A Project Report Entitled

Cinematic Sentiment Analysis : Exploring Emotions in Film Reviews

Project Report submitted to

RAYAT SHIKSHAN SANSTHA's,

SADGURU GADAGE MAHARAJ COLLEGE, KARAD

(An Autonomous College)

DEPARTMENT OF STATISTICS

FOR THE PARTIAL FULFILLMENT OF THE DEGREE

MASTER OF SCIENCE

IN

STATISTICS

SUBMITTED BY,

Miss. Patil Pallavi Pandurang

M.Sc. II (Statistics)

Under the guidance of

Miss. S.S.Jagtap

2023-2024

CERTIFICATE

This is to certify that the project report entitled ‘**Cinematic Sentiment Analysis : Exploring Emotions in Film Reviews’** being Submitted by **Miss. Patil Pallavi Pandurang** as partial fulfillment for the award of degree of M.Sc. (Statistics) is a record of work carried out by her under my supervision and guidance. To the best of my knowledge the matter presented in the survey has not been submitted earlier.

Place: Karad
Date:

Miss. S.S. Jagtap
Project guide

Dr. Mrs. Patil S. P.
PG Co-ordinator
Department of Statistics,
Sadguru Gadage Maharaj College, Karad.

ACKNOWLEDGEMENT

We have taken efforts in this survey however it would not have been possible without the kind of support & help of many individuals. We would like to extend our sincere thanks to all of them. It’s our great pleasure to express our sincere thanks with deep sense to Miss. K. H. Powar for their valuable guidance & constant encouragement during the course of work & for providing us necessary facilities as well as for providing necessary information regarding the survey. Also, we would like to thank all non-teaching staff ofour department for their help & Co-operation.

I thank to all teachers for spending their valuable time. We thank all my friends and research students for their co-operation and help which we received from them.

Introduction

In the vast expanse of cinematic storytelling, emotions serve as the silent orchestrators, conducting a symphony of human experiences that unfold on the silver screen. "Cinematic Sentiment Analysis" ventures into the realm of film reviews, where audiences articulate the often ineffable, translating their visceral responses into written expressions.

Movies, as a medium, transcend language and cultural barriers, tapping into the universal language of emotions. This project is not a mere categorization of sentiments; it's a nuanced exploration seeking to decode the intricate nuances of joy, sadness, excitement, and more within the realm of cinematic engagement.

At its essence, this endeavor is a journey into the viewer's emotional landscape—a reflection on how narratives, performances, and visuals coalesce to evoke a myriad of feelings. The project leverages natural language processing (NLP) to decipher the sentiments embedded in reviews, bridging the gap between the subjective experiences of movie enthusiasts and the objective analysis of their expressions.

This exploration celebrates the unspoken language of emotions, offering a fresh perspective on the dialogue between filmmakers and their audience. It transcends the binary classifications of positive or negative, aiming to capture the diverse emotional spectrum that contributes to the rich tapestry of cinematic engagement.

As technology converges with art, "Cinematic Sentiment Analysis" stands at the crossroads, delving into the intersection of human emotions and cinematic storytelling. Beyond the surface of categorizations, this project aspires to contribute to a profound understanding of the emotional impact of cinema—a celebration of the multifaceted responses that films inspire.

Embarking on this exploration invites contemplation on the dynamic interplay between technology and emotion within the cinematic landscape. Through sentiment analysis, we embark on a journey to unveil the unspoken language of emotions, unraveling the profound connections forged between storytellers and their audience in the enchanting world of cinema.

Background

The landscape of filmmaking has undergone a profound transformation in the digital age, with advancements in technology influencing not only the production process but also the way audiences engage with cinematic narratives. In this evolving paradigm, the exploration of audience sentiments and emotional responses through film reviews has become increasingly pertinent.

Traditionally, gauging audience reactions relied on subjective assessments and anecdotal evidence. However, with the rise of digital platforms and the prevalence of user-generated content, film reviews have become a rich source of raw, unfiltered expressions of viewer emotions. This project emerges from the realization that these expressions, often buried within the textual fabric of reviews, can be systematically decoded and analyzed to provide valuable insights into the emotional impact of films.

The advent of natural language processing (NLP) and machine learning techniques has opened new avenues for extracting meaningful patterns from vast datasets of textual information. By applying sentiment analysis to film reviews, we can move beyond surface-level evaluations and delve into the nuanced emotional nuances that shape audience perceptions.

Furthermore, the project acknowledges the dynamic nature of cinematic storytelling. Films, as an art form, have the power to evoke a myriad of emotions, and understanding how these emotions are articulated by audiences contributes to a holistic appreciation of the cinematic experience. By contextualizing sentiment analysis within the realm of film reviews, this exploration seeks to bridge the gap between the subjective nature of viewer responses and the objective insights derived from data-driven analysis.

Significance of the Dataset

"Deciphering the intricate tapestry of audience sentiments towards cinematic productions is a complex undertaking in the landscape of entertainment analysis. Traditional analytical methods often struggle to capture the richness and diversity of individual experiences expressed in movie reviews. The dataset harnessed for this sentiment analysis project emerges as a distinctive gateway, inviting us to unravel the subtle nuances embedded within the natural language expressions found in authentic movie critiques.

This dataset's significance transcends the constraints of conventional research approaches, offering a more comprehensive comprehension of audience reactions to films. By immersing ourselves in the authentic language wielded by moviegoers and discerning the myriad sentiments encapsulated in their reviews, the dataset emerges as a trove of insights for constructing a robust sentiment classification system. This system, poised to uncover patterns, emotions, and nuanced indicators, stands poised to reveal layers of audience reactions that conventional research methodologies may inadvertently overlook.

In essence, this dataset stands as a valuable repository of genuine audience sentiments, providing an authentic lens through which filmmakers, analysts, and researchers can delve into the effectiveness, challenges, and emotional dimensions associated with specific cinematic experiences. The authenticity and depth embedded in this dataset significantly contribute to advancing our understanding of audience perspectives, elevating the discourse on films and their impact on diverse cinematic genres."

Data Description

The dataset utilized in this study has been curated from Kaggle, a prominent platform for data science and machine learning competitions. Specifically, the dataset is sourced from the IMDB dataset, a comprehensive collection tailored for natural language processing and text analytics. Below is an overview of the key characteristics and structure of the dataset:

- Dataset Origin:Kaggle
- Source Dataset: IMDB dataset
- Dataset Size: 50,000 movie reviews

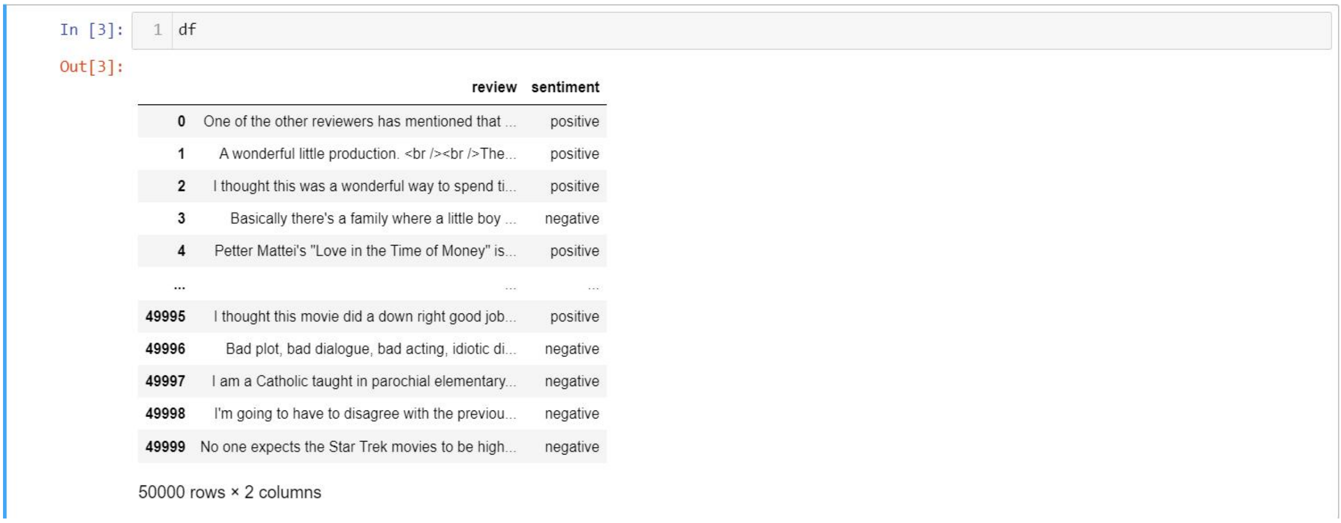
Columns:

1. Review (Text Data):

i. The dataset is primarily composed of textual data, representing movie reviews.
ii. Each entry in this column encapsulates the expressive and subjective responses of viewers to the respective movies.

2. Sentiment (Binary Classification):

a) The dataset is designed for binary sentiment classification.
b) Each review is associated with a sentiment label indicating whether it is categorized as positive or negative.
c) There are 25,000 Positive and 25,000 Negative reviews.
d) This binary sentiment classification serves as the target variable for predictive modeling.



```
In [3]:  df
Out[3]:
```

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. <br /><br />The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive
...	...	...
49995	I thought this movie did a down right good job...	positive
49996	Bad plot, bad dialogue, bad acting, idiotic di...	negative
49997	I am a Catholic taught in parochial elementary...	negative
49998	I'm going to have to disagree with the previou...	negative
49999	No one expects the Star Trek movies to be high...	negative

50000 rows × 2 columns

Scope Of The Project

The "Cinematic Sentiment Analysis: Exploring Emotions in Film Reviews" project harnesses advanced Natural Language Processing (NLP) techniques to dissect sentiments within a vast dataset comprising 50,000 movie reviews from IMDB. The primary focus is on constructing a robust sentiment classification system, extracting nuanced patterns and emotions from the authentic language used by viewers.

Diverging from traditional methods, this initiative acknowledges the multifaceted nature of cinematic experiences, aiming to capture the richness and diversity of sentiments expressed in unfiltered moviegoer reviews. Through the lens of NLP, the project aspires to distill valuable insights into the emotional dimensions associated with positive and negative sentiments, contributing to a more holistic comprehension of audience perspectives.

In essence, the project extends beyond conventional analyses, offering a comprehensive exploration of cinematic sentiments through the nuanced expressions of natural language. It seeks to enrich our understanding of audience reactions, providing valuable insights that transcend standard research methodologies and contribute to the broader discourse on filmmaking and its impact on diverse audiences.

Methodology

- . **Data Collection:**

Assemble a comprehensive dataset, emphasizing key columns like "Review," to capture diverse sentiments expressed in movie reviews.

- . **Text Preprocessing:**

- Conduct thorough text cleansing to eliminate noise and standardize formats, optimizing the quality of textual data for subsequent analysis.

- **Bag of Words (BoW)**

- Implement the Bag of Words model to represent movie reviews as a matrix of word frequencies. This technique simplifies the complex text into a numerical format for further analysis.

- **TF-IDF (Term Frequency-Inverse Document Frequency)**

- Utilize TF-IDF to weigh the importance of words in movie reviews. This method captures the significance of terms within the entire corpus, providing a nuanced representation of the textual content.

- **Word2Vec**

Apply Word2Vec embedding techniques to represent words as vectors in a continuous vector space. This approach captures semantic relationships between words, enhancing the model's understanding of context.

- **Data Visualization:**

- Employ visualizations to present the outcomes of each technique, facilitating a comparative analysis. Visual representations contribute to a more intuitive understanding of sentiment patterns.

- **Classification Model Development:**

- Implement machine learning algorithms, such as Decision Trees or XGBoost, on Bag of Words, TF-IDF, and Word2Vec representations to develop robust sentiment classification models.

- **Performance Evaluation:**

- Assess the performance of each model using metrics like precision, recall, and F1 score. This ensures the reliability of sentiment predictions across different techniques.

- **Iterative Refinement:**

- Iteratively refine models based on feedback and insights gained from each technique. Optimize the performance of Bag of Words, TF-IDF, and Word2Vec models over successive iterations.

- **Iterative Analysis:**

- Conduct iterative reviews and enhancements to the sentiment analysis methodology. Adapt models to evolving patterns and improve accuracy continuously throughout the project lifecycle.

This methodology leverages the strengths of Bag of Words, TF-IDF, and Word2Vec to comprehensively analyze sentiments in movie reviews, providing a nuanced understanding of the textual content.

Objectives

- The project aims to predict whether a given movie review expresses a positive or negative sentiment using machine learning.
- to determine the optimal feature selection method among Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and Word2Vec to enhance the accuracy of sentiment predictions.

Data Preprocessing

- **Lowercasing:**

Convert all text to lowercase to ensure consistency, as NLP models treat "word" and "Word" as different tokens.

- **Tokenization:**

Break the text into individual words, phrases, or tokens. Tokenization is essential for further analysis, as it segments the text into meaningful units.

- **Removing Punctuation:**

Remove punctuation marks (e.g., periods, commas, question marks) from the text as they often don't carry significant meaning in many NLP tasks.

- **Removing Numbers:**

Depending on the task, you may choose to remove or retain numerical values. For some tasks, like sentiment analysis, numbers might not be relevant and can be removed.

- **Stopword Removal:**

Stopwords are common words (e.g., "and," "the," "in") that occur frequently in the language but may not carry important meanings for certain tasks. Removing stopwords can reduce noise in the data.

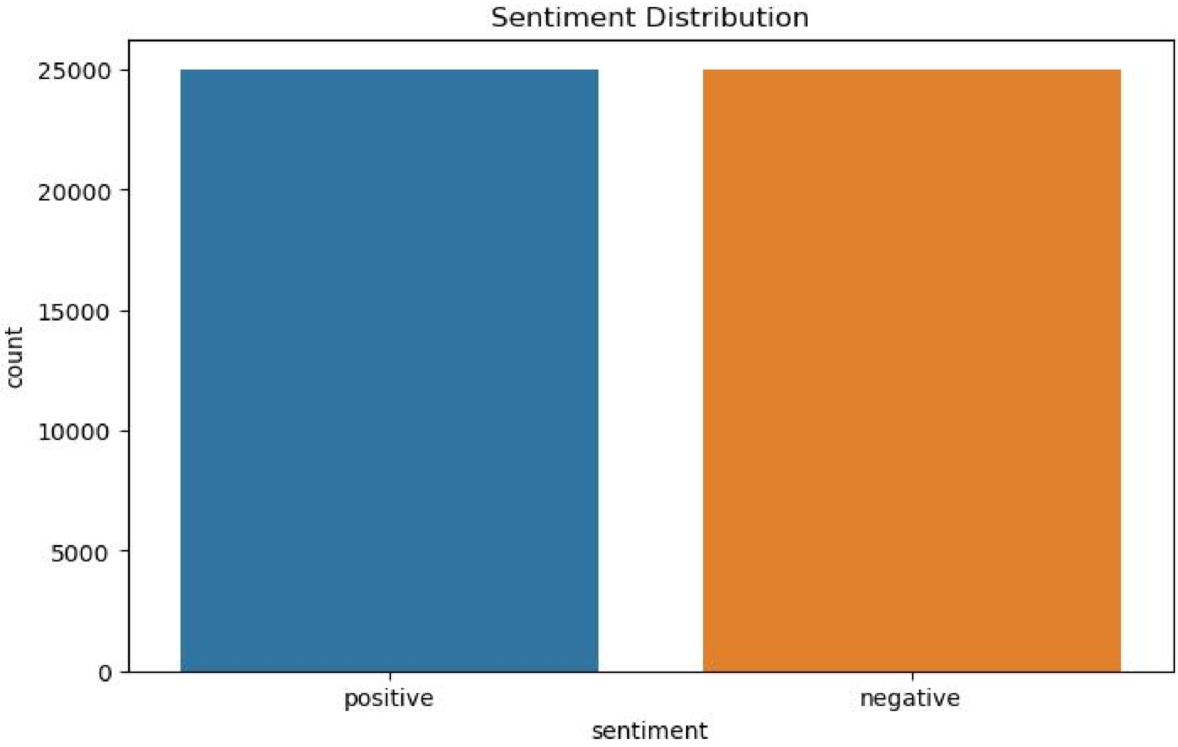
- **Stemming and Lemmatization:**

Stemming reduces words to their root form by removing suffixes (e.g., "running" to "run"). Lemmatization also reduces words to their base or dictionary form (e.g., "better" to "good"). These techniques help standardize word forms.

Exploratory Data Analysis (EDA)

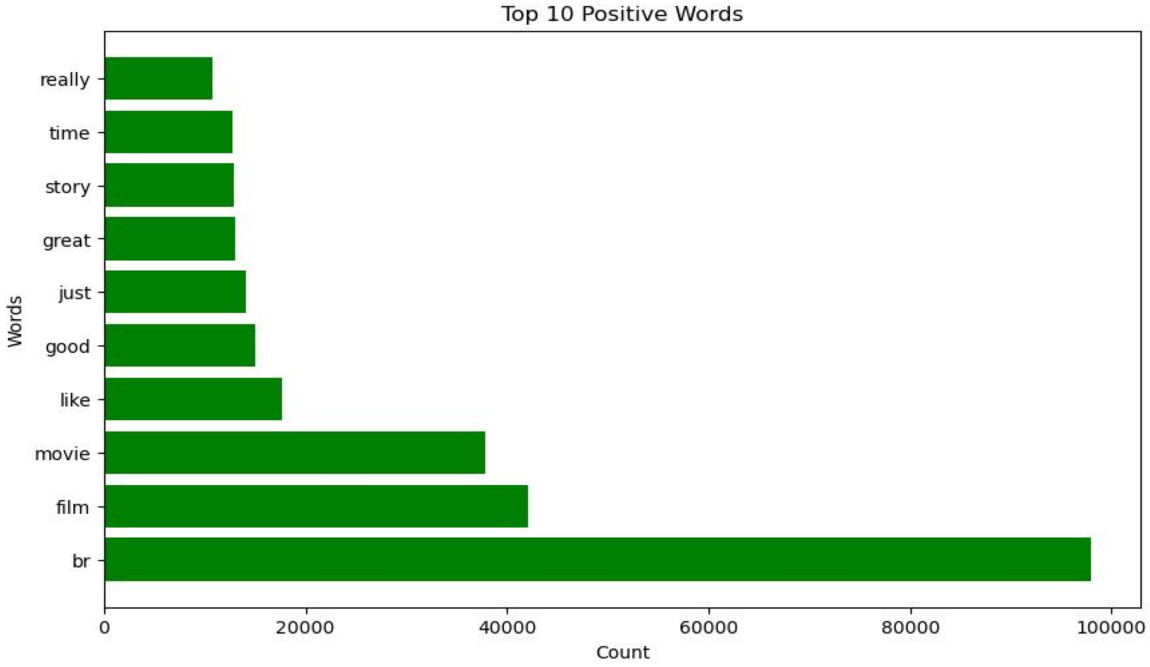
Exploratory Data Analysis (EDA) is a crucial step in understanding the characteristics and patterns within IMDB movie review dataset.

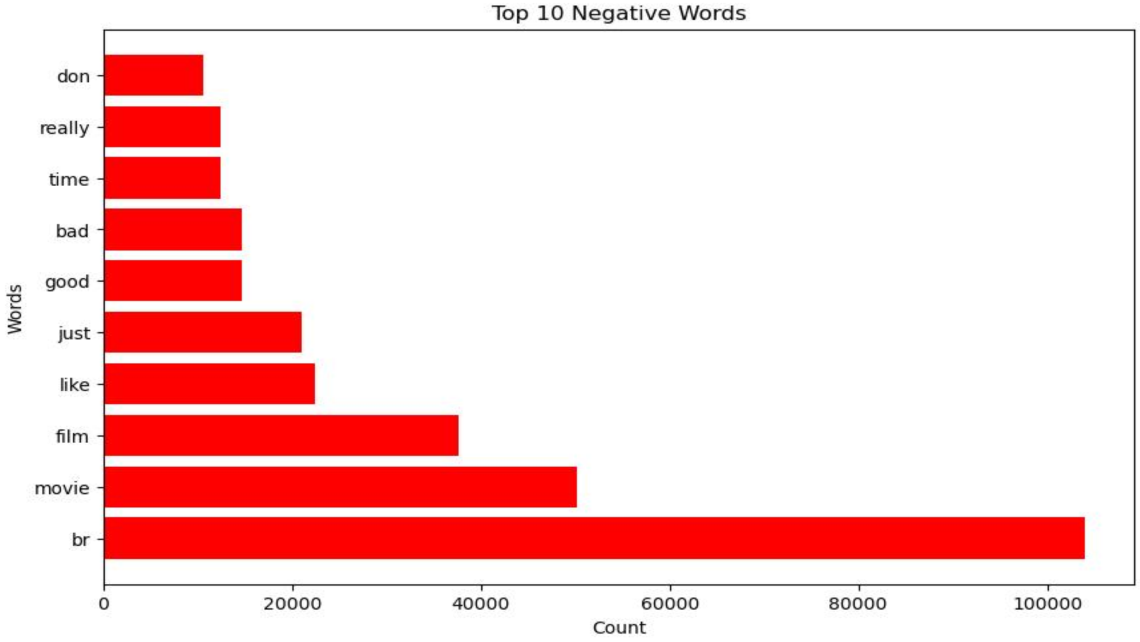
1) Distribution of Sentiment:



The conclusion from this plot is that the dataset is balanced in terms of sentiment labels. It has an equal number of positive and negative reviews. This balance is important to ensure that a machine learning model trained on this dataset does not have a bias towards any particular sentiment.

2)Sentiment Analysis through Top Words Exploration

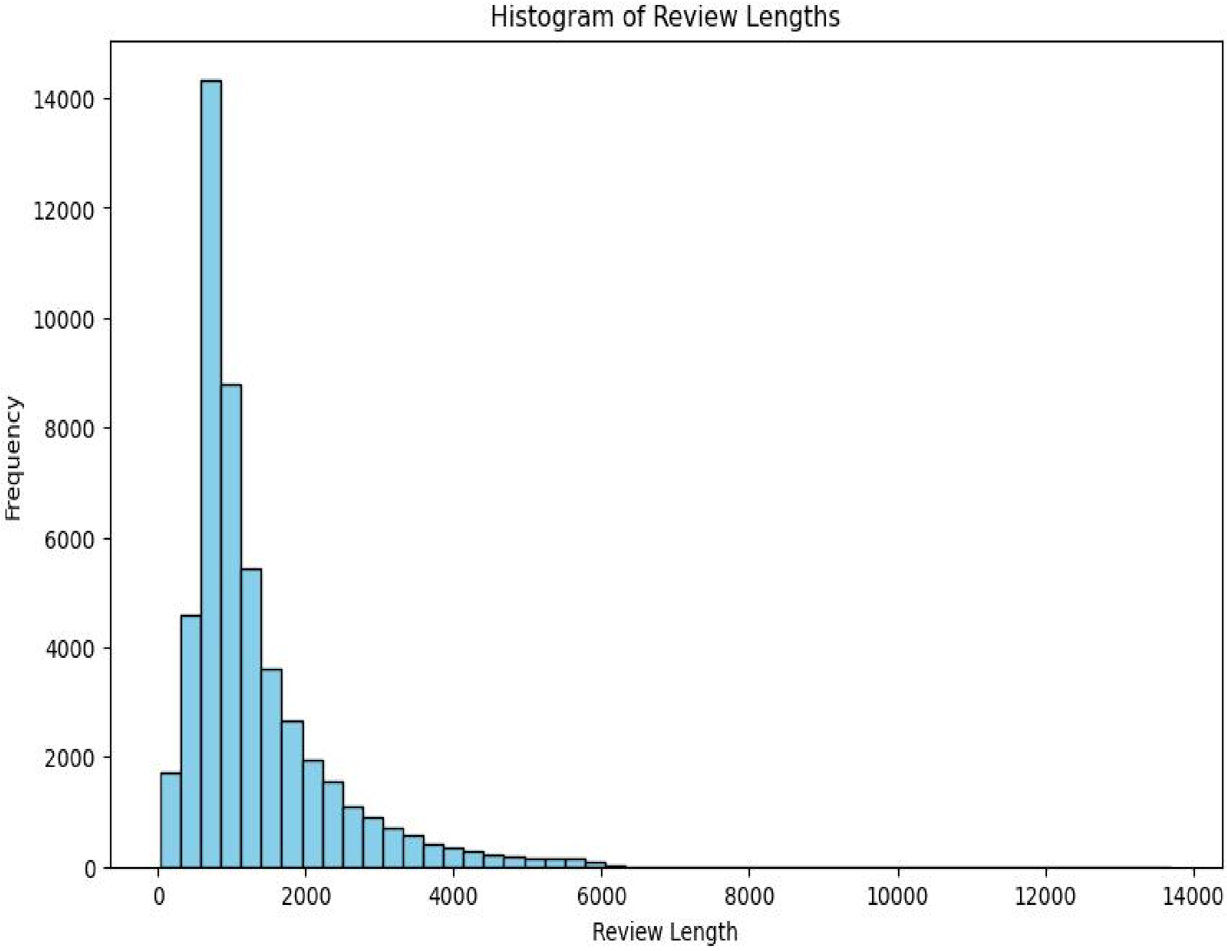




Observations:

- The word "br" appears in both positive and negative words, potentially due to HTML tags or formatting in the text data.
- Positive words include terms such as "good," "great," and "really," indicating favorable opinions.
- Negative words include terms like "bad" and "don," suggesting criticism or negative sentiments.
- Context about the dataset and specific analysis would provide a deeper understanding of the sentiment expressed in the text.

3)Review Length Exploration:

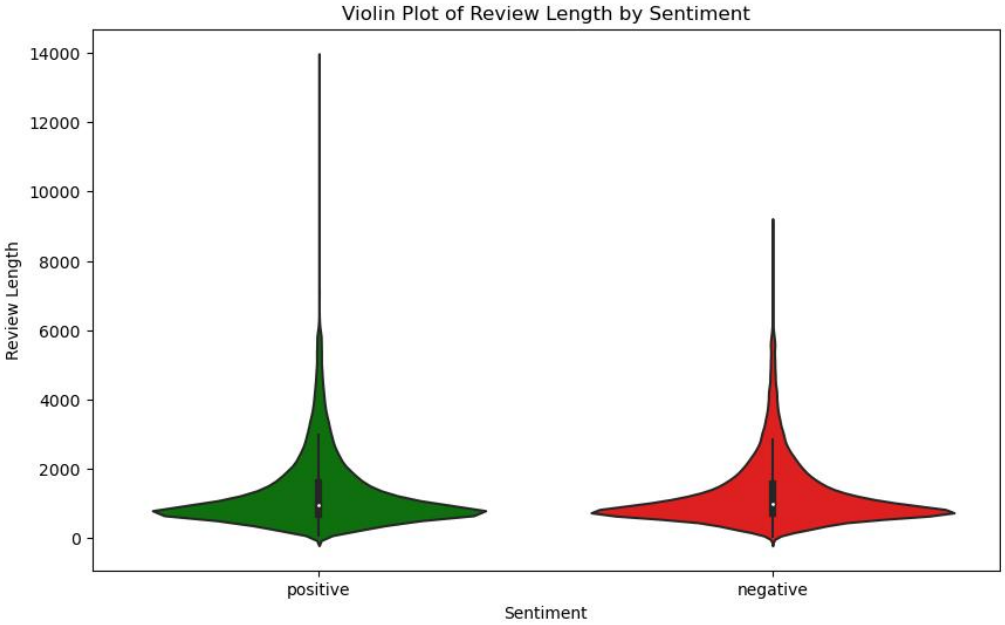


Observations:

- Reviews exhibit a right-skewed distribution.
- Majority of reviews are concise, with lengths ranging from 32 to 13,704 characters.
- Average review length is approximately 1,309 characters.

4)Violin Plot for Review Length by Sentiment:

Violin Plot:
A violin plot is a statistical visualization merging box plots and kernel density plots. It illustrates the distribution of a numeric variable across categories, depicting probability density via the width of the "violin." Wider sections denote higher data density, providing insights into the variable's distribution.



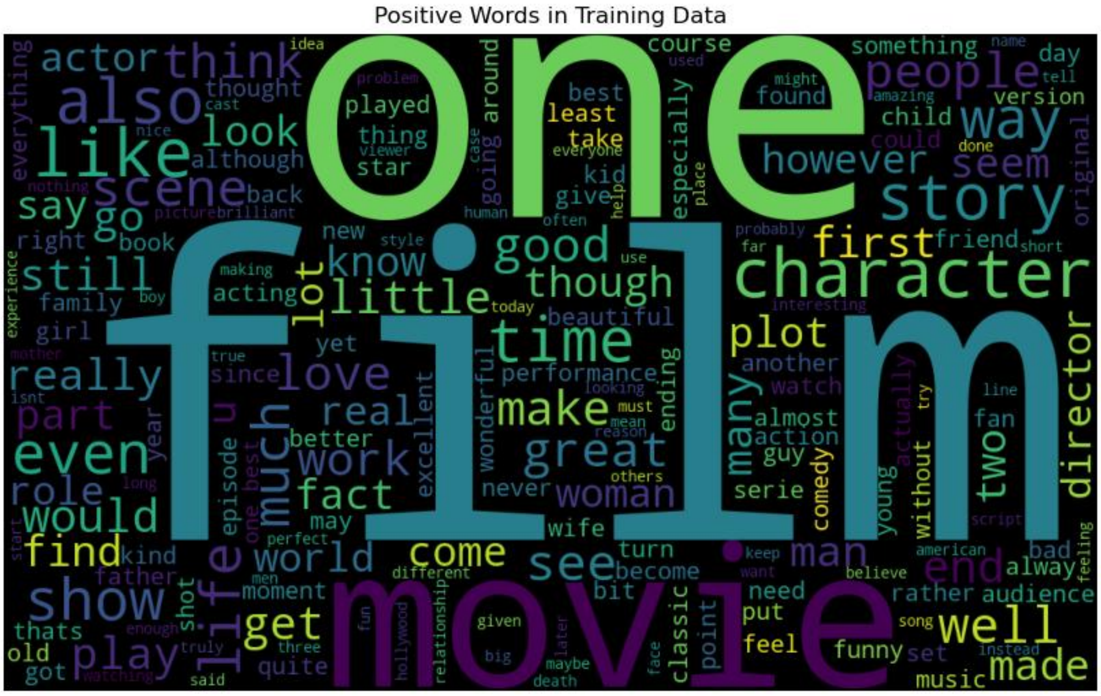
Conclusion :

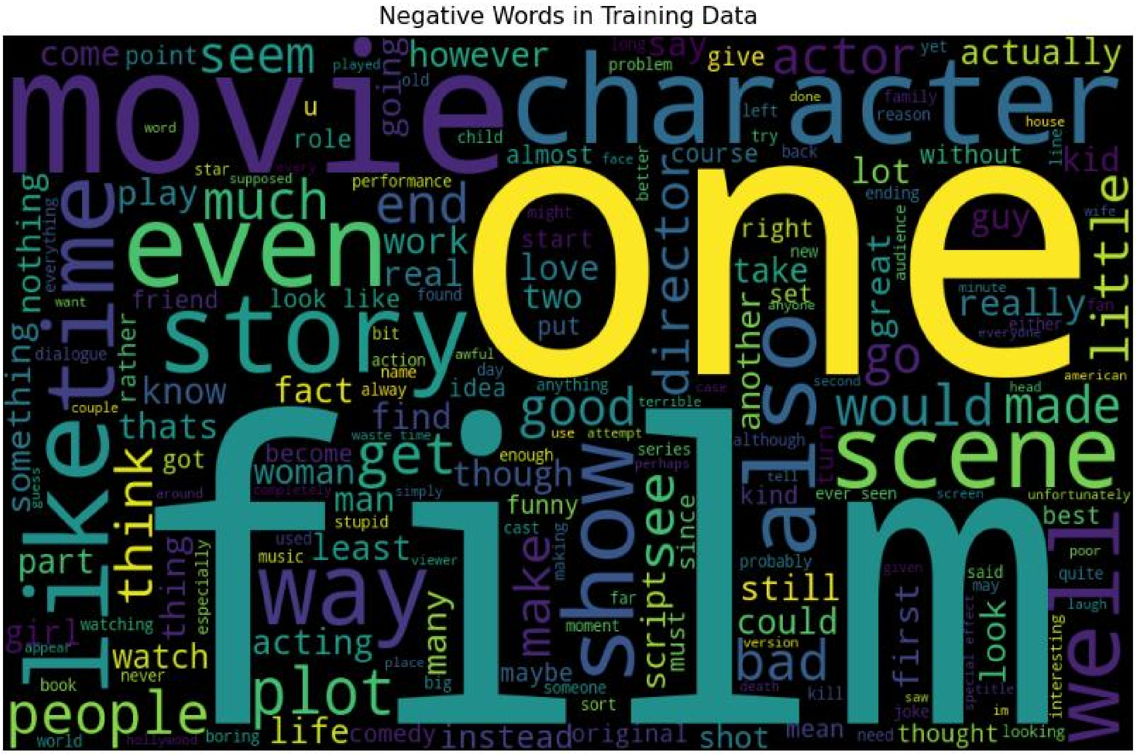
- **Distribution of Review Length**: The distribution of review lengths for both positive and negative sentiments appears to have a similar shape, with a single peak and a long tail extending towards higher review lengths. This suggests that while most reviews are relatively short, there are some with much longer lengths.
- **Median Review Length**: The median review length (indicated by the white dot within the thick black bar in the centre of each violin) appears to be slightly higher for negative reviews than for positive ones. This could suggest that when people have negative sentiments, they might provide more detailed feedback..
- **Extreme Values**: The presence of extreme values or outliers is more pronounced in the positive sentiment category, where the tails of the distribution stretch much longer than those of the negative sentiment. This suggests that some positive reviews are exceptionally lengthy.

From this plot, one might infer that people tend to write more and provide more extensive feedback when they are expressing negative sentiments compared to when they are expressing positive sentiments.

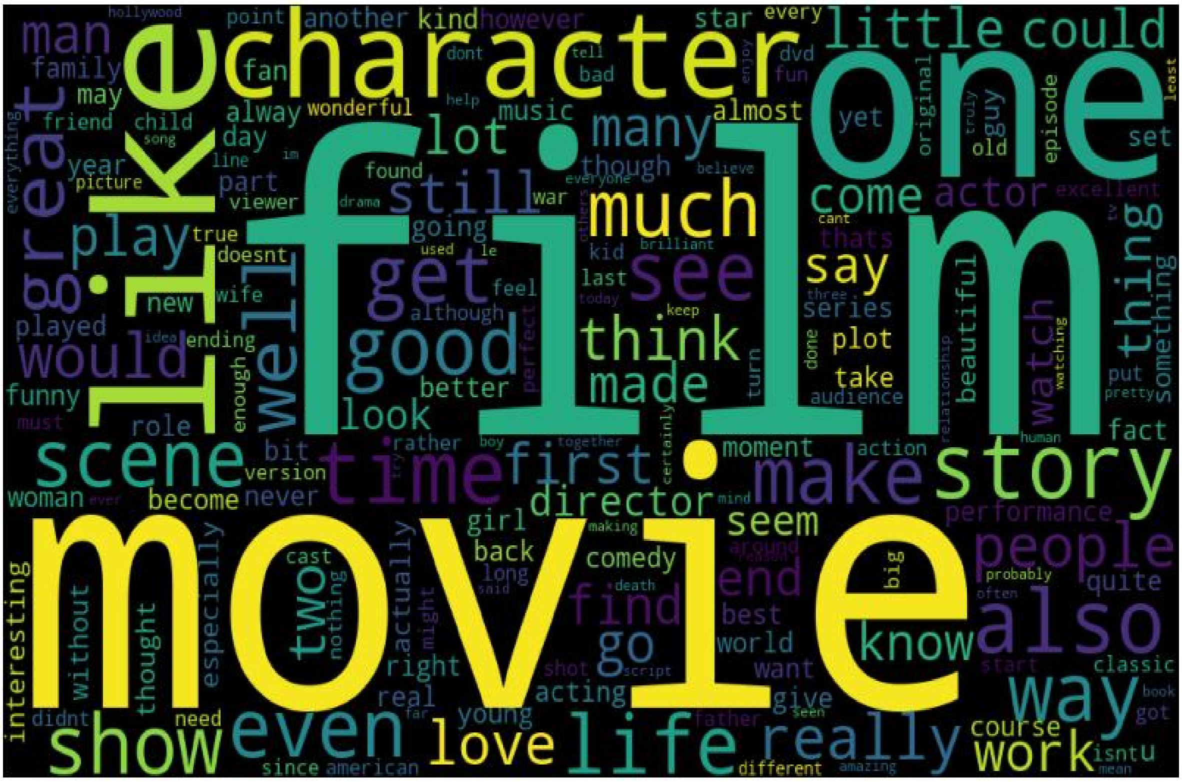
2) Word Cloud :

Word clouds or tag clouds are graphical representations of word frequency that give greater prominence to words that appear more frequently in a source text. The larger the word in the visual the more common the word was in the document.

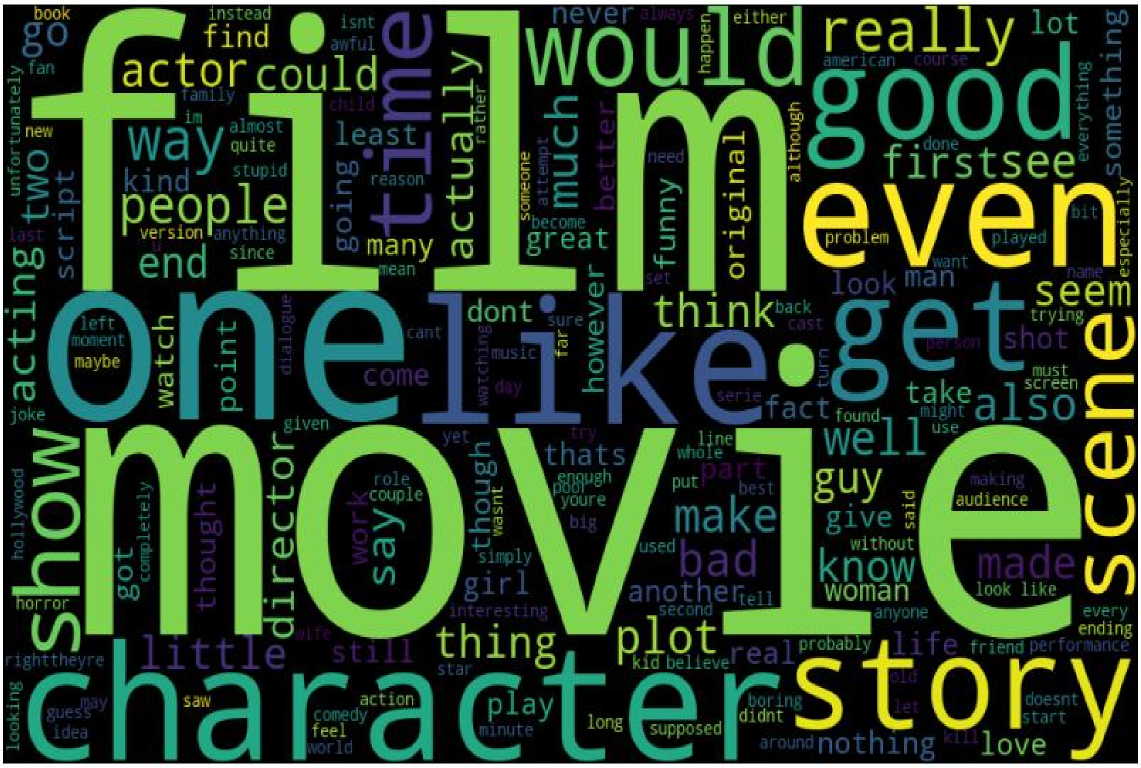

Positive Words in Training Data


Negative Words in Training Data

Positive Words in Test Data



Negative Words in Test Data



➢ Bag of Words (BoW):

The Bag of Words (BoW) model is a pivotal concept in natural language processing (NLP) and machine learning, offering a simplified yet effective approach to transform textual data into a format suitable for computational analysis. BoW treats each document as an unordered collection of words, discarding grammar and word order, and representing text through the frequency distribution of its constituent terms.

Essential Components of Bag of Words:

1. **Tokenization:**

The process initiates with tokenization, breaking down each document into individual words or tokens. Punctuation is typically removed, and the text is converted to lowercase to ensure uniformity. The result is a list of words for each document.
Let's consider two simple documents to showcase the BoW process:

- Document 1: "The cat sat on the mat."
- Document 2: "The dog barked loudly."
-

Document 1 tokens: ["the", "cat", "sat", "on", "the", "mat"]
Document 2 tokens: ["the", "dog", "barked", "loudly"]

2. Vocabulary Creation:

A collective vocabulary is formed by amalgamating all unique words across the entire corpus of documents. Each distinct word in the vocabulary is assigned a unique index, forming the basis for subsequent vectorization.

- Vocabulary: ["the", "cat", "sat", "on", "mat", "dog", "barked", "loudly"]

3. Vectorization:

Vectorization is the crux of BoW, where each document is represented as a numerical vector. The length of the vector corresponds to the size of the vocabulary, and each element denotes the frequency of a specific word in the document. The order of words is disregarded, emphasizing the frequency distribution.

Vectorization:

- Document 1 vector: [2, 1, 1, 1, 1, 0, 0, 0]

The word "the" appears twice, "cat" once, and so forth.

- Document 2 vector: [1, 0, 0, 0, 0, 1, 1, 1]

"The" appears once, "dog" once, "barked" once, and "loudly" once.

- **TF-IDF (Term Frequency-Inverse Document Frequency):**

TF-IDF, like the Bag of Words (BoW), plays a crucial role in natural language processing (NLP) and machine learning, providing a sophisticated method for text representation. It addresses the limitations of BoW by considering not only term frequency but also the significance of terms across a document corpus.

Essential Components of TF-IDF:

1. Tokenization

Similar to BoW, TF-IDF begins with tokenization, breaking down each document into individual words or tokens. Punctuation is removed, and uniformity is maintained by converting text to lowercase.

- Document 1 tokens: ["the", "cat", "sat", "on", "the", "mat"]
- Document 2 tokens: ["the", "dog", "barked", "loudly"]

2. TF (Term Frequency):

Term frequency measures how often a term appears in a document. It is calculated by dividing the number of occurrences of a term by the total number of terms in the document.

Example:

- TF for Document 1: ["the": 2/6, "cat": 1/6, ...]

3.IDF (Inverse Document Frequency):

IDF evaluates the significance of a term across the entire corpus. It is computed as the logarithm of the total number of documents divided by the number of documents containing the term.

- IDF for "the": log(2/2) = 0
- IDF for "cat": log(2/1) = 0.301

4. TF-IDF Calculation:

5. TF-IDF is obtained by multiplying the TF and IDF values for each term in a document.

- TF-IDF for Document 1: ["the": 0, "cat": 0.301, ...]

Significance of TF-IDF:

- TF-IDF, unlike BoW, considers the importance of terms in the entire corpus.
- It mitigates the impact of frequently occurring terms and highlights those with significance across documents.
- TF-IDF is pivotal in tasks like text classification, sentiment analysis, and information retrieval, offering nuanced representations despite the inherent complexities of language.

- **Word2Vec:**

Word2Vec, a transformative model in natural language processing (NLP) and machine learning, surpasses the limitations of traditional approaches like Bag of Words (BoW) by capturing semantic relationships between words. Unlike BoW's focus on frequency, Word2Vec embeds words into continuous vector spaces, enabling a more nuanced understanding of language.

Essential Components of Word2Vec:

1. Training on Large Corpora:

Word2Vec relies on extensive training on large text corpora to learn word embeddings. This involves predicting the context of a word within sentences.

2. Continuous Vector Spaces:

Words are represented as vectors in continuous spaces, preserving semantic similarities. The distance and direction between vectors reflect the relationships between words.

Example:
Vector("king") - Vector("man") + Vector("woman") ≈ Vector("queen")

Significance of Word2Vec:
Semantic Understanding:
- Word2Vec excels in capturing semantic relationships between words, allowing it to understand context and meaning more effectively.

Contextual Similarities:
- The model can identify similarities between words based on context, even if they don't share identical terms.

Nuanced Representations:
- Unlike BoW's simplistic frequency-based approach, Word2Vec provides nuanced representations that align with the intricate structure of language.

Applications in NLP:
- Word2Vec's embeddings are invaluable in various NLP applications, including sentiment analysis, document clustering, and information retrieval.

Mitigating Semantic Gaps:
- It mitigates the semantic gap present in BoW, where similar words may have disparate vector representations.

In summary, Word2Vec stands as a sophisticated advancement in text representation, offering a dynamic understanding of language that extends beyond the confines of traditional methods. Its ability to capture semantic nuances positions it as a pivotal tool in modern NLP and machine learning applications.

Model Building:

Once we have completed the Exploratory Data Analysis (EDA) and Feature Engineering stages for our sentiment analysis project on IMDB movie reviews, the subsequent step involves fitting the data, utilizing various feature selection methods, to different machine learning algorithms. Our ultimate goal is to discern the sentiment expressed in the reviews, determining whether they convey a positive or negative sentiment. For this purpose, we will employ three distinct feature selection techniques: Bag of Words (BoW), TF-IDF, and Word2Vec.

•

i) Decision Tree

ii) Random Forest

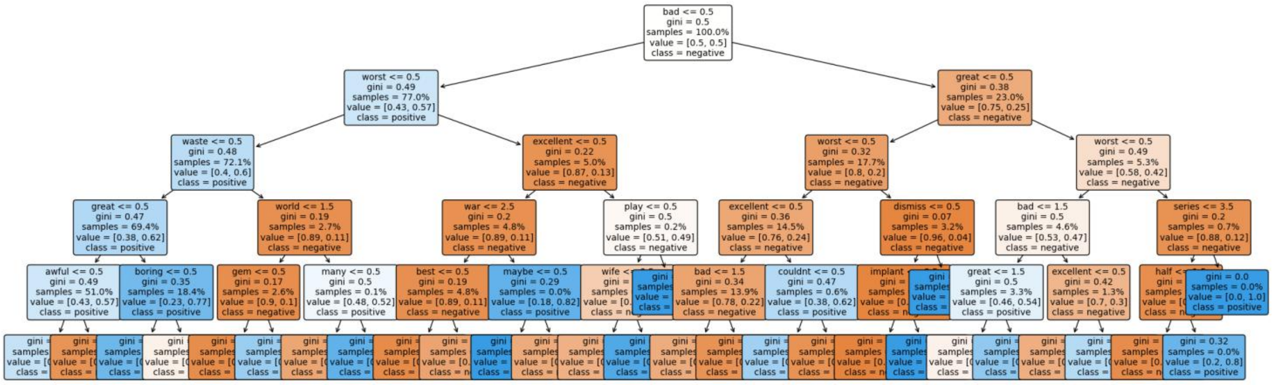
iii) XGBoost Classifier

iv) Logistic Regression

i. Decision Tree:

A Decision Tree is a supervised machine learning algorithm designed for both classification and regression tasks. Its structure resembles an inverted tree where each internal node represents a decision based on a feature, leading to subsequent branches representing possible outcomes. The leaves of the tree signify the final predicted class or value. Decision Trees are versatile and intuitive, making them suitable for tasks ranging from identifying spam emails and predicting customer churn to diagnosing medical conditions like diabetes and determining the likelihood of heart failure. The algorithm excels in handling both categorical and numerical data, providing interpretable results and facilitating decision-making processes in diverse domains.

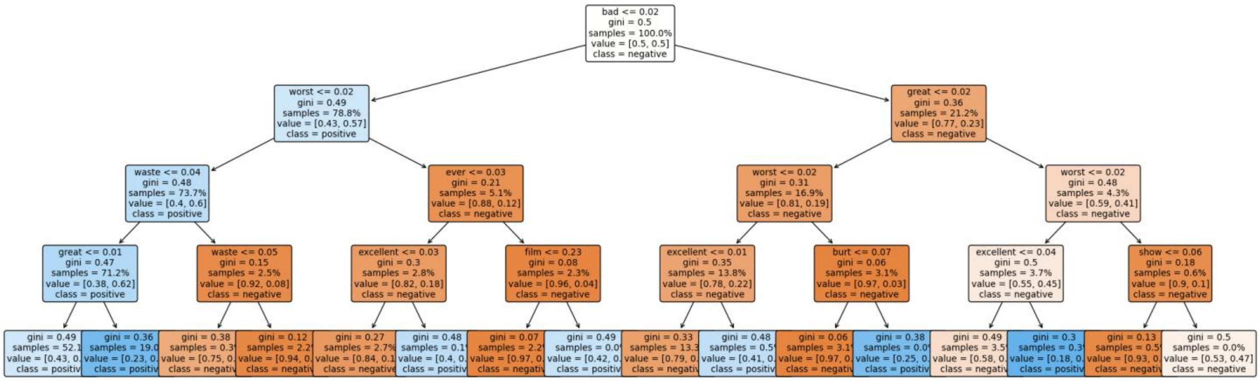
a. Decision tree structure for BOW :-



Interpretation:-

From the above Decision tree we observed that the word ‘bad’ is selected as Root Node. The criteria used here is Gini.

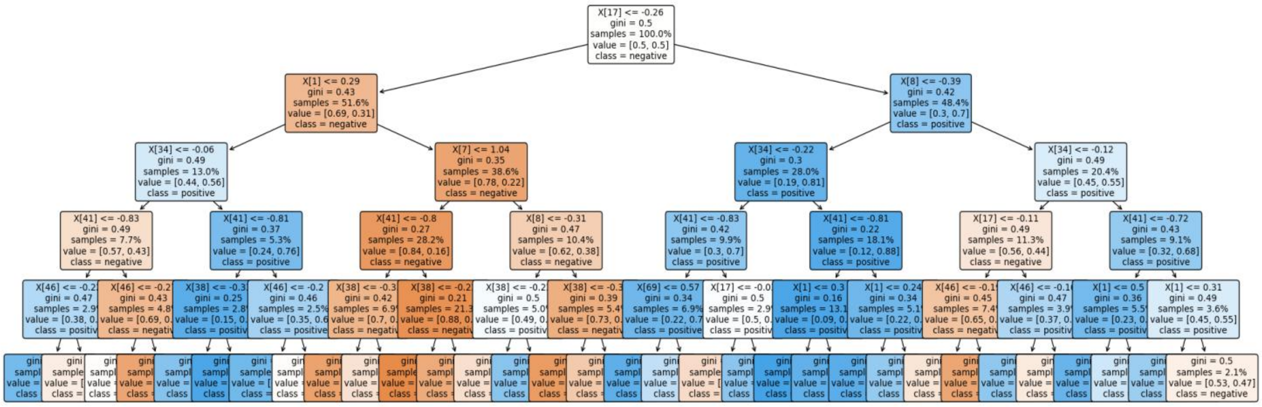
b. Decision tree structure for tf-idf :-



Interpretation:-

From the above Decision tree we observed that the word ‘bad’ is selected as Root Node. The criteria used here is Gini.

c. Decision tree structure for word2vec :-

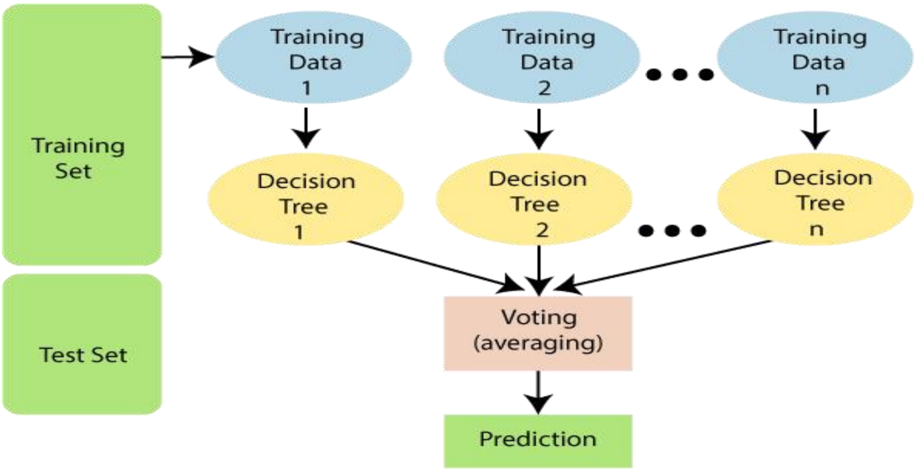


Interpretation:-

1. The term "root node" in this context refers to the representation of data, denoted as `x[34]`, within the Word2Vec model.
2. This designation reflects the model's nuanced sentiment analysis, discerning subtle contextual nuances, and stands in contrast to the more rigid representations of Bag of Words (BoW) or Term Frequency-Inverse Document Frequency (TF-IDF) methods.

ii. Random Forest:

Random Forest is an ensemble learning algorithm that operates by constructing a multitude of decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. It leverages the strength of multiple decision trees to enhance predictive accuracy and robustness. Each tree in the forest is built on a random subset of the dataset, and during prediction, the ensemble's combined wisdom mitigates overfitting and contributes to a more reliable outcome. Random Forest finds applications in various domains, from finance and marketing to healthcare. Its ability to handle large datasets, accommodate both categorical and numerical features, and provide feature importance rankings makes it a go-to choice for complex classification and regression tasks, such as credit scoring, stock price prediction, and disease diagnosis.



ii. XGBoost

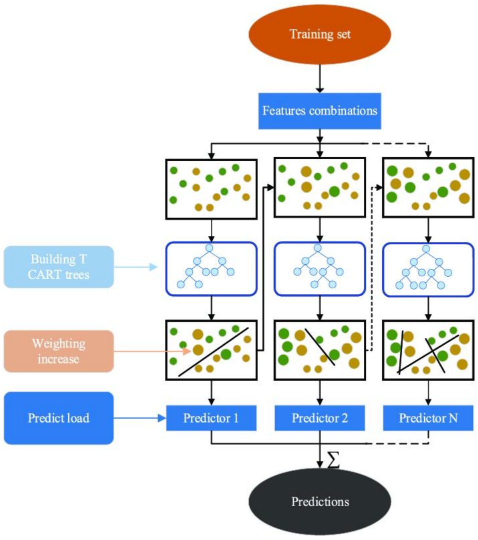
XGBoost, short for eXtreme Gradient Boosting, stands out as a powerful and efficient machine learning algorithm that belongs to the gradient boosting framework. It excels in predictive modeling for classification and regression tasks, leveraging the strengths of decision trees. XGBoost iteratively builds a series of weak learners, typically decision trees, and combines their predictions to enhance overall accuracy.

Key features of XGBoost include its ability to handle missing values, regularization techniques to prevent overfitting, and flexibility in accommodating various data types. The algorithm is known for its speed and scalability, making it suitable for large datasets and complex problems.

XGBoost finds widespread use in diverse applications, including click-through rate prediction, fraud detection, and healthcare analytics.

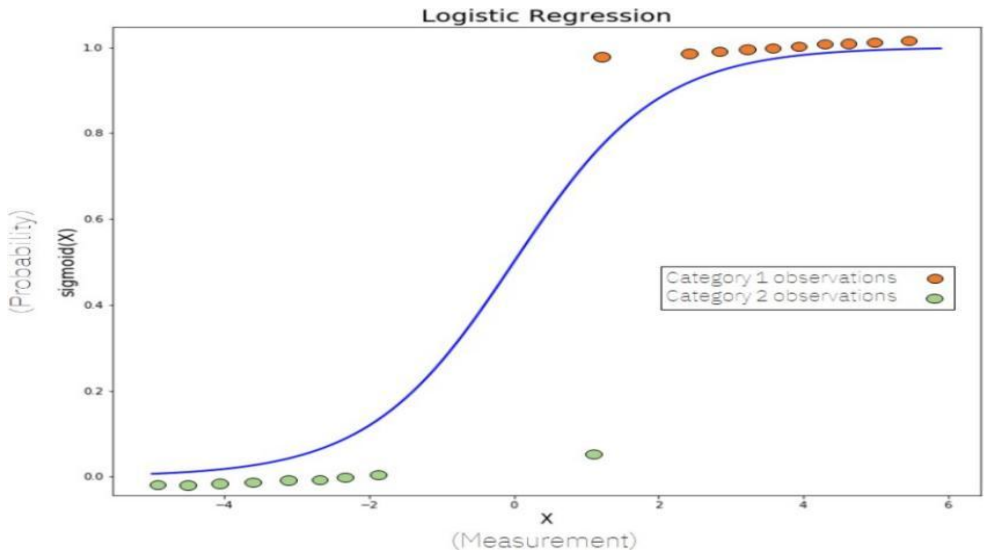
Its popularity stems from its effectiveness in achieving high predictive accuracy, handling non-linearity in data, and providing insights into feature importance.

In summary, XGBoost has become a go-to algorithm in the machine learning community, delivering robust performance across a spectrum of tasks and earning its place as a valuable tool for practitioners in both research and industry.



iii. Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable isdichotomous, which means there would be only two possible classes.
In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).



Mathematically, a logistic regression model predicts P(Y=1) as a function of X.It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection, heart failure prediction etc.

Result:-

Precision measures the accuracy of the model in correctly identifying negative reviews among all instances predicted as negative. It is calculated using the formula:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positives}}$$

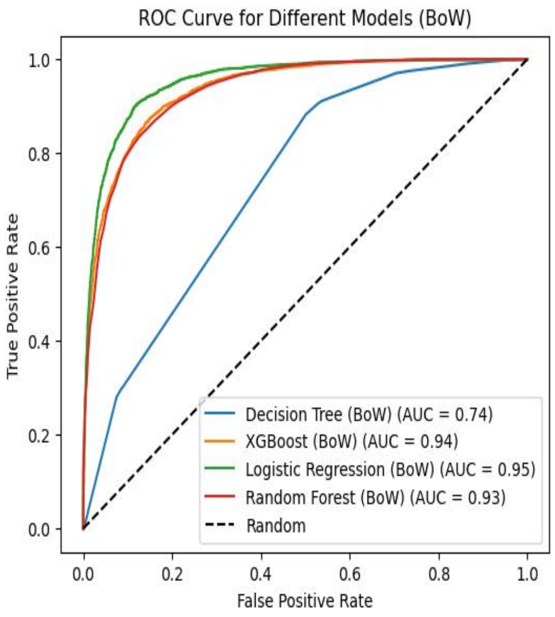
A higher precision score indicates that when the model predicts a review as negative, it is more likely to be correct. This metric is crucial for projects that prioritize accurately predicting negatives as negatives and minimizing the risk of falsely identifying positive reviews as negative.

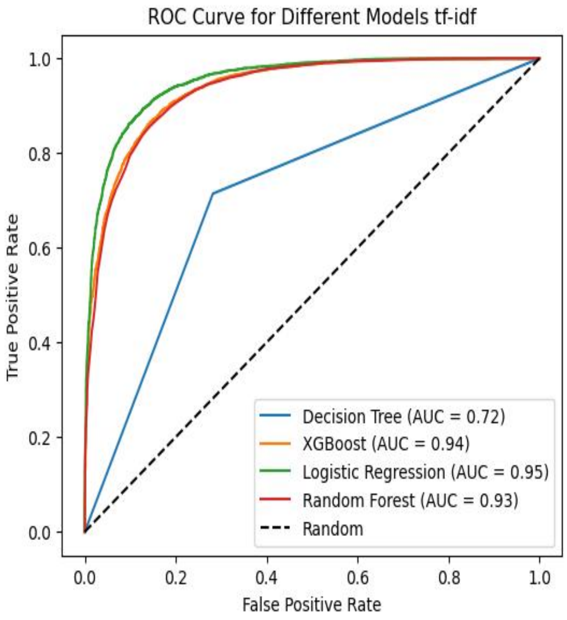
The Precision obtained for each model is given by,

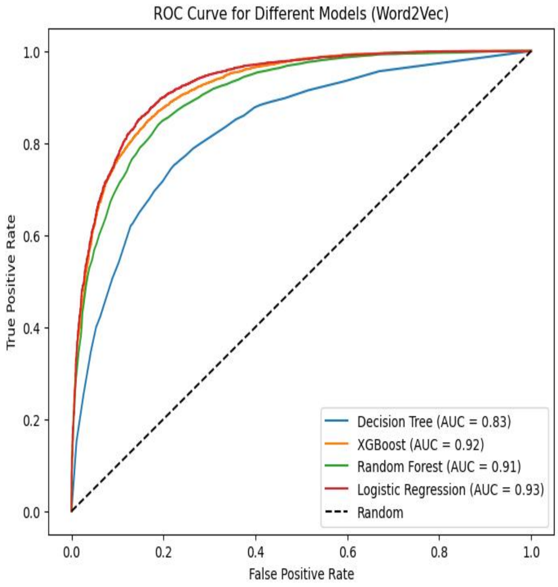
Sr. No.	Classifier	Precision		
		BOW	TF-IDF	Word2Vec
1.	Decision Tree	0.72	0.71	0.72
2.	Random Forest	0.85	0.85	0.83
3.	XGBoost	0.88	0.87	0.85
4.	Logistic Regression	0.91	0.92	0.86

Logistic Regression paired with TF-IDF achieves a precision of 92% for predicting negative reviews.To choose the best model we will take the help of AUC and ROC curves.Below tables shows the AUC score obtained for each model.

Sr. No.	Classifier	AUC		
		BOW	TF-IDF	Word2Vec
1.	Decision Tree	0.74	0.72	0.83
2.	Random Forest	0.93	0.93	0.91
3.	XGBoost	0.94	0.94	0.92
4.	Logistic Regression	0.95	0.95	0.93







Now, by looking at AUC scores, we can conclude that **Logistic Regression with TF-IDF** is the optimal choice. Consistently achieving AUC values of , this combination excels in capturing important patterns, with TF-IDF specifically emphasizing crucial words for improved precision.

Conclusion and Future Scope:-

"Cinematic Sentiment Analysis" delves into the intricacies of film reviews, unraveling the emotional tapestry woven by audiences through NLP techniques like Bag of Words, TF-IDF, and Word2Vec. Opting for Logistic Regression with TF-IDF, the project achieved a remarkable 92% precision in identifying negative sentiments. The exploratory analysis provided profound insights into sentiment distributions and review patterns, enriching our comprehension of audience responses. Moving forward, the project could explore deeper avenues, incorporating deep learning, genre-specific analysis, and user-generated content integration. Continuous strides in feature engineering and semantic understanding promise enhanced insights, bridging the gap between filmmakers and audiences. The future holds potential for groundbreaking contributions to both the cinematic landscape and sentiment analysis research, ushering in a new era of emotional connectivity in film critique.

References:

[1] Sentiment Analysis – Wikipedia – https://en.wikipedia.org/wiki/Sentiment_analysis

[2] IMDB Movie Review Dataset – https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

[3] Andrew L Mass, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng and Christopher Potts (2011). Learning Word Vectors for Sentiment Analysis

[4] Kigon Lyu Korea University, Korea "Sentiment Analysis Using Word Polarity of Social Media",Springer, 2016

[5] Monu Kumar Thapar University, Patiala "Analyzing Twitter sentiments through big data", IEEE, 2016

[6] Minhoe Hur Seoul National University "Box-office forecasting based on sentiments of movie reviews and Independent subspace method", Information Sciences, 2016

[7] Jorge A Balazs University of Chile "Opinion Mining and Information Fusion- A survey", 2015

[8] Pang, Bo; Lee, Lillian; Vaithyanathan, Shivakumar (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques". Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

[9] Turney, Peter (2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". Proceedings of the Association for Computational Linguistics.

[10] Tumasjan, Andranik; O.Sprenger, Timm; G.Sandner, Philipp; M.Welpe, Isabell (2010). "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment". "Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media"

[11] Natural Language Processing from Scratch - http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/35671.pdf

[12] Cambria, Erik; Schuller, Björn; Xia, Yunqing; Havasi, Catherine (2013). "New Avenues in Opinion Mining and Sentiment Analysis". IEEE Intelligent Systems

[13] Snyder, Benjamin; Barzilay, Regina (2007). "Multiple Aspect Ranking using the Good Grief Algorithm". Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference

[14] Chirag Sangani Stanford University, USA "Sentiment Analysis of App Store Reviews", 2013