Sentiment Analysis of IMDB Movie Reviews

ABHISHEK
Asst. Professor, ECE Department
KIET GROUP OF INSTITUTIONS, GHAZIABAD
abhishek.ece@kiet.edu

Prakhar Singhania
Student, ECE Department
KIET GROUP OF INSTITUTIONS
prakhar.1923ec1135@kiet.edu

Mahak Singh
Student, ECE Department
KIET GROUP OF INSTITUTIONS
mahak.1923ec1184@kiet.edu

Abstract - The interpretation and classification of emotions in text data via textual analysis techniques is commonly referred to as sentiment analysis. Not only polarity positive, negative, or neutral, but sentiment and emotions, as well as intentions, such as interested v. not interested, are taken into account by sentiment analysis models. Large businesses invest considerable amounts in predicting the outcome of their activities.

In the field of marketing, customer support and clinical medicine sentiment analysis is often applied. It is intended to give voice to customer materials, such as feedback and survey results, information on the Internet and Social Media in addition to Health Information. Now, thanks to the development of deep language models like ROBERTa, these data sets can be analyzed in greater detail, for example news stories where authors often express their opinions and feelings without explicitly expressing them.

INTRODUCTION

Cinema is the most practical form for people to amuse themselves. Nevertheless, it's a small number of films that are very popular and profitable. One of the many rating sites is available to movie lovers, allowing them to choose which films they want to see and skip. IMDB, Rotten Tomatoes, etc. are among the most popular websites in this category.

These websites reward a film for its success with 10 stars based on the ratings given to it by viewers. However, there is no method to make predictions based on film reviews. The assessment of the popularity of a film, based upon its reviews, shall therefore be accompanied by sentiment analysis. On the other hand, based on its reviews, a film can't be predicted. A sentiment analysis shall therefore be used to assess the popularity of a film on the basis of its reviews.

LEVELS OF SENTIMENT ANALYSIS

Sentiment Analysis Levels:

Document-level: A review of the procedure applied to this document in its entirety. A paper on a particular subject is part of the classification level. The customer tends to compare two or more aspects in the analysis of document levels.

Sentence level: The classification of subjectivity is directly connected with the analysis of sentence levels. Identifying whether a word is positive, negative or neutral is the objective of the analysis of sentence level. For the analysis of sentence level sentiment, all classifiers used are those found in text analysis.

Aspect level: Sentiment analysis for the aspect level is utilized to detect sentiments regarding these people' aspects. Consider this example. "My car has good handling, but it is a little bit." In this situation of a vehicle, there is an opinion that the handling of a cat is good but the car is bad. The competitive comment is included in the SA on the aspect stage.

Phrase level: The phrase with the terms of opinion identifies the expression level classification. These have both advantages and disadvantages because the advantage is that they have an exact opinion. However, because of relational polarity, the outcome cannot be precise.

Feature Level: The product's feature is to define product qualities. A feeling analysis at a functional level is referred to in the text as an examination of these elements for the identification of feelings. The retrieved features determine whether the view is positive, negative, or neutral.

RELATED WORK

By anticipating the attitudes from the structure of phrases, namely the objectivity and subjectivity in the sentences, Verma et al. (2009) [1] have provided an intriguing viewpoint. They took into account the negation of a few adjectives and phrases and gave the entire text the greatest feeling subjectivity rating. SentiWordNet, which contains polarity scores for collections of synonyms in the English language, was used to extract the sentiment subjectivity values. They later produced additional TF-IDF vectors based on the emotion vectors, after which they computed information gain. After using the information gain-based pruning, their maximum accuracy was 83%

Numerous machine learning applications, such as logistic regression and linear regression, were adopted by Vr, Nithin, and Babu Pb, Sarath [2]. Their linear regression model has a 51% accuracy rate. Using logistic regression, they achieved an accuracy of roughly 42.2%. Movie sequels cannot be predicted using this technique. The success of a movie should be seen relative to other factors, therefore it cannot be predicted using only one factor, in this case, gross revenue.

The dataset utilized in this study is the one that Pang et al. offered to address the issue of rating similarity between films and the five-star rating, where certain ratings have a tendency to be extremely similar to others but have a tendency to award different ratings. To determine if a certain viewpoint is greater, less, or equal to other movie evaluations, they used human judges in a separate survey. They developed three models and made use of the term frequency feature to automate the same process. The work of Pang and Lee (2005) is renowned for its use of multiple ratings in numerical non-categorical data, which have since been widely employed in place of the binary classifications. This helped to address the problem by considering it an SVM problem.

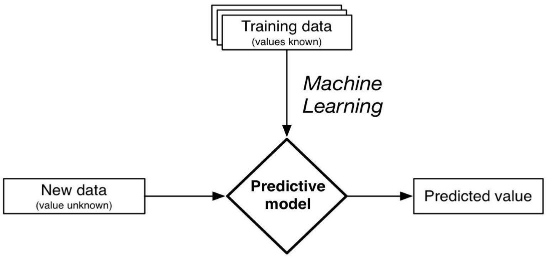
MACHINE LEARNING APPROACHES

A subset of artificial intelligence called machine learning focuses on teaching computers to learn from data and grow as a result of that data understanding. Applications for machine learning grow more effective as they use more data to learn from and develop. In order to train the model to find patterns and correlations between the features in huge datasets and to make predictions based on the training data, machine learning methods are utilized.

A variety of machine learning algorithms, strategies, and methodologies can be utilized to create models that use data to address real-world issues. supervised, unsupervised, and reinforcement learning approaches are the different categories for machine learning techniques.

The majority of **unsupervised learning** techniques use lexicon-based approaches, which create and update sentiment prediction using pre-existing lexical resources like WordNet and language-specific sentiment seed words. Unsupervised learning techniques can produce large emotions, but they are not able to gather context or domain specific information of the content without a corpus of previously classified data.

Machine learning is used in **supervised learning** methods to classify data that have already been approximated. These pre-classified datasets are often domain specific, so it is not possible to use models produced by them in a particular area. The data are then converted into intermediate models representing documents as vectors before feeding them to the machine learning algorithm.



Simplified representation of a supervised learning model.

By interacting with their surroundings, machines can learn using **reinforcement learning**. With this method, the computer learns by carrying out various tasks with rewards and penalties for each one. Every right behaviour will be rewarded, and every bad action will be punished, during the training of the model. Findinga tactic that maximizes the rewards is the main objective ofreinforcement learning. The machine operates independently and without human supervision in reinforcement learning. The development of games, marketing, and advertising are a few examples of the many uses of reinforcement learning.

CLASSIFICATION TECHNIQUES

We used Nave Bayes, Logistic Regression, and Support Vector Machine in our experiment. We trained our model using the aforementioned classifiers to determine whether a movie is goodor bad.

NAIVE BAYES

The ***Naive Bayes*** algorithm is a system to supervise learning that uses the Bayes theorem to solve classification problems. It's applied mainly for classification of text that includes a high quality training data set. The Nave Bayes Classifier is one of the simplest and most efficient classification algorithms that helps to build fast machine learning models, which allow for a rapid prediction. Bayes' theorems are also known as Bayes' Law or Bayes' Rule, which are used to determine the probability of any hypothesis based on prior knowledge. That's based on the conditional probability. The formula of Bayes' theorem is as follows:

$$\mathrm{P(A\,|\,B)} = \frac{P(B|A)P(A)}{P(B)}$$

P(A|B) : Posterior probability:

P(B|A) : Likelihood probability

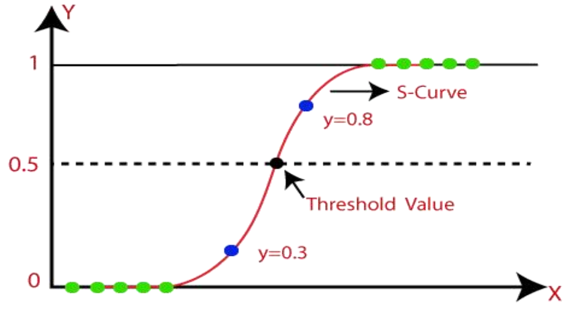
P(A) : Prior Probability

P(B) : Marginal Probability

LOGISTIC REGRESSION

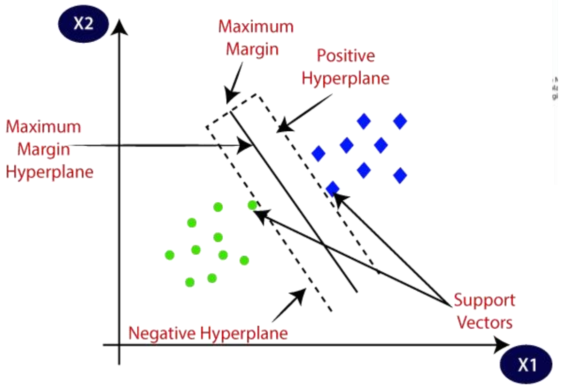
Logistic regression is a supervised machine learning algorithm for classification tasks, designed to predict the probability that an instance belongs to one of those classes. Its name is ***logistic regression*** and it's used for classification algorithms. It shall be called regression because by entering the linear regression result as input and using a **sigmoid** function to estimate the probability of this class, it has been referred to as regression. The difference between linear and logistic regression is that, while logistic regression can predict the likelihood of an instance belonging to a given class or not, linear regression outputs are continuous values which may be anything.

Sigmoid function



SVM (SUPPORT VECTOR MACHINE)

Support vector machines or SVM are the most widely used supervised learning algorithm for classification and regression problems. But it is primarily used to resolve classification problems in the area of machine learning. To make it easy for a new data point to find its correct category in the future, SVM's algorithm is dedicated to creating an optimal line or decision boundary that allows n dimensional space to be split according to classes. This best decision boundary is referred to by a hyperplane. SVM selects the extreme points and vectors that help create the hyperplane. Support vectors are those special cases, and their algorithm is called support vector machine. From this diagram we will be able to see that two categories are divided according to the decision boundary or hyperplane:



CONCLUSION

An important objective of this research is to develop a sentiment analysis model that will allow us to better understand the movie reviews we have collected.

One of the most significant pillars of the decision-making process is without a doubt the use of sentiment analysis techniques. To create useful services or products, many people rely on sentimental analysis. We started with a model that produced IMBD movie reviews fairly well. Consequently, the pre-trained language model idea outperformed the most recent academic research. A phrase that is used negatively might be connected with something positive, and vice versa, because human languages are rather challenging for robots to understand.

The primary focus of the paper is a comparative analysis of various machine learning techniques for sentiment analysis of reviews. LR performs better than Naive Bayes and SVM. We can draw the conclusion that an algorithm performs better in forecasting the success rate of films when the data is cleaner.

REFERENCES

[1]. B. P. Verma S, "Incorporating semantic knowledge for sentiment analysis. Proceedings of ICON," 2009.

[2]. Vr, Nithin & Babu Pb, Sarath. (2014). Predicting Movie Success Based on IMDB Data

[3]. Danny Varghese “Comparative Study on Classic Machine learning Algorithms, 2018”

[4].Wikipedia- “Sentiment Analysis”.