iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** XII **Month of publication:** December 2022

DOI: https://doi.org/10.22214/ijraset.2022.47795

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

Sentiment Analysis of IMDb Movie Reviews

Avi Ajmera
Department of Computer EngineeringMukesh Patel School of Technology Management & Engineering Mumbai, India

Abstract: Sentiment analysis is an artificial intelligence branch that focuses on expressing human emotions and opinions in the sense of data. We try to focus on our sentiment analysis work on the Stanford University-provided IMDB movie review database in this post. We look at the expression of emotions to determine whether a movie review is positive or negative, then standardize and use these features to train multiple label classifiers to identify movie reviews on the right label. Since there were no solid framework structures in a study of random jargon movies, a standardized N-gram approach was used. In addition, a comparison of the various categories' methods was conducted in order to determine the best classifier for our domain query. Our process, which employs separation techniques, has an accuracy of 83 percent.

Keywords: Natural language processing, Sentimental analysis, lemmatization, Stochastic Gradient Descent (SGD), Image Processing), Support Vector Machine (SVM)

I. INTRODUCTION

"What other people think" has always been the safest way to make well-informed shopping, service-seeking, and voting decisions. Because of the Internet, it is no longer limited to just friends, families, and neighbors; instead, it now includes complete strangers that we have never met before. On the other side, a growing number of people are turning to the internet as a means of communicating their opinions to those they do not know.[1]

Sentiment is a feature that decides whether a text's expressed opinion is positive or negative, or whether it is about a particular subject. "Logan is a really nice movie, highly recommended 10/10," for example, sharing positive feelings about the film Logan, which is also the title of this post. For example, in the text "I am amazed that so many people put Logan in their favorite movies, I felt a good watch but not that good," the author's feelings about the movie might be good, but not the same message. Above everything, Emotional analysis is a collection of processes, strategies, and tools for gathering and converting knowledge below the level of language, such as viewpoint and attitude.[2]

Emotional analysis has traditionally focused on the principle of harmony, or whether an individual has a positive, neutral, or negative perspective on something. The topic of emotional analysis was a product or service with publicly accessible online reviews.

This may explain why sentimental analysis and theoretical analysis are often used interchangeably; however, we believe that seeing emotions as emotionally charged ideas is more accurate. The aim of this paper is to categorize movie reviews by examining the magnitude (good or bad) of each category in the review [3]. We devised an approach to describe the polarity of the movie using these sentiment words because of the abundant usage of sentiment words in the study.

To determine which model works best for sentiment classification, we looked at network configuration and hyper parameters (batch size, number of epochs, reading rate, batch performance, depth, vocabulary size).

II. LITERATURE REVIEW

A. Natural Language Processing (NLP)

Natural language grammar, also known as NLP, refers to a computer program's capacity to comprehend human speech in its natural form. Natural language processing is a part of artificial intelligence. The development of natural language processing (NLP) applications is notoriously challenging since computers typically require people to "speak" to them in the language of an accurate, unambiguous, and well-structured system, or with a limited number of voice commands that are clearly written.

Human speech, on the other hand, is not always accurate; it is frequently vague, and the structures of languages can be influenced by a variety of factors, such as slang, regional languages, and the social context in which they are spoken. Syntax analysis and semantic naming are the two primary methods that are utilized in natural language processing. Syntactic analysis and semantic labelling are the two primary methods that are utilized in natural language processing. It is imperative that the word order in a syntax be understood to comprehend a method.

©IJRASET: All Rights are Reserved

B. Sentiment Analysis

Sentiment analysis, which can also be referred to as opinion mining, is a type of natural language processing (NLP) that determines the overall mood of a piece of written communication. The identification and organization of ideas regarding a product, service, or idea into categories is a common practice in business settings. It requires data analysis, machine learning (ML), and artificial intelligence (AI) to mine sensitive information and text that contains emotional content. [4]

Internet outlets such as emails, blog posts, service tickets, web chats, social media sites, forums, and tweets are some examples of places where organisations can use technology research tools to gather insights from informal and informal messages. It is possible to use algorithms in place of manual data analysis to perform legal, automatic, or hybrid methods of analysis. While working on this project, we made use of a few different steps to conduct research and think about both the meaning of the term and the context in which it is used. We began by inserting a summary token and splitting the terms into features. The random review was stopped in the first phase to ensure that there was no sexism in the procedure. We kept the movie summary in the list of positive or bad aspects in front of it until we ended the review. We then built a new list to store all the words from the view. The frequency distribution function was used in the second column to search the most frequent terms in all updates.

When the most familiar words offer a bad or positive idea, we now announce the root of these words. We've created a new list of the most famous terms and their origins, which will be used to train and validate our method. These terms can now be used as functions of our method to use various forms of algorithms.[5] We used separate Machine Learning Algorithms to discern any negative or positive feedback that you have added to it after separating the training and testing sets.

C. Text Pre-Processing and Normalization

Cleaning, pre-processing, and standardizing the text to bring text artefacts such as phrases and words to some degree format is one of the most critical steps before joining the state-of-the-art engineering and modelling process.

This allows for ranking over the entire text, which aids in the creation of meaningful functionality and reduces the noise that can be imported because of a variety of factors such as inactive characters, special features characters, XML and HTML tags, and so on. [9].

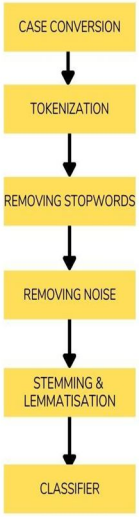


Fig. 1 Steps for Text Pre-processing

The following are the key components of our text normalisation pipeline:

1) *Cleaning Text:* Unnecessary content, such as HTML tags, often appears in our text and provides little meaning when evaluating emotions. Therefore, we must ensure that they are disabled before the functions are removed. The Beautiful Soup library is doing an excellent job of delivering the necessary facilities.
2) *Removing Accented Characters:* We work with changes in the English language in our database, but we need to make sure such characters survive and that other formats, such as computerized characters, are translated to ASCII characters.
3) *Expanding Contractions:* Problems are abbreviated forms of words or clusters of letters in the English language. By deleting certain letters and sounds from actual words or phrases, these shortened versions are created. Typically, vowels are replaced by English.

4) *Removing Special Characters:* Another critical aspect of text processing and design is the removal of special characters and icons that often add extra sound to meaningless text.
5) *Stemming and Lemmatization:* Word titles are always the simplest type of existing terms and can be created by adding prefixes and suffixes to the stem to create new words. Inflection is the term for this. A base is the method of acquiring a simple type of a name daily. Lemmatization is like stemming in that it eliminates word conjunctions to expose a word's basic form. The root word, rather than the root stem, is the basic form in this situation. The distinction is that the root word is always the proper dictionary word (which is also in the dictionary), but the root stem may not be.
6) *Removing Stopwords:* Stopwords or stop words are words that are of little or no value and are particularly important when creating significant elements from a text. When you do basic word or word frequency in a copy text, these are normally the words that end up in the plural.

D. *Feature Weighting Methods*

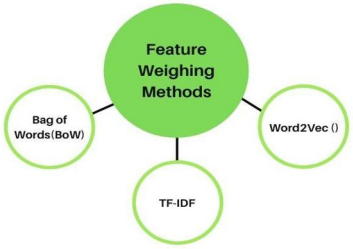


Fig. 2 Feature Weighing Methods

Feature Weighting is Selection of text-related functions depending on the number of continuous values. We sought to test the efficacy of weight-loss approaches in the sense of emotional isolation. We also tried to figure out what impact the split algorithm has. We picked the following three weight-measurement methods from among the exceptions.

1) *Bag of Words (BoW):* The number of times a word appears in a text is calculated by BoW. The paper is contrasted with this, and the comparisons are calculated in a variety of applications. The vector VT Nd is given to it as specified in the text T, with VTi denoting the number of times the word appears in the text T. D is a lexicon that includes all of the words in a series of revisions.
2) *TF-IDF:* Means a Document Frequency that can be adjusted over time. If it demonstrates the book's meaning of the expression. This is considered when evaluating diversity mining applications. Increase the number of times a word appears in the text thus removing the word's occurrence in the corpse.
3) *Word2Vec (W2V):* Term embedding is developed using Word2Vec. This is a model that uses two-layer neural networks to form word meaning. It accepts a broad corpus of text as input and generates a space vector of words as output. Each name in the corpus is assigned a spatial vector. The general meaning in the corpus is located very similar to the vector space, which is why names are shared. Prior to running the Word2Vec algorithm, the definitions are separated into sentences.

E. *Text Classifiers*

Text separators are used in sensory research to decide which movie reviews have the most votes. As a result, we analyse the influence of dividers on updates using the methods described below.

1) *Random Forest (RF):* It's a way for people to meet. Differentiated integration approaches are used to solve the approach used to boost performance. Several vulnerable students has been grouped together to form a group of good students. The decision tree is first read by a guarded machine. [6] When we cut under the tree, feedback is given at the apex, and information are obtained by breaking down the bucket into small sets.
2) *Decision Tree (DT):* A decision tree is a diagram layout that looks like a tree. Each internal hub is linked to a trait evaluation. The outcome of the test is addressed by each branch on the hub. The outcome is discussed by the leaf hub. It is fair in any new scenario that may occur. It is easy to understand and decode. Other preference methods can be used to converge a decision tree. The count becomes perplexing if the attributes are ambiguous or related.

3) *Naïve Bayes (NB):* Based on the Bayes principle. The appearance of one entity in the classroom has no bearing on the presence of another element. The NB model is straightforward and effective for massive data sets. The Bayes theorem is used to measure the back odds.
4) *P(c/x)= (P(c/x)*P(c))/P(x)*

P(c/x) → posterior probability of class given predictor.
P(c) → prior probability of class.
P(x/c) → likelihood.
P(x) → prior probability of predictor

5) *Support Vector Machine (SVM):* It's a machine learning algorithm that's supervised. It is commonly used for partitioning as well as the withdrawing side. To find a hyper plane, the concept is to partition the database into two sections. The space between the data point and the hyper plane is called the boundary. The aim is to assess the absolute limit of well-distributed new data potential. Smaller collections have higher accuracy, whereas massive data sharing takes longer.
6) *Neural Network (NN):* They work on a single record at a time. Compares record divisions of recognised classifications to learn. Errors from the first iteration are fed into the continuous flow, reducing errors. Back delivery uses forward simulation to change the weight of the link and is often used to train the network with proven good performance. It improves the precision of different data. Tuning and training on a massive database is challenging.
7) *K-Nearest Neighbor (KNN):* It employs all possible segmentation scenarios. The Euclidean parallel scale is used to establish new classifications. It's used for pattern analysis as well as statistical equations. The distance between two points in Euclidean space is known as the square root of the difference between the corresponding point links. If the complexity of the training database becomes larger, KNN becomes more complex.
8) *Stochastic Gradient Descent (SGD):* The increasing gradient is another name for it. Upgrade parameter values based on human training values. This method is useful when an optimization algorithm is looking for parameters. Instead of renewing coefficients at the completion of a series of parameters, it is done for each training event. [7] The gradient drop algorithm makes predictions for each data state, so large data sets take longer to process. The reduction of stochastic gradient produces a positive effect of massive data in a short amount of time.

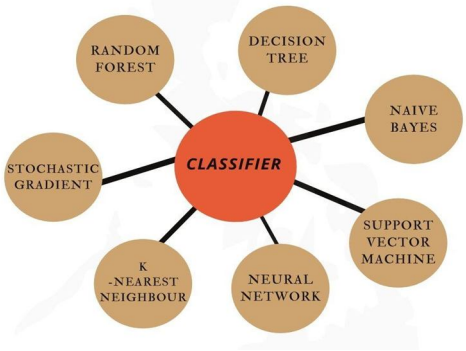


Fig 3 Classifiers

III. RESULTS AND DISCUSSION

As performance measurements, we used precision, recall, accuracy, and F1 ranking [17].

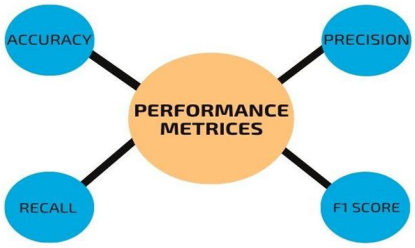


Fig 4 Performance Metrices

A. Accuracy

It is a subset of emotions that are distinct from the overall number of emotions. It's used to work out how effective a machine is.

Accuracy = (TP+TN)/(TP+TN+FP+FN)

where
TP → The number of positive emotions well classified as positive,
FP → The positive emotions, but separation does not classify you as good,
TN → The number of negative emotions well classified as negative,
FN → The negative emotions, but separation does not classify you as bad

B. Precision

It's a test of how many feelings have been correctly judged as right as opposed to the overall amount of well-separated emotions (P)

Precision = TP/(TP+FP)

C. Recall

It is the overall number of sensors in the absolute total that are correctly predicted (R)

Recall = TP/(TP+FN)

D. F1 Score

The F measure, which is used to incorporate application accuracy or memory, is a composite measure that measures a trade's accuracy and memory.

F1=(2*P*R)/(P+R)

1) The recall for the SGD classifier was 86.7 percent, while the accuracy, precision, and F1 ratings for the BoW + SVM combination were 84.4 percent, 83.3 percent, and 85 percent, respectively.
2) We obtained By combining TFIDF with other classifiers, we got good accuracy and precision for the SGD classifier (75.1 percent and 78.1 percent, respectively), while the TF-IDF + SVM combination generated a recall of 77.2 percent and an F1 score of 75.3 percent.
3) When we paired Word2Vec with other classifiers, we reached a high precision of 80.4 percent for the RF classifier, while the Word2Vec + SGD combination resulted in an accuracy of 82.2 percent, recall of 89.0 percent, and F1 score of 83.6 percent.
4) When we compared precision, we found::
a) In terms of BoW, SVM is very strong (83.3 percent)
b) In TF-IDF, SGD has a high percentage (80.9%).
c) In Word2Vec, Random Forest has a high score (76.6%).

	Accuracy	Precision	Recall	F_1 score
RF	0.821	0.818	0.829	0.822
DT	0.704	0.707	0.704	0.705
NB	0.787	0.809	0.756	0.792
SVM	**0.844**	**0.833**	0.864	**0.850**
NN	0.824	0.815	0.829	0.822
KNN	0.647	0.632	0.720	0.657
SGD	0.806	0.775	**0.867**	0.812

Table 1: Different classifiers with bag of words method

	Accuracy	Precision	Recall	F_1 score
RF	0.738	0.732	0.747	0.740
DT	0.624	0.620	0.633	0.626
NB	0.727	0.725	0.725	0.723
SVM	0.745	0.731	**0.772**	**0.753**
NN	**0.751**	0.738	0.761	0.749
KNN	0.666	0.659	0.682	0.668
SGD	**0.751**	**0.781**	0.693	0.751

Table 2: The performance of the TF-IDF feature weighting approach with various classifiers

	Accuracy	Precision	Recall	F_1 score
RF	0.809	**0.804**	0.817	0.826
DT	0.699	0.703	0.693	0.707
NB	0.748	0.763	0.723	0.755
SVM	0.725	0.746	0.684	0.735
NN	0.732	0.718	0.752	0.661
KNN	0.766	0.773	0.754	0.767
SGD	**0.822**	0.784	**0.890**	**0.836**

Table 3: On the basis of accuracy, a text classifier is compared to attribute weighting methods.

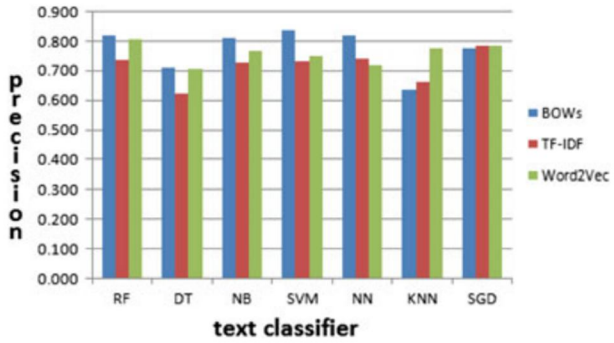


Fig 5: Comparison of text classifier with feature weighting methods based on precision

	BoW	TF-IDF	Word2Vec
RF	0.818	0.732	**0.804**
DT	0.707	0.620	0.703
NB	0.809	0.725	0.763
SVM	**0.833**	0.731	0.746
NN	0.815	0.738	0.718
KNN	0.632	0.659	0.773
SGD	0.775	**0.781**	0.784

Table 4: Precision metrics for different machine learning algorithms are compared.

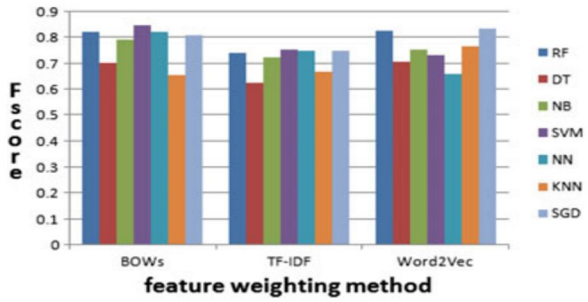


Fig 6: Focused on the F1 score, a text classifier is compared to feature weighting methods.

	BOWs	TF-IDF	Word2Vec
RF	0.822	0.740	0.826
DT	0.705	0.626	0.707
NB	0.792	0.723	0.755
SVM	**0.850**	**0.753**	0.735
NN	0.822	0.749	0.661
KNN	0.657	0.668	0.767
SGD	0.812	0.751	**0.836**

Table 5: F1 score metrics on different machine learning algorithms compared

IV. CONSTRAINTS

The below are the key obstacles for sentiment analysis: -

1) *Entity with a Name Recognize:* What precisely is the suser talking about, for example, is the phrase "300 Spartans" referring to a group of ancient Greeks or a movie?
2) *Anaphora Resolution:* The difficulty of determining what a pronoun or a noun phrase corresponds to is known as anaphora resolution. "We went to dinner and watched the movie; it was terrible." What exactly does "It" mean?
3) *Parsing:* What are the subject and object of this sentence, as well as the person or thing that the noun and/or adjective refer to?
4) *Sarcasm:* You won't know whether "bad" means "bad" or "good" if you don't know the author.
5) *Twitter:* Capitalization, Abbreviations, bad pronunciation, punctuation, and syntax are all problems on Twitter.

V. CONCLUSION

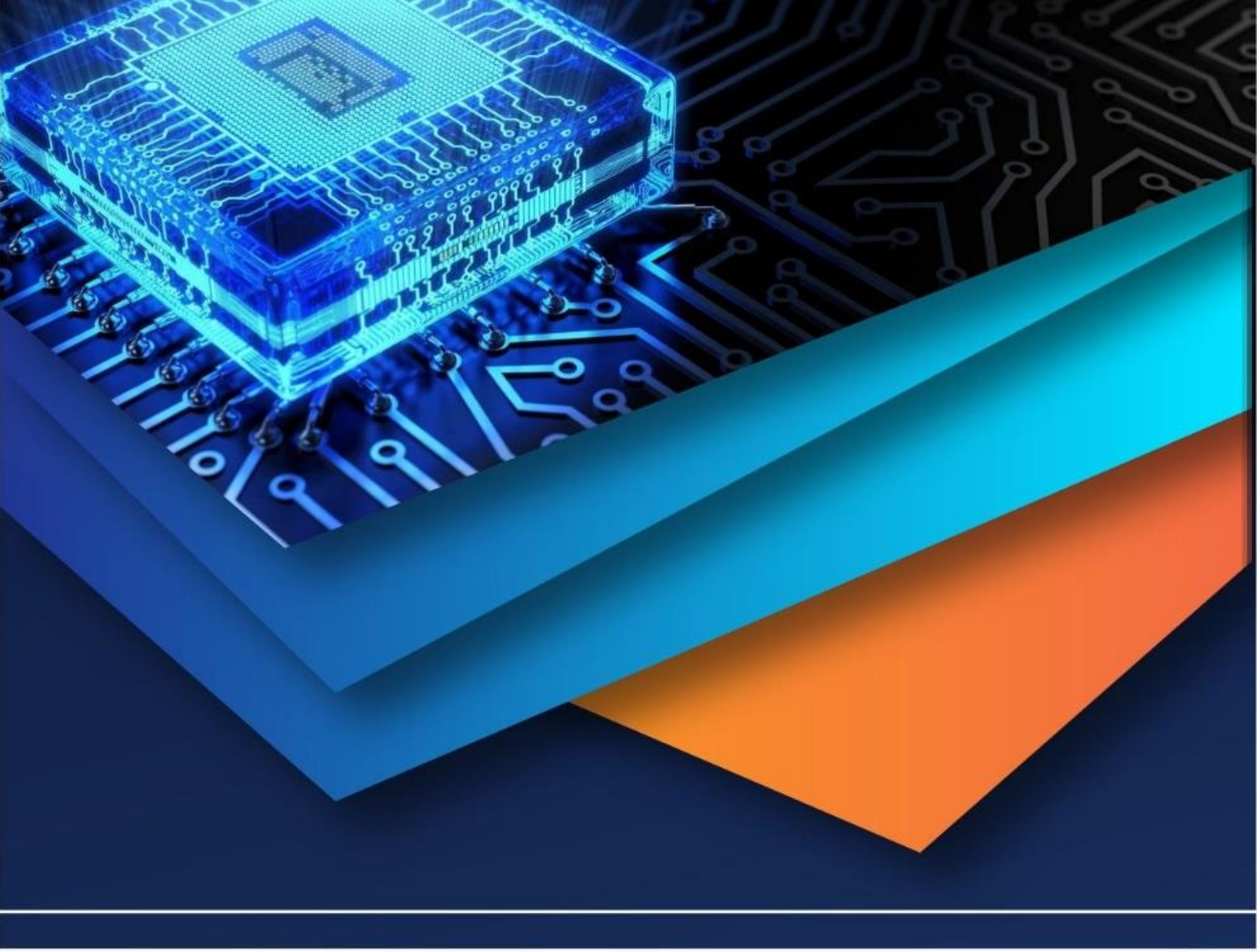
This paper examines the issue of sentiment analysis.

There are three influencing factors. The BoW, TF-IDF, and Word2Vec calculation methods were combined with a variety of classifier, including the Decision tree, Random Forest, Naive Bayes classifier, Support vector machine, , Decreased stochastic gradient, and checks, to determine emotions in the IMDB's database of movie reviews.

We believe that Word2Vec with SGD is the right mixture of emotions for the IMDB database segmentation issue based on test results. This form had the highest accuracy score of 89 percent and the highest F1 score of 83.6 percent, making it the best of all combinations. This research can be expanded to include the analysis of online feedback from social media platforms and other sophisticated algorithms.

REFRENCES

[1] NLTK: The Natural Language Toolkit. Edward Loper and Steven Bird Department of Computer and Information Science University of Pennsylvania, Philadelphia, PA 19104-6389, USA
[2] Steven Bird NLTK Documentation Release 3.2.5. (English) Sep 28, 2017
[3] Edward Loper,Steven Bird, NLTK: the natural language toolkit, Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, p.31-es, July 21-26, 2004, Barcelona, Spain
[4] Naive Bayes Scikit-Learnn.org. (English) Sep 28, 2017
[5] Cortes, Corinna;Vapnik,VladimirN.(1995)."Support-vectornetworks".Machine Learning. 20 (3): 273–297. CiteSeerX 10.1.1.15.9362
[6] Guerini Marco Lorenzo Gatti Marco Turchi Sentiment analysis: How to derive prior polarities from SentiWordN et vol. 5843 2013.
[7] Bhoir Purtata Shilpa Kolte "Sentiment analysis of movie reviews using lexicon approach" 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) pp. 1-6 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)