two classes, positive and negative. Positive label denotes that the document d expresses a positive opinion and negative label means that d expresses a negative opinion of the user. More sophisticated algorithms try to identify the sentiment at sentence-level, feature-level or entity-level.

There are broadly three types of approaches for sentiment classification of texts: (a) using a machine learning based text classifier -such as Naïve Bayes, SVM or kNN- with suitable feature selection scheme; (b) using the unsupervised semantic orientation scheme of extracting relevant n-grams of the text and then labeling them either as positive or negative and consequentially the document; and (c) using the SentiWordNet based publicly available library that provides positive, negative and neutral scores for words. Some of the relevant past works on sentiment classification can be found in [3], [4].

The new user-centric Web hosts a large volume of data created by various users. Users are now co-creators of web content, rather than being passive consumers. The social media is now a major part of the Web. The statistics shows that every four out of five users on the Internet use some form of social media. The user contributions to social media range from blog posts, tweets, reviews and photo/ video uploads etc. A large amount of the data on the Web is unstructured text. Opinions expressed in social media in form of reviews or posts constitute an important and interesting area worth exploration and exploitation. With increase in accessibility of opinion resource such as movie reviews, product reviews, blog reviews, social network tweets, and the new challenging task is to mine large volume of texts and devise suitable algorithms to understand the opinion of others. This information is of immense potential to companies which try to know the feedback about their products or services. This feedback helps them in taking informed decisions. In addition to be useful for companies, the reviews and opinion mined from them, is helpful for users as well. For example, reviews about hotels in a city may help a user visiting that city locating a good hotel. Similarly, movie reviews help other users in deciding whether the movie is worth watch or not.

Rest of the paper is organized as follows: section II explains background of sentiment analysis, in section III literature review of previous research in the field of movies review sing machine learning is analyzed, in section IV we explained our Proposed method with algorithm and flow graph, in section V experimental parameters and results are discussed, in last we conclude our work in section VI followed by references used in this work.

II. Background

Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be their judgment or evaluation, affective state which is the emotional state while writing, or the intended emotional communication which is the emotional effect the writer wishes to have on the reader. Sentiment is defined as any kind of emotion and in the context of sentiment analysis is the opinion that is being expressed in the form of text or speech.

A. Feature Selection

Most researchers apply standard feature selection in their approach to improve computational performance with a handful using more sophisticated approaches. Papers focusing entirely on feature selection to improve sentiment analysis are few. Among them, the famous one was Pang & Lee [5], who removed objective sentences on a testbed consisting of objective and subjective text trained on SVM. Initially they found that sentiment classification result actually abated. They then concluded it was more likely that sentences adjacent to discarded sentences improved classification result over their baseline. Another work used sophisticated feature selection and found that using either information gain (IG) or genetic algorithm (GA) results in an improvement in accuracy.

B. SentiWordNet

SentiWordNet [6] (SWN) is an extension of WordNet [7] that was developed by Esuli & Se- bastiani, which augments the information in WordNet with sentiment of the words in them. Each synset in SWN comprises of sentiment scores that are positive and negative score along with an objectivity score. The summation of these three scores gives the relative strength of positivity, negativity and objectivity of each synset. These values have been obtained by using many semi-supervised ternary classifiers, with the capability of determining whether a word was positive, negative, or objective. If all the classifiers settled on a result then the highest value are assigned for the analogous score, else the values for the positive, negative and objective scores were proportional to the number of classifiers that assigned the word to each class.

III. Related Work

A large number of works have been carried out previously on opinion mining and sentiment analysis.

Nagamma P et al. [8] proposed different data mining techniques for classification of movie reviews and it also predicts the box office collection for the movie. Classification accuracy for pretending was improved substantially by clustering method. The online movie review data collected from IMDB dataset, the box office collection and the success or failure of the movie is predicted based on the reviews. Pang et al. [5] applied the machine learning technique for classification of reviews present on IMDB movie reviews database, by forming the list of 14 keywords which are useful in finding the baseline for classification accuracy. The machine learning techniques like Naïve bayes, SVM, achieves higher accuracy over the baseline. J. Erman et al. [9] studied three types of clustering algorithms namely K-Means, DBSCAN and AutoClass algorithm for the classification of network traffic problem. This study is based on the ability of each algorithm for forming clusters having higher predictive power of a single traffic class and for determining the ability of each algorithm to generate small number of clusters that has many connections. The AutoClass algorithm is compared with DBSCAN and K-Means algorithm and the result indicates that both K-Means and DBSCAN work faster than AutoClass algorithm. Turney et al. [10] studied the unsupervised learning algorithm for sentiment classification process. They determined the similarity of words with help of NEAR operation and developed a classifier for finding polarity result.

Stefano, Andrea and Fabrizio [6] present SentiWordNet 3.0, it is a lexical resource developed for sentiment classification. SentiWordNet 3.0 is an open resource platform for all researchers all over the world, for different types of research projects it has supported more than 300 research groups worldwide. Rudy Prabowo et al. [11] studied the hybrid SVM classification method for sentiment classification. They used Sentiment Analysis Tool for achieving good level of effectiveness. Rui Yao et al. [12] proposed a simplified version of sentiment aware autoregressive model this model can be used for producing the good accuracy for prediction of box office sale revenue using online movie review data. NB classifier is used for the sentiment classification. Mullen and Collier et al. [13] proposed an approach for classification of text data into positive or negative polarity using SVM. Their work involved extraction of value phrases (two word phrases conforming to a particular part-of-speech) and assigning them sentiment orientation values by using point wise mutual information.

Godbole, Manjunath & Stevens in their work [14] present a system that quantifies positive or negative opinion to each distinct entity in the text corpus. Their system consists of two phases, a sentiment recognition phase where opinion expressing entities are determined and a scoring phase where a relative score for each entity is determined. In the work by Annett & Kondark [15] it was observed that ML technique of sentimental classification on movie reviews is quite successful and it was also observed that the type of features that are chosen have a dramatic impact on accuracy of the classifier. As there is an upper bound on the accuracy level that a dictionary based approach has as demonstrated in lexical approach.

However, the three machine learning methods they employed (Naive Bayes, maximum entropy classification, and support vector machines) do not give as good results on sentiment classification as on traditional topic-based categorization. They further extracting these portions [16] and implementing efficient techniques for finding minimum cuts in graphs; this greatly favors indulgence of cross-sentence contextual constraints, which provides an efficient means for integrating inter sentence level contextual information with traditional dictionary of words features.

Singh et al. [17] presents experimental work on performance evaluation of the SentiWordNet approach for document-level sentiment classification of Movie reviews and Blog posts. They did variations in semantic features, scoring schemes and thresholds of SentiWordNet approach along with two popular machine learning approaches: Naive Bayes and SVM for sentiment classification. The comparative performance of the approaches for both movie reviews and blog posts is illustrated through standard performance evaluation metrics of Accuracy, F-measure and Entropy.

IV. Proposed Work

One of the challenging tasks in machine learning is sentiment analysis. Because people are writing their reviews on different areas and we have to predict their emotions or reviews which are written in their own language as positive or negative. So in our work also we have done the sentiment analysis. But our work is based on improved Naïve Bayes classifier for classifying the movie review sentiments. The improvement in Naïve Bayes is done with the help of Tf-IDF and the movie dataset is taken from IMDB. Now we will discuss about the data set and Tf-IDF.

A. Dataset- The reviews of movies are taken from IMDB websites. We will perform the classification on dataset sizes of 2300 and 3400 reviews.
TF-IDF (Term Frequency-Inverse Document Frequency) - It is a text mining technique used to categorize documents. It stands for term frequency * inverse document frequency and is a method for emphasizing words that occur frequently in a given document, while at the same time de-emphasizing words that occur frequently in many documents.
In order to compute the TF-IDF more efficiently, we performed the following pre-processing methods on the dataset:

- We removed most frequent words from documents. In fact, we discarded words appearing in more than 80% of the documents because they don't give information and can bias the result.
- We have also removed rare words: discard words appearing in less than 5 documents. It can cause some kind of noise.
-

B. Algorithm of Proposed Work

In this section we will discuss the proposed work

```
1.	Import all the necessary libraries like pandas numpy, matplolib
2.	Now load the file from IMDB dataset, let it be
moviedir = imdbmovie_reviews'
3.	Import the libraries like Count Vectorizer and NLTK(Natural Language Toolkit)
Feature Extraction= Count Vectorizer
4.	Tokenize each word, let
movie_vec = CountVectorizer(min_df=2, tokenizer=nltk.word_tokenize)
movie_counts = movie_vec.fit_transform(movie_train.data)
5.	Call function Tf-IDF from from sklearn.feature_extraction.text
6.	movie_tfidf = tfidf_transformer.fit_transform(movie_counts)
7.	Import Naïve Bayes, Call function train_test_split() and Naïve Bayes
// Splitting the data into training and testing data
docs_train, docs_test, y_train, y_test = train_test_split( movie_tfidf, movie_train.target, test_size = 0.20, random_state = 12)
clf = MultinomialNB().fit(docs_train, y_train)
8.	Predicting the dataset
y_pred = clf.predict(docs_test)
9.	Computing accuracy, mean square error, precision, recall and F1 score
Accuracy=accuracy_score(y_test, y_pred)
mse1 = ((y_pred - y_test) ** 2).mean()
Precision= precision_score(y_test,y_pred)
Recall= recall_score(y_test,y_pred)
F1_score= f1_score(y_test,y_pred)
```

In this work we will use python 3.6, which is one of the best Tools available today for sentiment analysis purpose.

C. Data Source

In this phase, we contribute the perfunctory details about the datasets that are used in our implementation. We extricate the latest reviews of movies from IMDB websites. We have performed demonstration on above declare corpus and our own datasets. Table1 shows the dataset details that are used in proposed work.

Table 1: IMDB dataset details

Datasets	*No. of Reviews*
IMDb movie reviews *(http://www.imdb.com/)*	*2300*
IMDb movie reviews *(http://www.imdb.com/)*	*3400*

D. Data Preprocessing

For the pre-processing of the dataset, it will be applied some data filtering techniques to make that raw data into structured format. Pre-processing involves several steps such as tokenization, Removal of stop words and Case Normalization.

1) Tokenization: Tokenization is a process in which text of a document is cleaved into series of token. The data that is extracted from online reviews hold noise such as symbols, scripts etc which is not necessary and not used for machine learning. In order to retain only text, these noises are to be removed so as to improve the performance of the classifier.

2) Removal of stop words: It is a process to minimize the size of document by eliminating the most usual words according to stop word list. Stop words are those words that are not compulsory for the sentences or opinions. Stop word list consist of preposition and determiners. For example: "Rahul is a good boy" will be processed to "Rahul is good boy".

3) Case Normalization: Most of the reviews are in the compound form that is uppercase and lowercase and it needs to convert whole document into uppercase or lowercase. Case Normalization is the process in which all the characters in a document convert either in the uppercase or lowercase.

E. Feature Selection

Feature selection is a process that performs the selection of features in the data, out of which majority data is related to the current predictive modeling problem. It is a process of choosing a reduced relevant features that improves classification by searching for the best feature subset, from the fixed set of original features according to a given classification accuracy [11]. It removes irrelevant or redundant features. Feature selection is also known as attribute selection, variable selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. In this work, we are used Gini Index method for feature selection. The Gini Index of node impurity is the measure most common chosen for classification-type problem. If a dataset T contains examples from n classes, Gini Index, Gini (T) is defined as: Σ (1)

F. Classification

Classification is a technique in data mining that assigns items in a set to target classes or categories. The purpose of classification is to get the forecast of the target class for every case in the data. The algorithm will try to find out relationships between the attributes/variables that will ensure it is possible to forecast the outcome. In the part of classification, SVM and RF algorithm is used for classify movie reviews that are collected from internet. In this work algorithms are used for identification and compared for the detailed evaluate the results. Classification is done by SVM and RF. Support Vector Machines are supervised learning models that are used hyperplanes or set of hyperplanes for separation of classes by evaluating maximum margin from both classes. A hyperplane is represented by the following equation [6]: On the other hand, Random Forest algorithm works as a large collection of de-correlated decision trees. Random Forest is applicable to high-dimensional data with a low number of observations.

G. Performance Measure

After completing the classification of features, performance is measured by using various parameters such as Accuracy, Root Mean Square Error, Precision, Recall and F-Measure.

H. Flow Graph

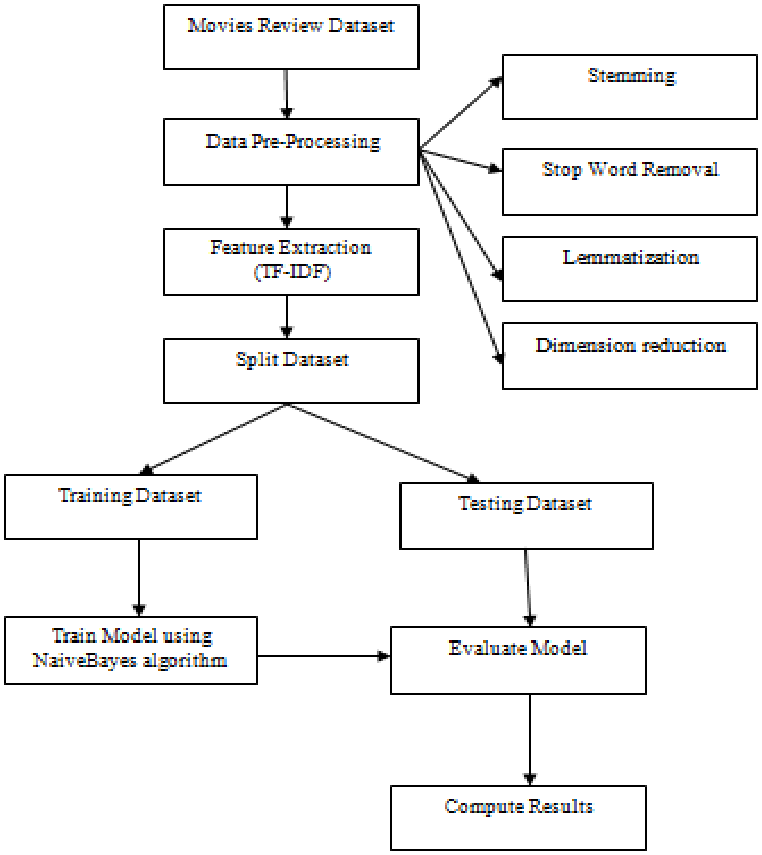

Fig.1: flow graph of the proposed method.

V. Result Analysis

Our proposed model is implemented on movie reviews datasets. There are two datasets. The first and second dataset consider 2300 and 3400 reviews of latest movies that are collected from IMDB web site. These both datasets are having both positive and negative reviews that are delivered by viewers on websites. After applying string to vector filtration for preprocessing of data, feature selection methods are used to select the most relevant features. Random Forest and Support Vector Machine (Linear) algorithms are applying for classification of movie reviews. We are using six parameters to comparing the results that is Accuracy, Root Mean Square Error, Precision, Recall, and F-Measure.

A. Evaluation Metrics

1) Accuracy : Accuracy is the performance evaluation parameters in which the true outcomes such as true positive and true negative among the total cases are examined such as true positive, true negative, false positive and false negative.

2) Root Mean Square Error: The square root of the arithmetic mean of the squares of a set of the values. The Root Mean Square Error (RMSE) (also called root mean square deviation, RMSD) is the frequently used measure of the difference between values predicted by the model (y) and the values actually observed from the environment (yi).

3) Precision: Precision is defined as the fraction of documents that are retrieved and that are relevant from that retrieved documents to the Query.

4) Recall: Recall is defined as the division of the documents that are matches to the query that are retrieved successfully.

5) F-Measure: The F1 score (also F-score or F-measure), in statistical analysis of binary classification is a measure of a test's accuracy. It computes the score considering both precision p and recall r.

B. Result of Algorithms for Movies Reviews Datasets

1) Accuracy: For Accuracy, table2 shows that overall result and performance are best proved for different number of movie reviews with Random Forest algorithm with Gini Index based feature selection than other techniques with 63.6364 (in case of 1100 reviews), 71.6087(in case of 2300 movie reviews), and on case of proposed method it is 64.42 & 78.47 respectively This shows that the proposed model is performing better in terms of Accuracy and all other measures.

Table 2: Comparing various algorithms by **Accuracy** for different Movie Reviews Dataset

ALGORITHMS	***2300 Reviews***	***3400 Reviews***
GI+SVM	*69.4*	*71.2*
IG+RF	*71.1*	*76.0*
GI+RF	*71.6*	*78.0*
Proposed	*78.4*	***86.1***

Table 3: Comparing various algorithms by **RMSE** for different Movie Reviews Dataset

ALGORITHMS	***2300 Reviews***	***3400 Reviews***
GI+SVM	*55.3*	*53.5*
IG+RF	*43.2*	*39.6*
GI+RF	*42.7*	*39.0*
Proposed	*42.3*	***37.18***

Table 4: Comparing various algorithms by **Precision** for different Movie Reviews Dataset

ALGORITHMS	***2300 Reviews***	***3400 Reviews***
GI+SVM	*69.4*	*71.3*
IG+RF	*71.9*	*76.4*
GI+RF	*72.1*	*78.3*
Proposed	*80.9*	***86.7***

Table 5: Comparing various algorithms by **F-Measure** for different Movie Reviews Datasets

ALGORITHMS	*2300 Reviews*	*3400 Reviews*
GI+SVM	*69.4*	*71.3*
IG+RF	*70.9*	*76.0*
GI+RF	*71.5*	*78.0*
Proposed	*76.4*	***85.6***

VI. Conclusion

Various kinds of information are produced by some social media organizations, and these businesses have to be prepared to notice people's perspectives on various things, including products, objects, and movie reviews, among other things. It is common knowledge that the attributes that contain a huge number of values are the source of the issue with information gathering. In this work, we identified new variables that have a major impact on identifying the polarity of the movie reviews and employed computation linguistic approaches for the preprocessing of the data. Additionally, we used the data to create a model that predicts the likelihood of a reviewer giving a positive or negative rating for a movie. In the study that we have proposed, not only have we demonstrated that increasing the size of the dataset will also enhance the accuracy and other metrics, but we have also demonstrated that the accuracy can be improved by utilizing Naive Bayes in conjunction with Tf-IDF. In the future, one of our goals is to determine whether or whether the features and methods that have been proposed for sentiment classification are beneficial when applied to other tasks, such as sentiment classification. In order to make a more accurate prediction of the polarity of the document, we would like to apply more advanced principles from NLP. We would also like to apply this method to more areas of opinion mining, such as newspaper articles, product evaluations, political discussion forums, and so on.

References

[1.] Gregory, Michelle L., Nancy Chinchor, Paul Whitney, Richard Carter, Elizabeth Hetzler, and Alan Turner. "User-directed sentiment analysis: Visualizing the affective content of documents." In Proceedings of the Workshop on Sentiment and Subjectivity in Text, pp. 23-30. Association for Computational Linguistics, 2006.

[2.] Pak, Alexander, and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining." In LREc, vol. 10, no. 2010, pp. 1320-1326. 2010.

[3.] Singh, V. K., R. Piryani, Ahsan Uddin, and P. Waila. "Sentiment analysis of Movie reviews and Blog posts." In 2013 3rd IEEE International Advance Computing Conference (IACC), pp. 893-898. IEEE, 2013.

[4.] Durant, Kathleen T., and Michael D. Smith. "Mining sentiment classification from political web logs." In Proceedings of Workshop on Web Mining and Web Usage Analysis of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (WebKDD-2006), Philadelphia, PA. 2006.

[5.] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pp. 79-86. Association for Computational Linguistics, 2002.

[6.] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." In Lrec, vol. 10, no. 2010, pp. 2200-2204. 2010.

[7.] Adreevskaia, Alina, and Sabine Bergler. "Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses." In 11th conference of the European chapter of the Association for Computational Linguistics. 2006.

[8.] Nagamma, P., H. R. Pruthvi, K. K. Nisha, and N. H. Shwetha. "An improved sentiment analysis of online movie reviews based on clustering for box-office prediction." In Computing, Communication & Automation (ICCCA), 2015 International Conference on, pp. 933-937. IEEE, 2015.

[9.] Erman, Jeffrey, Martin Arlitt, and Anirban Mahanti. "Traffic classification using clustering algorithms." In Proceedings of the 2006 SIGCOMM workshop on Mining network data, pp. 281-286. ACM, 2006.

[10.] Turney, Peter D., and Michael L. Littman. "Unsupervised learning of semantic orientation from a hundred-billion-word corpus." arXiv preprint cs/0212012 (2002).

[11.] Prabowo, Rudy, and Mike Thelwall. "Sentiment analysis: A combined approach." Journal of Informetrics 3, no. 2 (2009): 143-157.

[12.] Yao, Rui, and Jianhua Chen. "Predicting movie sales revenue using online reviews." In Granular Computing (GrC), 2013 IEEE International Conference on, pp. 396-401. IEEE, 2013.

[13.] Mullen, Tony, and Nigel Collier. "Sentiment analysis using support vector machines with diverse information sources." In Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.

[14.] Godbole, Namrata, Manja Srinivasaiah, and Steven Skiena. "Large-Scale Sentiment Analysis for News and Blogs." Icwsm7, no. 21 (2007): 219-222.

[15.] Annett, Michelle, and Grzegorz Kondrak. "A comparison of sentiment analysis techniques: Polarizing movie blogs." In Conference of the Canadian Society for Computational Studies of Intelligence, pp. 25-35. Springer, Berlin, Heidelberg, 2008.

[16.] Pang, Bo, and Lillian Lee. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." In Proceedings of the 42nd annual meeting on Association for Computational Linguistics, p. 271. Association for Computational Linguistics, 2004.

[17.] Singh, Vivek Kumar, Rajesh Piryani, Ashraf Uddin, and Pranav Waila. "Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification." In Automation, computing, communication, control and compressed sensing (iMac4s), 2013 international multi-conference on, pp. 712-717. IEEE, 2013.