JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Sentimental Analysis: Machine Learning Approaches

[1]Khushi Mehta, [2]Riya Patel, [3]Heli Parikh, [4]Neel Patel, [5]Sejal Thakkar

[1,2,3,4] U.G Student, Department of Information Technology Engineering, IITE, Ahmedabad, Gujarat, India
[5] Assistant Professor, Dept. of Computer Engineering, IITE, Ahmedabad, Gujarat, India

Abstract: IMDb is a well-known database for movies, video games, internet streaming content, and series. One is given access to sources and reviews of the particular film they are looking for, as well as the choice of whether or not they want to write their own. After analyzing all the comments and categorizing them based on positive and negative views, classification models are used to predict the sentiment of all the positive or negative assessments. This helps provide the vast datasets for the implementation of sentiment analysis on IMDB. The primary goal is to provide the best practicable review based on the reviews and to make the appropriate approximations. The models are trained for several classifiers using a combination of unique text values and sentiment analysis is conducted.

***Keywords* – Sentiment Analysis, Machine learning, Lexicons, Polarity, Sentiments, Social Media**

1. Introduction

There were roughly 11 million reviews on 8.4 million titles in the given database as of the most recent stories from February 2022. Users can quickly post reviews on a specific video game, television show, or movie using the services provided, providing the necessary opinions and reviews for sentiment scrutinization. The primary perspective of an analysis of people is, in essence, the rating of a certain video or television show. This is utilized generally for the classification of sentiment analysis.

The search feature on Netflix might be used by someone who wants to watch a movie but is unsure about what to choose. The Training Set contains nothing more than a textbook-style collection of good and negative IMDb movie reviews. Included in the test set is The Textbook of Unlabeled Movie Reviews. We chose various subsets of the training set to be the confirmation set for evaluation and model choice using the K-fold Cross Confirmation technique. By eliminating stop words, we cleaned up the data for preprocessing. Keeping the words inside a certain frequency range helped to remove noise and improve our scores. Moreover, the number of occurrences of a word does not invariably indicate its significance throughout the documents. Stemming and N-grams (up to trigrams) were eventually used in some models. Humans can choose what they want to see by using this approach as a suggestion tool. Decision trees, logistic regression, Multinomial Naive Bayes, and support vector machines are some of the classifiers available in the SciKit Learn package.

2. Needs for sentimental analysis

Sentiment analysis is amazing because it enables individuals and organizations to comprehend the feelings of their clients and employees. Sentiment analysis refers to the methods and tactics that let businesses look at the information on how subscribers change the way they feel about various movie genres. Positive or negative emotions are determined by their polarity. Sentiment analysis is a technique that automatically examines statements made in natural language, identifies the key ideas they convey, and groups them into categories based on how they make you feel.

Tools for sentiment analysis is crucial for identifying and comprehending client sentiment. Companies may enhance CX by using these methods to learn how consumers feel. Tools for sentiment analysis provide insights on how businesses might improve customer service and experience. Customer satisfaction analysis using sentiment analysis: The client uses comments in natural language to share his experience with a product and to express his thoughts and attitudes toward it. This gives us vital information about whether the customer is happy and, if not, how we can enhance the product. Determine problems in real-time and take action. A customer can rapidly express his dissatisfaction to the entire globe through social media.

3. DIFFERENT APPROACHES FOR SENTIMENTAL ANALYSIS

Many researchers use sentiment analysis in their work, and there are numerous ways to do so. Because of its significance in this situation, numerous studies are still being conducted to identify better options. This paper illustrates the main 4 ways Lexical based, Machine Learning, Hybrid Approach, etc.

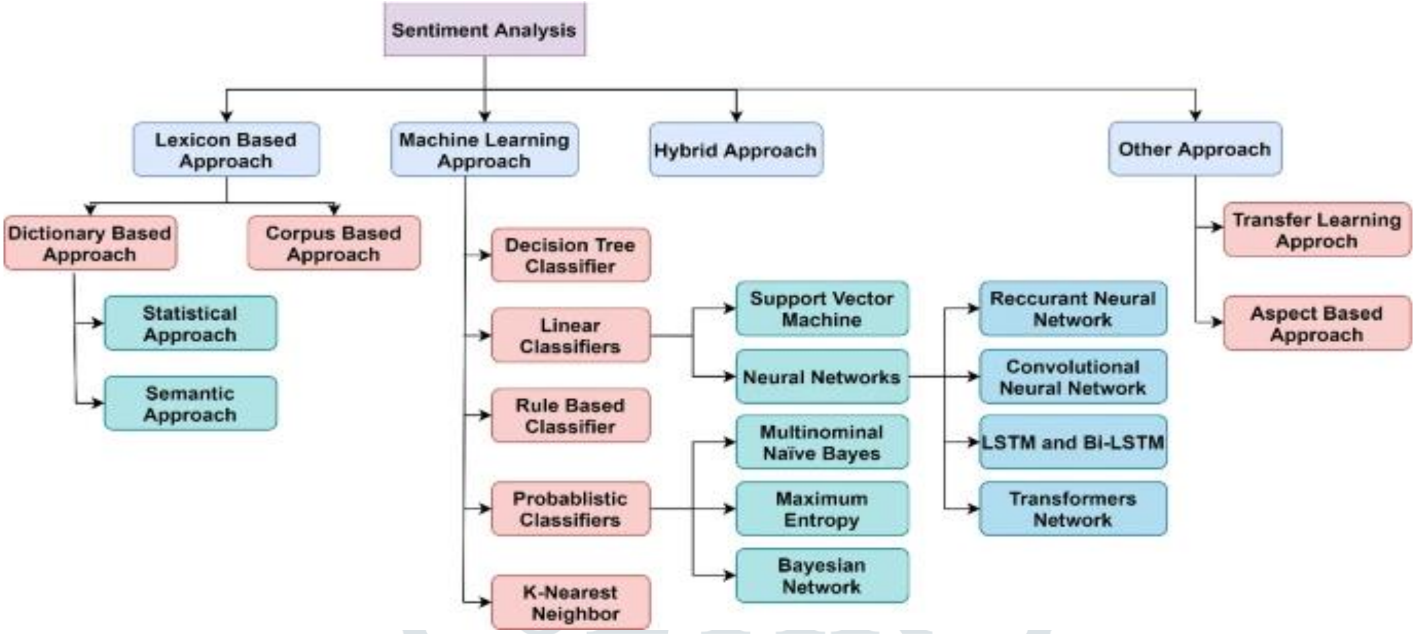

Fig 1: Different Sentiment Analysis Finding Approach

3.1 An Approach Based on Rules

A rule-based technique is employed by defining several rules for obtaining the opinion. These rules are produced by tokenizing each sentence in each document and then verifying each token, or word, for its occurrence. If the word is present and has a favorable connotation, it earns a +1 rating. Each comment begins with a neutral score of 0 and is rated as positive. The rule-based approach's output is verified or challenged afterward to determine whether it is accurate.

3.2 Lexical-Based Method

Lexicon-based approaches operate under the presumption that the total polarity of a statement or document is equal to the sum of the polarities of all its identities. The lexicon-based method proposed was applied at the seminar ROMIP 2012. The emotional research used to create sentiment analysis dictionaries for each area is the foundation of this method. The word-modifier modifies (increases or reduces) the weight of the appraisal word that comes after it by a specific percentage. Word-negation changes the weight of the succeeding assessment word by a specific offset: positive words become heavier, while negative words get heavier. A one-dimensional emotional environment is created for all the messages. Utilizing the cross-validation method, the percentage of deletions was calculated. The average text weights for each sentiment class for training texts were then determined. The class closest to the one-dimensional emotional space was referred to in the categorized text.

3.3 Machine Learning Approach

Natural language processing, text analysis, computational linguistics, and other methods are used in sentiment analysis to identify and measure the sentiment of text or voice. Machine learning approaches initially train the algorithm using known inputs and results so that it can later deal with novel unknown data. Sentiment analysis algorithms are trained to read beyond simple definitions to comprehend relevant information, irony, and wrongheaded words. The most well-known machine learning-based creations are as follows.

4. VARIOUS MACHINE LEARNING ALGORITHMS FOR SENTIMENTAL ANALYSIS

4.1 Naive Bayes

The simplest and fastest classification approach for a huge amount of data is naive Bayes. Naive Bayes classifier is extensively employed in several applications, namely fake news detection, text classification, sentiment analysis, and virtual assistants. For predicting unknown classes, the Bayes probability theorem is utilized. The Naive Bayes classification performs well for textual data analysis, such as Natural Language Processing. Given the individual probabilities of events A and B as well as the conditional chance that event B happens, the conditional probability that event A occurs. It is presumed that characteristics are independent in this context. Mostly utilized when the training set size is less.

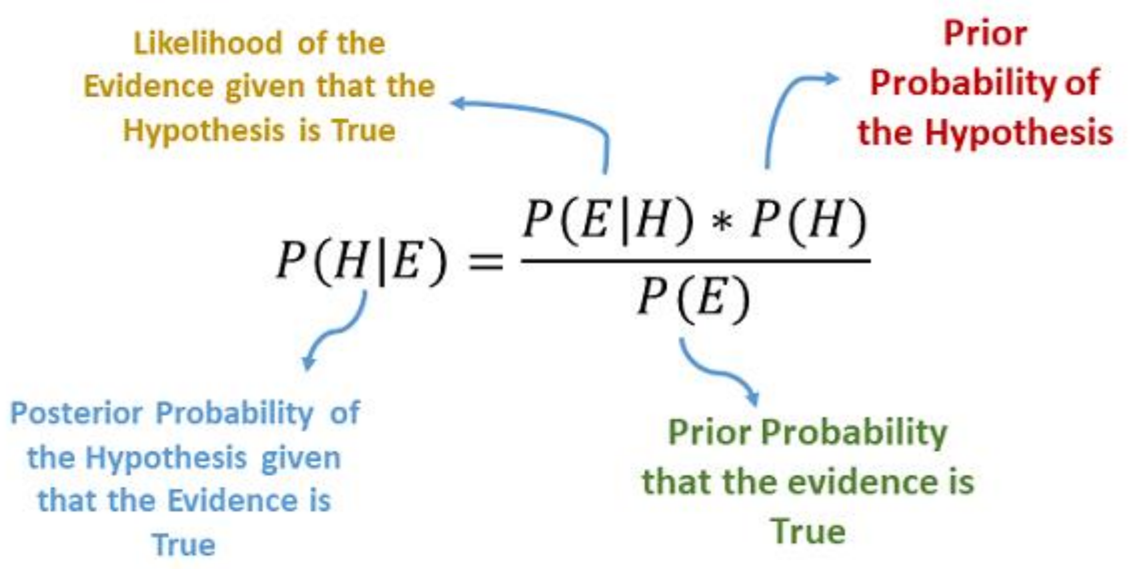
Likelihood of the Evidence given that the Hypothesis is True
Prior Probability of the Hypothesis
$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$
Posterior Probability of the Hypothesis given that the Evidence is True
Prior Probability that the evidence is True

Fig 2: Naïve Bayes Formula

P(H)/(E) = (P(H) * P(E/H)) / P(E)

Therefore, the equation is changed into the following to obtain the emotion. Where, Sentiment words = H and Sentence = F

4.2 Logistic regression (LR)

By dividing an input value by a weight value, the logistic regression machine learning method operates. It is a classifier that discovers which attributes of the input aid distinguish between positive and negative classes. A probabilistic regression approach used for classification problems is called logistic regression. Logistic regression is frequently used for applications that need binary categorization. The ratio of odds is calculated using logistic regression when there are several explanatory factors. Maximum-likelihood is used in logistic regression to choose the optimal parameters. The independent variables might be either continuous or discrete, for example (ordinal and nominal).

4.3 Support Vector Machine

It is a non-probabilistic classifier that requires a sizable training set. Points are categorized using a (d-1)-dimensional hyperplane. The main goal of SVM is to identify the hyperplane that best divides the data into several groups. SVM then looks for the hyperplane with the largest practical margin. In this, the border is defined by the dividing line, and the items either belong to the red or green class. Here, the original items are transformed or mapped using the kernel mathematical function, which is referred to as mapping. After transformation, the mapped items are linearly separable, which allows for the avoidance of complicated structures that require curves to divide the objects.

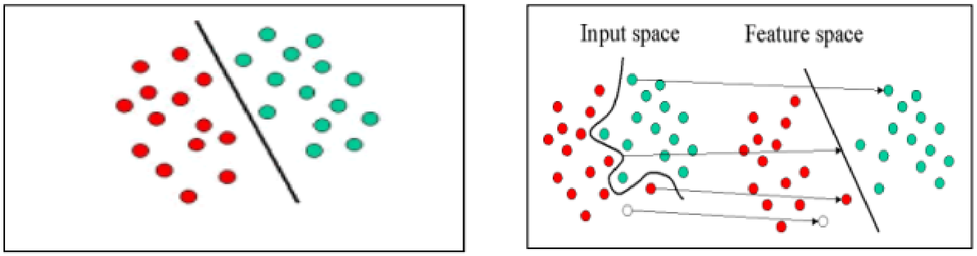

Fig 1. (a). Linear classifier (b). SVM illustration.

Fig 3: Linear Classifier and SVM graph

4.4 K-NN and Weighted K-NN

Finding the K closest matches in the training set of data, followed by utilizing the label of those matches to forecast, is how the KNN algorithm classifies the components of data.

The K-Nearest Neighbor approach is based on the idea that instances close to one another in a vector space will be classified roughly similarly. Further study was done on the weighted k-Nearest Neighbor approach, in which the training set's components were given weights, and the weights were then used to calculate the sentiment of the text word by word. In this case, the is used to determine the score.

Positive Score = ($\sum$Positive Score + $\sum$Negative Score) / $\sum$Q Maximum Score

Here, Q is equal to P plus N, or the combined count of positive and negative. Prior to tokenizing the phrases and removing the stop words from the tweets they have collected, the weighted k-NN approach tokenizes the sentences. Two parses are used to execute the method suggested by the authors of [8]. After the initial parsing, each review receives a favorable score. This is sent for a second processing and a neutral review input is provided. This is used to modify the score as needed. The goal is to improve positivism determination, and the result is an output file with the review ID and its positive score.

4.5 Multilingual Sentiment Analysis

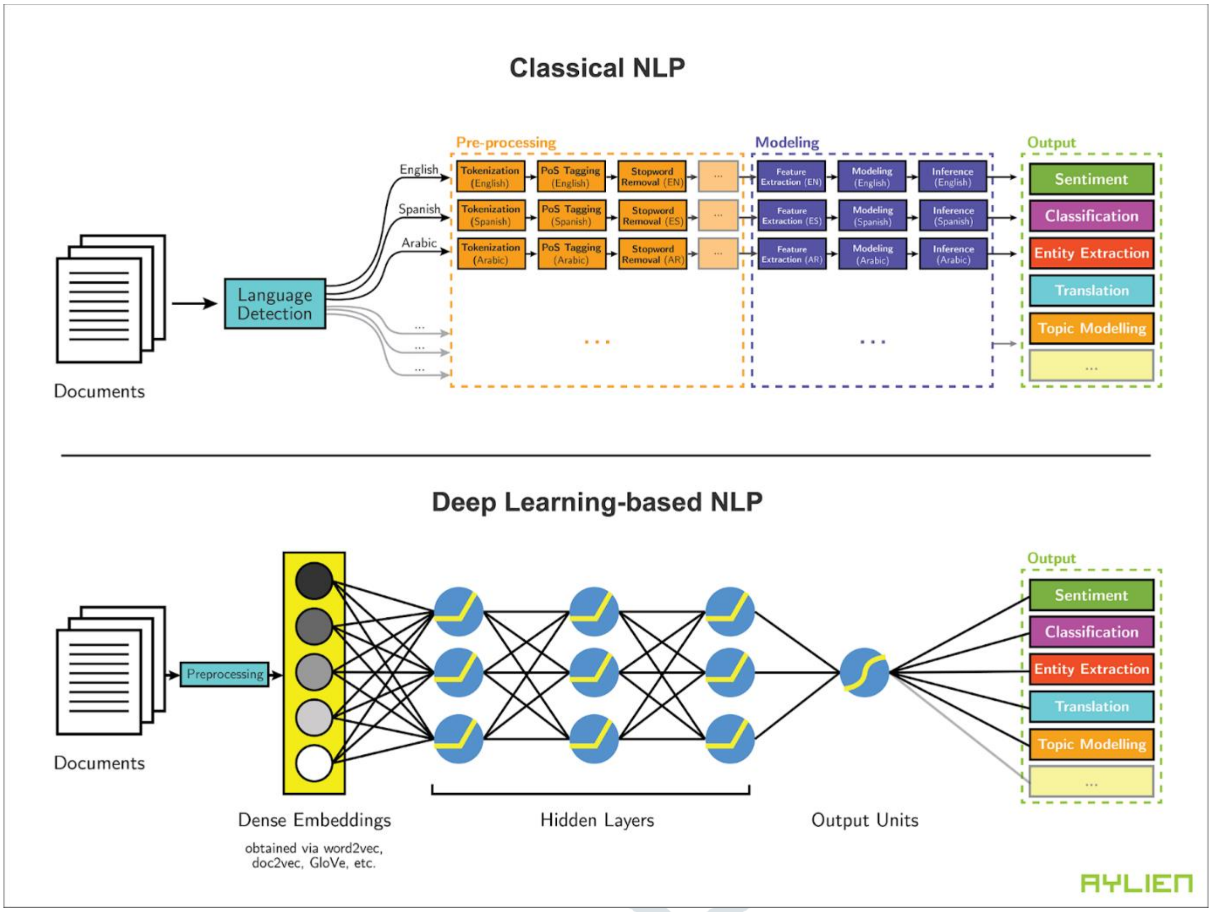

Fig 4: Multilanguage Sentimental Analysis

In order to complete the task of detecting the polarity of the text, a technique inside a multilingual framework is described and explained. Several Natural Language Tool Kits are used to complete it. Using language models, the first word in this language is identified. Using common translation software, the language is first identified and then translated to English. For the purpose of translation, they are using PROMT exceptional Translation (XT) Technology. They then go on to the sentiment classification phase.

5. Conclusion

The use of sentimental analysis in today's digital era has been appreciated. This article gives a brief about various sentiment analysis methods and the different levels of analyzing sentiments that are available today. The best methods for the implementation are k-fold cross-validation and Naive Bayes. Generally, reviews are based on two categories, positive and negative. With the help of this approach shown here, we can easily make a decision based on the huge amount of reviews gathered online. This not only benefits the user but all movie makers to see the status of the film. In a world where people depend on social media or blogs, reviews, and comments, sentiment analysis is very useful and beneficial for a better decision-making process. We have come across some other methods like Approaches based on rules and lexical-based methods.

6. REFERENCES

[1] Wankhade, Mayur, et al. “A Survey on Sentiment Analysis Methods, Applications, and Challenges - Artificial Intelligence Review.” *SpringerLink*, doi.org, 7 Feb. 2022, https://doi.org/10.1007/s10462-022-10144-1.

[2] Gamal, B. (2021, April 11). Naïve Bayes algorithm. Medium. https://medium.com/analytics-vidhya/na%C3%AFve-bayes-algorithm-5bf31e9032a2

[3] Devika, M., Sunitha, C., & Ganesh, A. (2016). Sentiment analysis: A comparative study on different approaches. Procedia Computer Science, 87, 44-49. https://doi.org/10.1016/j.procs.2016.05.124

[4] Leveraging deep learning for multilingual sentiment analysis. (2016, August 14). Analytics & IIoT. https://analyticks.wordpress.com/2016/08/14/leveraging-deep-learning-for-multilingual-sentiment-analysis/

[5] Performing sentiment analysis with naive Bayes classifier! (2022, August 4). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/07/performing-sentiment-analysis-with-naive-bayes-classifier/

[6] Sentiment analysis using product review data. (2015, June 16). SpringerOpen. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2

[7] (n.d.). Research India Publications was established in 1998 and now today we are one of the leading International Publishers. https://www.ripublication.com/ijaer18/ijaerv13n16_53.pdf

[8] L. Nahar, Z. Sultana, N. Iqbal and A. Chowdhury, "Sentiment Analysis and Emotion Extraction: A Review of Research Paradigm," 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 2019, pp. 1-8, doi: 10.1109/ICASERT.2019.8934654.

[9] IEEE Style Citation: Rajwinder Kaur, Prince Verma, “Sentiment Analysis of Movie Reviews: A Study of Machine Learning Algorithms with Various Feature Selection Methods,” International Journal of Computer Sciences and Engineering, Vol.5, Issue.9, pp.113-121, 2017.