International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

SENTIMENTAL ANALYSIS OF IMDB USER REVIEWS

1. Mohit Khodse, 2. Rohit Kumar, 3. Hrishikesh Raj.

D. Y. Patil College of Engineering, Pune

Abstract— The sentiment analysis is an emerging analyzed, to generate useful insights in regards to a specific topic. It is an effective tool which can serve government ,corporations and even consumers. In this project, machine learning-based approaches are applied to accurately determine the sentiment behind the IMDb user reviews to evaluate whether user has positive remarks or negative remarks regarding the movie. Multiple Machine learning technique i.e. Logistic Regression, Decision Tree, Random Forest, Naive Bayes and Support Vector Machines are employed to assess the polarity of each sentiment present in the dataset.

Keywords: Sentiment Analysis, Machine Learning (ML), Natural Language Processing (NLP), Movie Reviews

INTRODUCTION

The sentiment analysis is the process of using natural language processing and computational linguistics to extract, identify and categorize different opinions expressed in text format. The sentiment analysis is the process of using natural language processing and computational linguistics to extract, identify and categorize different opinions expressed in text format. Sentiment analysis has been investigated on levels ranging from document level, where the polarity of a sentiment of the whole document can be determined; to sentence level sentiment analysis, where polarity is determined sentence-wise; going deeper towards phrase-level sentiment analysis and finally aspect level sentiment analysis, where sentiments are selected on the basis of the aspects.

MOTIVATION

Significance and importance of sentiment analysis are well-understood since it helps businesses understand their consumer's sentiment towards their brand or product. Having this done automatically, stakeholders can use them to make well-informed decisions. Sentiment analysis may be redefined as follows- it refers to the methods and strategies that enable firms to examine data about how their customer base feels about a given service or product. It is a process that automatically analyzes natural language occurrences, mines essential aims or views and classifies them according to emotional attitude. It has applications for a wide range of problems .It tackles cl business demands as mentioned above. Customer satisfaction analysis is also possible via sentiment analysis .Then the analysis can identify and help stakeholders act in real-time problems.

PROBLEM DEFINITION

The task of sentiment analysis is very crucial and needs detailed attention. As discussed earlier, emoticons play a major role in deriving sentiments from one's review. Their presence or absence can have a major impact on model evaluation parameters and hence are important for consideration. So the formulation of the problem statement is as follows - What is the effect of emoticons on sentiment analysis? This project utilizes various Machine Learning algorithms mentioned above to examine and compare the evaluation parameters and answer the problem statement.

RELATED WORK

Data Collection-

Data collection is an important step in the task of sentiment analysis as it will depend upon the collected data to determine how to perform sentiment analysis or in other words, which approach is to be performed. Data can be gathered from a wide range of options from the Internet like web scraping, social media, news channels, E-commerce websites, Forums, blogs, Weblogs, etc.

2.Feature Extraction-

The problem accessed in this project is basically a binary classification problem i.e. all input reviews are needed to be correctly classified into two classes either positive or negative. Developing a classification model requires the identification of relevant features in the dataset such that a reiew can be partitioned into base words during the training of the model and then appended into the feature vector. The basic technique involves the usage of unigrams, bigrams or trigrams.

Methodology

As discussed in the above sections, there are mainly three approaches employed for the task of Sentiment Analysis namely Lexicon Based Approach, Machine Based Approach and Hybrid Approach. Each approach is derived on the basis of the kind of data at its disposal. If the data is structured, Machine learning techniques are suitable for performance. If the data is unstructured, Lexicon-based approaches are suitable for employment since their performance is solely dependent upon the predefined value associated with the respective lexicon. If the data is structured, we employ machine learning based or deep learning based approaches to perform the task of sentiment analysis and this requires the least human effort in nature hence is suitable for automated sentiment analysis.

If the data is semi-structured, hybrid approaches come to play. Hybrid approches involves combination of lexicon and machine learning approaches since both kind of data need to be targeted for this type of data.

Data Processing

After selecting the data for the task of sentiment analysis, some preprocessing of data is required to make it model-ready with target of achieving maximum accuracy. Data after data selection is in raw form i.e. natural form. Data needs to go through various steps to be model-ready. Accuracy of a model depends directly on this step. All the feature extraction techniques mentioned above are employed on this step.

This project utilizes tf-idf vectorizer as a feature. TF-IDF vectorization involves calculating the TF-IDF score for every word in the corpus relative to that document (in this case, reviews) and then putting that information into a vector. This helps in transformation of the text into sensible representation of integers which may be used to fit machine learning algorithm for predictions. The term frequency is the number of occurrences of a specific term in a document. Term frequency indicates how important a specific term in a document. Document frequency is the number of documents containing a specific term signifying how common the term is.

Inverse document frequency (IDF) is the weight of a term, it aims to reduce the weight of a term if the term's occurrences are scattered throughout all the documents. It can be calculated using

$$Idfi=log(n/dfi)$$

Finally ,tf-idf score can be calculated using

$$Wi,j = tfi,j * idfi$$

Where wi,j is tf-idf score for term i in review j, former term on R.H.S. is term frequency and latter term is inverse document frequency.

Standard preprocessing steps are employed further to convert raw data into meaningful and processible format. This is achieved in steps mentioned below. First step is done by removing all the HTML tags and markups from the reviews. Second step has two versions, one involves removal of punctuation marks without saving emoticons for later appending and other involves saving emoticons for later appending. This step is crucial for differentiating the affect of emoticons on sentimental analysis and differences on evaluation parameters is done on the basis of emoticon appending. Next step involves converting all letters to lowercase for text normalization, with and without saved emoticons appended. Then for each word in a review, tokenization and stemming is performed. Stemming is removal of suffixes and prefixes from a word to bring it into its root form, while tokenization is breaking down of sentences into words. This is the final step of data preprocessing and now the data is ready to be employed into the models .

Finally we split our preprocessed data into training and testing sets keeping ratio of 7:3 with stop words removed paralelly. This is achieved by using a function train test split from the sklearn module for python. Training data is the data with which the models are trained. Testing data is the data on which testing is performed to evaluate all the parameters of evaluation of a model.

Logistic Regression

Logistic Regression is an example of supervised learning. It is used to forecast the probability of a binary event occuring. It is a type of statistical model often used for classification and predictive analysis. It estimates the probability of an event occuring based on a given dataset of independent variables. That is, it uses X_train for training and based on that it predicts the class label of X_test. Since the outcome is a probability, the dependent variable is bounded between 0 and 1.In logistic regression, the probability of success divided by the probability of failure. It utilizes sigmoid function to classify the reviews into the positive and negative. Weight value is multiplied with the input value resulting into the values between 0 and 1. If the value is above 0.5, the review classifies as positive, otherwise as negative. Sklearn module provides the model and trains on X_train..

Decision Tree

Decision Trees are a type of a non parametric Supervised Machine Learning where the data is continuously split according to a certain parameter. Parameter is a condition or a test that makes subsets based on satisfied or not satisfied. This algorithm is well employable on the binary classification problem, the kind this project demands of. Decision tree has two parts namely internal nodes and leaf nodes. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. In the case of consideration that is positive and negative reviews are the leaf nodes and decision nodes formation is the part of machine learning.

Random Forest-

Random Forest is a powerful and versatile supervised machine learning algorithm that grows and combines multiple decision trees to create a forest. This algorithm is an enhancement of decision trees where we use multiple trees instead of using just one decision tree. This algorithm is based on ensemble learning since it combines multiple classifiers to solve a problem and hence improving its performance. Using single decision tree has drawbacks like overfitting. Random Forest overcomes overfitting by building multiple decision trees over random feature subsets (hence random forest). More the number of trees, better the performance of model. Output of all the decision trees is then aggregated (averaged in case of regression, maximum one winner in case of classification) and forecast is displayed. Having actual inputs in the model helps gaining higher accuracy. It also takes less time as compared to the other algorithms. Works even on large datasets. The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets). Step-3: Choose the number N for decision trees that you want to build. Step-4: Repeat Step 1 & 2. Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes. The only disadvantage of this algorithm is that it does not perform well on regression tasks**.**

Naive Bayes-

Naive Bayes is one of the simplest machine learning algorithm. It is a probabilistic Machine Learning model used for classification and is based on the theorem of Bayes. Bayes theorem states that probability of event A can be determined, given the conditional probability of occurrence of an event B given both events are -

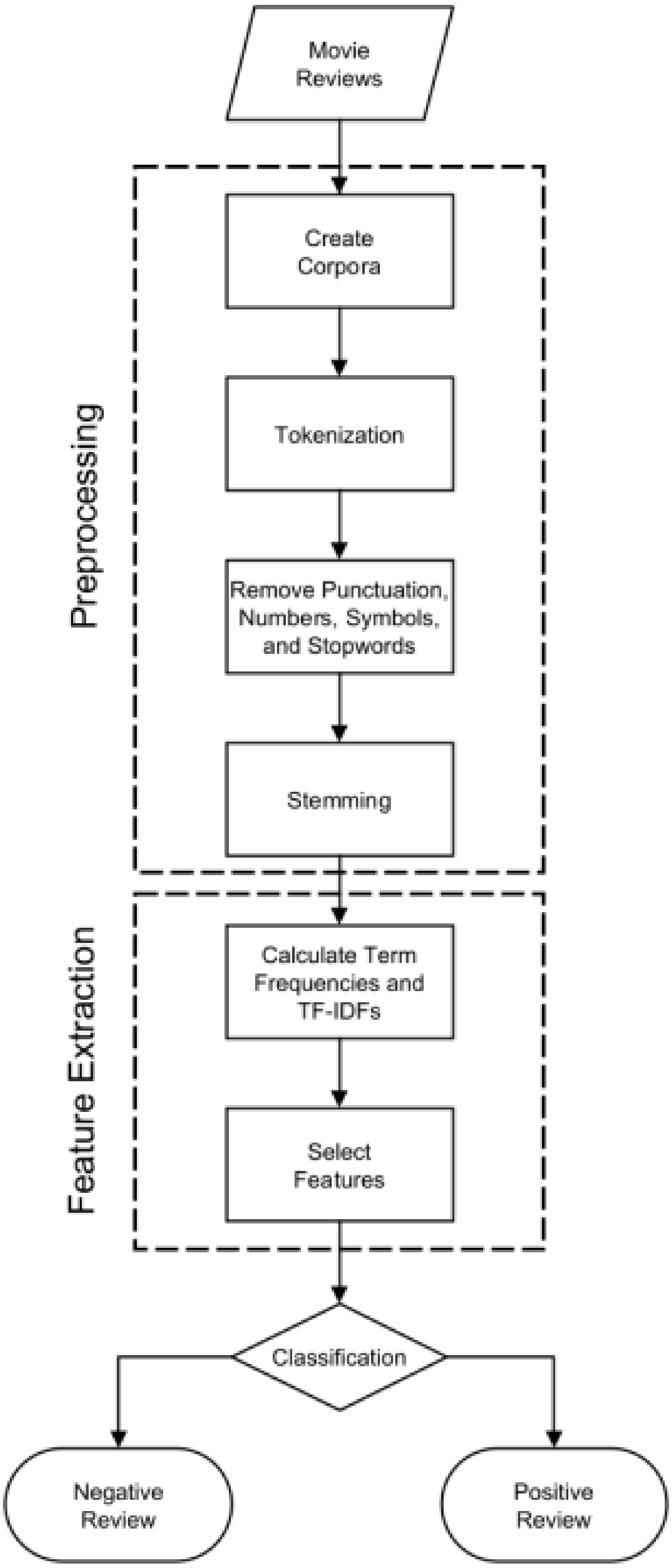
$$P(A/B) =P(B/A)p(A)/P(B)$$

Where A and B are events and P(B) != 0

With regards to the given dataset, the above theorem can be applied as follows –

$$P(y/X) =P(X/y)P(y)/P(X)$$

where, y is class variable and X is a dependent feature vector (of size n) where: X =(x1 ,x2,....) \
This algorithm is generally applied when training dataset is small.



Support Vector Machines

Support Vector Machine (SVM) is a robust classification and regression technique that maximizes the predictive accuracy of a model without overfitting the training data. SVM is particularly suited to analyzing data with very large numbers (for example, thousands) of predictor fields. SVM has applications in many disciplines, including customer relationship (CRM), facial and other image recognition, bioinformatics, text mining concept extraction, intrusion detection, protein structure prediction, and voice and speech recognition.

SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, then the data are transformed in such a way that the separator could be drawn as a hyperplane. Following this, characteristics of new data can be used to predict the group to which a new record should belong. In this case, it predicts class label i.e.positive or negative.

Evaluation Parameters

The majority of state-of-the-art sentiment analysis makes use of accuracy, F1 score, and precision. Sentiment analysis using deep learning architectures: a review utilizes recall and accuracy as performance metrics. These metrics are as follows
:
True Positive(TP): The number of positive reviews that have been correctly classified.
True Negative(TN):The number of negative reviews correctly classified as negative.
False Positive(FP): Number of incorrectly classified positive review. False Negative(FN): Number of incorrectly classified negative review.

Precision

Precision is defined as the ratio of correctly classified positive samples to the total number of samples predicted as positive. This metric can be used to indicate the strength of the prediction. i.e., if a model has 100 percent precision, all the samples evaluated as positive are confidently positive. Precision = TP/(TP+FP).

Recall

Recall is also known as sensitivity. It is defined as the ratio of actual positive instances out of a total number of positive instances present in the classification. It measures the misclassifications done by the model. Precision and recall are inversely proportional to each other. Therefore it is impossible to increase both Precision and Recall at the same time. A recall is used in cases where the capture of a class is dominant. Recall = TP/(TP+FN).

F1 score

F1 score is the harmonic mean of Recall and Precision. It is the most used metric after Accuracy. It is used when we are unable to choose between Precision or Recall. F1 score manages the trade-off between recall and precision.

Accuracy

Accuracy is the most commonly used metric in all the classification tasks. Accuracy defines how accurate the model is. It is the ratio of correct classification to total predictions done by the model. Accuracy is a good metric to use for sentiment classification for a balanced dataset.

Confusion matrix

A confusion matrix is a table that is frequently used to evaluate a categorization model's (or "classifer's") effectiveness on a set of training test data values are known.

TF-IDF

Term Frequency refers to the number of times a term is present in a document. TF which counts the number of times a term word appears in the document Because each document is varied in length, it is likely that a term will appear far more frequently in longer documents than in shorter ones. As a result, the phrase frequency is frequently divided by the document length.

Conclusion

1. We obtained high accuracies with all the machine learning methods indicating ML algorithms can be well used to perform Sentiment analysis.
2. Support Vector Machine Classifier has the highest accuracy but takes training time.
3. Decision Tree has the lowest accuracy and low training time.
4. Results can be further improved by adding more steps in data preprocessing or try different feature extraction method than tf-idf vectorizer. 5. Appending emoticons increased the accuracy of the models

References

Acheampong FA, Nunoo-Mensah H, Chen W (2021) Transformer models fortextbased emotion detection: a review of BERT-based approaches.

Artif Intell Rev 54:5789–5829 Adomavicius G, Kwon Y (2011) Improving aggregate recommendation diversity using ranking-based techniques. IEEE Trans Knowl Data Eng 24(5):896–911

Ahmad S, Asghar MZ, Alotaibi FM, Awan I (2019) Detection and classifcation of social media-based extremist afliations using sentiment analysis techniques. Hum Centric Comput Inf Sci 9(1):1–23

Ahmad SR, Bakar AA, Yaakub MR (2019) A review of feature selection techniques in sentiment analysis. Intell Data Anal 23(1):159–189

Akhtar MS, Ekbal A, Cambria E (2020) How intense are you? predicting intensities of emotions and sentiments using stacked ensemble [application notes]. IEEE Comput Intell Mag 15(1):64–75

Akhtar N, Zubair N, Kumar A, Ahmad T (2017) Aspect based sentiment oriented summarization of hotel reviews. Procedia Comput Sci 115:563–571

Al Amrani Y, Lazaar M, El Kadiri KE (2018) Random forest and support vector machine based hybrid approach to sentiment analysis. Procedia Comput Sci 127:511–520

Birjali M, Kasri M, Beni-Hssane A (2021) A comprehensive survey on sentiment analysis: approaches, challenges and trends. Knowl-Based Syst 226:107134

Bhaskar J, Sruthi K, Nedungadi P (2015) Hybrid approach for emotion classifcation of audio conversation based on text and speech mining. Procedia Comput Sci 46:635–643

Bhatia P, Ji Y, Eisenstein J (2015) Better document-level sentiment analysis from rst discourse parsing. arXiv preprint arXiv:150901599

Calders T, Verwer S (2010) Three naive bayes approaches for discrimination-free classifcation. Data Min Knowl Disc 21(2):277–292

Cambria E, Das D, Bandyopadhyay S, Feraco A (2017) Afective computing and sentiment analysis. In: A practical guide to sentiment analysis. Springer, pp 1–10

Cao Q, Duan W, Gan Q (2011) Exploring determinants of voting for the “helpfulness’ of online user reviews: a text mining approach. Decis Support Syst 50(2):511–521