Sentiment Analysis and Topic Modeling on Amazon Movie Reviews

CSE572 Final Project Report

Rakshith Chandrashekar
Arizona State University
Tempe, AZ, USA
rchand44@asu.edu

Aditya Sajeev
Arizona State University
Tempe, AZ, USA
asajeev4@asu.edu

Adesh Chougule
Arizona State University
Tempe, AZ, USA
amchougu@asu.edu

Aanshi Patwari
Arizona State University
Tempe, AZ, USA
apatwari@asu.edu

Abstract

Movie reviews serve as a valuable resource for understanding audience perceptions, providing nuanced qualitative insights alongside ratings. However, the unstructured and succinct nature of these reviews poses challenges for traditional sentiment analysis and topic extraction methods. In this project, we leverage machine learning techniques to extract topics and perform sentiment analysis on a dataset comprising movie reviews from Amazon. Our objective is twofold: first, to discern subjective opinions on various elements of films such as storyline, acting, cinematography, and overall enjoyment through sentiment analysis. Second, to identify recurring themes and topics within the reviews using topic extraction methodologies. By employing natural language processing (NLP) and short-text mining techniques, we aim to derive meaningful insights into audience sentiments towards movies.

We hypothesize that employing state-of-the-art (SOTA) NLP techniques will yield the most effective approach for sentiment analysis, particularly considering the inherent imbalance in our dataset. To address this, we propose integrating methods like undersampling, oversampling, and cost-sensitive learning with existing techniques to enhance accuracy. For topic modeling, traditional approaches like Latent Dirichlet Allocation (LDA) may struggle with the brevity and lack of structure in our data. Hence, we suggest utilizing metrics such as the Davies–Bouldin index to evaluate and compare different clustering algorithms, aiming to identify the most suitable algorithm for extracting meaningful topics from the dataset.

Keywords

Data Mining, Sentiment Analysis, Topic Modelling, Machine Learning, Neural Networks

1 Introduction

In the contemporary era of rapid technological advancements, the proliferation of digital platforms has led to an explosion in the generation of textual data across various forms such as news articles, webpages, social media posts, and blogs. Within this vast pool of data, movie reviews stand as a valuable source of qualitative information, providing insights into audience perceptions and preferences. However, the unstructured nature of textual data[17], coupled with the brevity of reviews, poses significant challenges in extracting meaningful insights through traditional methods.

The availability of data in unstructured formats presents a formidable obstacle in harnessing its potential for analysis. Movie reviews, often expressed in short, informal texts scattered across diverse platforms, lack the coherence and structure required for conventional analysis techniques. As a result, processing this data necessitates sophisticated methodologies capable of deciphering the latent patterns and sentiments embedded within.

Sentiment analysis emerges as a pivotal technique in the realm of textual data analysis, particularly for movie reviews. This computational approach, also known as opinion mining[25], involves the extraction and classification of sentiments expressed in textual data. By discerning the polarity of opinions—whether positive, negative, or neutral—towards various aspects of films, sentiment analysis provides invaluable insights into audience reactions and preferences. Moreover, sentiment analysis extends beyond movie reviews to encompass a wide array of textual data, including social media posts[20], news articles, and customer feedback, thereby offering versatile applications across industries.

Complementary to sentiment analysis, topic modeling offers a systematic framework for uncovering latent themes and subjects within textual data. In the context of movie reviews, topic modeling endeavors to identify recurring topics or motifs discussed across reviews, providing a deeper understanding of audience perceptions and preferences. By automatically extracting topics[2] inherent within a corpus of documents, topic modeling facilitates the organization, search, and summarization of textual data, thereby enabling researchers and analysts to navigate through the vast sea of information effectively.

Despite the promise offered by sentiment analysis and topic modeling, several challenges persist, particularly concerning unstructured data and short-length texts. Traditional methods often falter in handling the polysemy, context, and noise inherent in such data, leading to suboptimal results. To address these shortcomings, we propose a comprehensive approach leveraging machine learning models, word embeddings, and deep learning methods, alongside innovative sampling methods and advanced evaluation metrics.

Our approach intends to integrate machine learning and word embeddings with classical methods to improve the precision and efficiency of sentiment analysis and topic modeling. We plan to train algorithms on well-defined features and utilize word embeddings to grasp the subtleties of language and context[3] found in movie reviews. Additionally, we aim to refine our methodologies through advanced sampling techniques and robust evaluation metrics to ensure accurate and reliable results. We also propose using clustering to uncover thematic patterns in the data, providing a more holistic understanding of the content. Overall, we are combining established practices with modern innovations to overcome the complexities of analyzing unstructured and concise text data.

The structure of this paper is outlined as follows: Section 2 reviews recent advancements in research focusing traditional and innovative models utilised for performing task of sentiment analysis and topic modeling. Section 3 elaborates on the dataset employed and its details. Section 4 describes about preprocessing steps, dataset splitting and implemented models. In Section 5, comprehensive results for the utilized models are presented. Section 6 delves into further discussion about modifications which can be implemented in future. Finally, Section 7 encapsulates our project's concluding remarks.

2 Related Works

In the current research landscape, sentiment analysis has garnered significant attention, particularly with the advent of social media platforms like Twitter, which present unique challenges due to the colloquial and dynamic nature of the content. Traditional sentiment analysis approaches, including lexicon-based methods [4], [21], and machine learning techniques [13], [15], have been extensively explored, with hybrid methods showing promising enhancements in performance [5], [9]. Notably, Go et al. [7] and Pang et al. [18] have pioneered binary classifications of tweets, employing distant supervision and analyzing the effectiveness of negation with POS tags, respectively. Mohammad et al. [12] and Kiritchenko et al. [10] further advanced the field by employing SVM classifiers to achieve superior performance in sentiment classification tasks. The exploration of ensemble classifiers by Da Silva et al. [5] and lexicon-based methods [24], [23] has contributed to improving classification accuracy, albeit with certain limitations highlighted by the unique characteristics of tweets [24], [23].

Parallelly, topic modeling has emerged as a crucial technique in natural language processing, offering insights into the semantic structures of large text corpora. George and Birla [6] provided a comprehensive overview of topic modeling methods, from Latent Semantic Analysis (LSA) to more complex models like Latent Dirichlet Allocation (LDA) and its extensions [Correlated LDA, Dynamic Topic Model, Hierarchical LDA], addressing the challenges in analyzing untagged textual data. The advancements in domain-specific sentiment analysis, as demonstrated by the Domain Attention Model (DAM) by Yuan et al. [26], highlight the significance of incorporating domain knowledge to enhance sentiment classification accuracy across various domains.

Moreover, the study by Vanaja and Belwal on aspect-level sentiment analysis [25] emphasizes the importance of a granular analysis of customer feedback on e-commerce platforms, utilizing Amazon customer reviews to derive insights into consumer sentiments regarding product features or services. This approach marks a departure from traditional methods, focusing instead on extracting actionable insights from customer feedback to align offerings closely with consumer expectations.

The advent of deep learning has ushered in a new era in sentiment analysis, with neural network language models [1], skip-gram models [11], [8], and deep contextual embeddings [14], [13], [16] being applied to enhance the performance of sentiment classification tasks. These studies underscore the potential of deep learning techniques in addressing the complexities of sentiment analysis, particularly in the context of social media.

Lastly, the survey by Qiang et al. [19] on short text topic modeling techniques underscores the challenges posed by the sparse nature of short texts, such as tweets. By categorizing approaches into Dirichlet Multinomial Mixture (DMM) based methods, global word co-occurrences, and self-aggregation strategies, this survey highlights the advancements in tackling the sparsity issue prevalent in short texts, thus enhancing the semantic understanding of such texts. The development of the Short Text Topic Modeling (STTM) Java library [19] represents a significant step forward, offering a unified platform for applying these algorithms to real-world datasets, thereby facilitating further research and application in the field.

This overview of recent research developments in sentiment analysis and topic modeling illustrates the dynamic nature of the field, highlighting the ongoing efforts to refine methodologies and improve the accuracy and applicability of sentiment classification and topic discovery in the face of evolving data characteristics.

3 Data

The dataset[22] comprises movie reviews from Amazon spanning a period of over 10 years, encompassing approximately 8 million reviews collected from August 1997 up to October 2012. It includes product and user information, ratings, and plaintext reviews, along with reviews from other Amazon categories.

The dataset statistics are as follows: it contains 7,911,684 reviews (1,084,731 negative reviews, 791,594 neutral reviews, and 6,035,359 positive reviews.), contributed by 889,176 users for 253,059 products. Among these users, 16,341 have provided more than 50 reviews each. The median number of words per review is 101, reflecting the dataset's textual richness.

Below is the table presenting the attributes and their descriptions:

Attribute	Description
product/productId	Unique identifier for the product (e.g., B00006HAXW)
review/userId	Unique identifier for the user providing the review
review/profileName	Name of the user providing the review
review/helpfulness	Fraction of users who found the review helpful
review/score	Rating of the product on a scale of 1 to 5
review/time	Time of the review (Unix time)
review/summary	Summary of the review
review/text	Full text of the review

Table 1: Attributes and their Descriptions

Here is an example illustrating the attribute values:

- product/productId: B00006HAXW
- review/userId: A1RSDE90N6RSZF
- review/profileName: Joseph M. Kotow
- review/helpfulness: 9/9
- review/score: 5.0
- review/time: 1042502400
- review/summary: Pittsburgh - Home of the OLDIES
- review/text: "I have all of the doo wop DVD's and this one is as good or better than the 1st ones. Remember once these performers are gone, we'll never get to see them again. Rhino did an excellent job and if you like or love doo wop and Rock n Roll you'll LOVE this DVD !!"

4 Methods

4.1 Data Preparation and Exploration

4.1.1 CSV formation

The dataset available is in the form of a text file where each record is organized in row form with multiple lines of key-value pairs per record. We performed a line-by-line processing to extract and aggregate data. This transformed the dataset into a structured CSV file, facilitating analysis. Cleaning and validation ensured data integrity, resulting in a finalized dataset ready for analysis.

4.1.2 Dataset Pruning

Given the considerable volume of 8 million reviews, processing and modeling on the full dataset posed computational challenges. To facilitate binary classification between positive and negative sentiments, we excluded neutral reviews from our analysis. Additionally, to prevent model bias due to class imbalance, we deliberately balanced the dataset by limiting the number of positive reviews to approximately 1 million, aligning closely with the quantity of negative reviews. This approach ensures a more even distribution of classes, allowing for unbiased model training and evaluation.

4.1.3 Data Exploration

The provided histogram illustrates the distribution of review lengths within a particular dataset. This graphical representation is vital for understanding the verbosity of the reviews, which can have implications for text processing and analysis. The histogram indicates that a significant majority of the reviews are brief, with lengths concentrated toward the lower end of the scale.

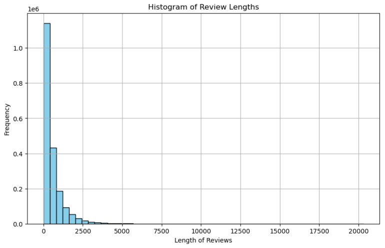


Figure 1: Distribution of review lengths histogram

4.2 Task 1 - Sentiment Analysis

4.2.1 Preprocessing

Preprocessing data, for our task is a critical step to ensure that the models receive clean and standardised input, which helps in improving both the accuracy and efficiency of the models. We have formed a comprehensive preprocessing pipeline. Each of these steps prepares the data for efficient and effective analysis, modeling, or both, depending on the specific requirements of the task.

Punctuations and Stopwords Removal: Punctuation marks can often be irrelevant for many tasks and can add unnecessary noise to the data. While stop words are common words (e.g., "the", "is", "in") that usually do not carry significant meaning and are often removed from the text to reduce the dataset size and improve processing speed. Removing them simplifies the dataset and can make further processing steps more effective. For sentiment analysis, we have used regular expressions or built-in string manipulation functions in Python to remove punctuation marks from our text data and utilised a predefined list of stop words to filter them out from our tokens.

Tokenization: Tokenization involves breaking down text into smaller units (tokens), which can be as small as words or as large as sentences. This converts unstructured text into a structured form that algorithms can understand and analyze. There are many tokenization methods, from simple space-based splitting to sophisticated algorithms that can correctly identify tokens in complex linguistic contexts. For sentiment analysis, we have used the nltk tokenizer which is a ready-to-use tokenizer.

4.2.2 Feature Engineering

Feature engineering for textual data involves transforming raw text into structured numerical features that machine learning models can understand.

Feature Extraction: To convert text data into numerical features, we have used the CountVectorizer, TFIDFVectorizer from scikit-learn library and Glove embeddings. CountVectorizer was used to convert a collection of text documents into a matrix of token counts.

TFIDF used to evaluate the importance of a word in a document relative to a collection of documents. Glove was used to represent words as dense vectors in a continuous vector space and lower-dimensional word embeddings.

Feature Scaling: Feature scaling is crucial for machine learning algorithms relying on distance or optimization. It ensures equal contribution from all features during model fitting, aiding convergence and reducing bias. We have used the StandardScaler from scikit-learn to standardize features by scaling them to unit variance.

Feature Reduction: The feature extraction step using Vectorizer gives approx. 37K features. To reduce the dimensionality for efficient processing, we have used Principal Component Analysis (PCA) technique. PCA transforms the original high-dimensional data into a lower-dimensional space by finding orthogonal axes (principal components) that capture the maximum variance in data.

4.2.3 Dataset split

The dataset after preprocessing features 1,980,243 Amazon movie reviews from 889,176 users, with a notable textual richness and a median of 101 words per review.

For the purposes of sentiment analysis, we have partitioned the dataset into three subsets: training, validation, and testing. The training set, comprising 60% of the data(1,188,145 reviews), is used to train our models, enabling them to learn the intricacies of sentiment present in the reviews. The validation set accounts for 20% of the data(396,049 reviews), serving as a tool to fine-tune model parameters and prevent overfitting. Finally, the testing set, also representing 20% of the data(396,049 reviews), is employed to evaluate the performance of our models, providing insights into their generalization capabilities on unseen data.

4.2.4 Models

Logistic Regression: Logistic Regression(LR) is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable in which there are only two possible outcomes.
Logistic Regression models the probability of the positive class using the logistic function:

$$P(y = 1|\mathbf{x}; \mathbf{w}, b) = \frac{1}{1 + e^{-(\mathbf{w}\cdot\mathbf{x}+b)}}$$

Where: $\mathbf{x}$ is the input feature vector, $\mathbf{w}$ is the weight vector, b is the bias term, e is the base of the natural logarithm.
It provides the probability for a given input belonging to a class (e.g., the probability of a text being positive). This probabilistic output can be very informative for understanding model decisions. It is relatively simple and quick to train, yet very effective for linearly separable problems.We chose this as it can provide insights into the importance of different features (words or n-grams) for predicting sentiment, which can be useful for understanding and improving the model.

Multinomial Naive Bayes: Multinomial Naive Bayes(MNB) is a probabilistic learning method frequently used in Natural Language Processing (NLP). The model calculates the probability of a document belonging to a class based on the presence of terms in it, assuming independence between the features.
The class conditional probability for each feature given the class label is estimated as:

$$P(x_j|y) = \frac{\sum_{i=1}^{N} \mathrm{I}(x_j, x_{ij})}{\sum_{k=1}^{K} \sum_{i=1}^{N} \mathrm{I}(x_k, x_{ik})}$$

Where: x_j is the j-th feature, y is the class label, x_{ij} is the j-th feature value of the i-th sample, N is the number of training samples, K is the number of unique feature values.
MNB applies Bayes' theorem with the naive assumption of independence between every pair of features. We chose MNB because it is computationally efficient, making it fast for training and prediction, which is beneficial for large datasets often encountered in sentiment analysis. It works particularly well with count-based feature vectors, such as word counts, commonly used in text processing.

Support Vector Machines: Support Vector Machines(SVM) is a supervised learning model used for classification and regression. It tries to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points.
The decision function of SVM can be defined as:

$$f(\mathbf{x}) = \mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + b)$$

Where: $\mathbf{x}$ is the input feature vector, $\mathbf{w}$ is the weight vector, b is the bias term, $\mathrm{sign}(\cdot)$ is the sign function.
The objective function of SVM is given by:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \max(0, 1 - y_i(\mathbf{w} \cdot \mathbf{x}_i + b))$$

Where: N is the number of training samples, $\mathbf{x}_i$ is the i-th input feature vector, y_i is the true class label of the i-th sample, C is the regularization parameter.
We chose SVM as it is particularly well-suited for sentiment analysis and it performs well with high-dimensional data, which is typical in text data where each word or n-gram can be considered a dimension. SVM aims to find the optimal separating hyperplane that maximizes the margin between different classes (in our case positive and negative sentiments), leading to better generalization on unseen data.

Multi Layer Perceptron Neural Networks: Multi Layer Perceptron(MLP) Neural networks are computational models inspired by the structure and function of the human brain. They consist of interconnected nodes organized into layers, where each node performs a computation on its input and passes the result to the next layer.
The output $\hat{y}$ of a neural network can be represented as:

$$\hat{y} = f(W_3 \cdot f(W_2 \cdot f(W_1 \cdot x + b_1) + b_2) + b_3)$$

where x is the input data, W_i are the weight matrices, b_i are the bias vectors, and f is the activation function.
By feeding textual inputs into the network and training it on labeled data, the model learns to recognize patterns and associations

between words and sentiments. The choice of a neural network architecture, such as the one described with three dense layers, along with specific activation functions and optimization techniques, significantly influences the sentiment analysis performance.
Here, we used Rectified Linear Unit (ReLU) activation functions in hidden layers introduces non-linearity, enabling the network to capture complex relationships within the text data. We also used the sigmoid activation function in the output layer as it is suitable for binary classification tasks like sentiment analysis, where it outputs probabilities indicating the likelihood of each sentiment class. Futher we compiled the model with the Adam optimizer and binary cross-entropy loss function to enhance training efficiency and accuracy by adjusting the model's parameters to minimize prediction errors.

4.3 Task 2 - Topic Modeling

First, review text was splitted into 2 separate dataframes namely positive and negative reviews dataframe for review wise topic modeling. Then, preprocessing(tokenization, stopword removal and lemmetization) was applied, followed by feature engineering which formed the input for topic modeling purposes.

4.3.1 Data Preprocessing

Stopwords Removal: For topic modeling purposes, we have removed all 1-letter, commonly used 2-letter and 3-letter words to avoid them in the word cloud formation.

Tokenization: For topic modeling, we have used tokenization to convert the dataframe rows into documents format for further feature engineering.

Lemmetization: Lemmatization is a text normalization technique that reduces words to their base or dictionary form, known as the lemma, to consolidate different forms of the same word. For topic modeling, lemmatization was used to used the variations of words.

4.3.2 Feature Engineering

Dictionary Formation: It involves compiling a structured inventory of words with associated linguistic properties. Here, this process is done by extracting unique words from a corpus and assigning frequency counts forming a table like structure.

Corpus Creation: Corpus creation is the process of assembling a structured collection of texts representative of a language or domain. We have converted each document (review) into a bag-of-words representation to form a corpus for positive and negative reviews.

4.3.3 Models

Latent Dirichlet Allocation: Latent Dirichlet Allocation(LDA) is a statistical model that explains sets of observations as generated from hidden groups, specifically catering to how certain parts of data are grouped together. LDA is particularly used for identifying latent topics from a collection of documents, where each document can be considered a mixture of various topics and each topic as a collection of words with certain probability distributions.
The generative process for a corpus D consisting of M documents each of length N_i is as follows:

1. Choose $\theta_i \sim \text{Dir}(\alpha)$, where $i = 1, \dots, M$.
2. For each of the word positions $j = 1, \dots, N_i$, in document i:
   (a) Choose a topic $z_{ij} \sim \text{Multinomial}(\theta_i)$.
   (b) Choose a word $w_{ij} \sim \text{Multinomial}(\beta_{z_{ij}})$.

Where: θ_i is the topic distribution for document i, α is the parameter of the Dirichlet prior on the topic distributions, z_{ij} is the topic for the j-th word in document i, β is the word distribution for a given topic, w_{ij} is the specific word.
LDA is extensively used in topic modeling, particularly in the analysis of large volumes of textual data. Out core idea is to infer the topics present in a collection of documents automatically with each topic being characterized by a set of words, and each document is considered a mixture of these topics. By applying LDA, we were able to discern the underlying thematic structure of a corpus, assign topics to documents, and list out the most significant words per topic. This capability makes LDA invaluable for tasks such as organizing large archives of documents, summarizing texts, and guiding the exploration of key themes in large textual datasets.

Latent Semantic Analysis: Latent Semantic Analysis(LSA) is a technique in natural language processing and information retrieval used to extract and represent the contextual-usage meaning of words by statistical computations applied to a large corpus of text. LSA is based on the principle that words used in similar contexts tend to have similar meanings. Essentially, it reduces the dimensionality of text data by applying singular value decomposition (SVD) to the term-document matrix, which describes the occurrences of terms in documents.
Latent Semantic Analysis (LSA) uses Singular Value Decomposition (SVD) to reduce the dimensions of the term-document matrix in text analysis. The SVD of a term-document matrix A is given by:

$$A = U\Sigma V^T$$

where U contains the left singular vectors, Σ is a diagonal matrix of singular values, and V^T contains the right singular vectors.
For our topic modeling purpose, LSA was used to identify patterns in the relationships among the terms and documents in a dataset, effectively uncovering the latent topics within the texts. By decomposing the original term-document matrix into matrices representing singular vectors and singular values, LSA allows for the approximation of the matrix by retaining only the top k singular values and corresponding vectors. This reduced representation captures the most significant semantic structures, making it easier to interpret and categorize the underlying topics in the documents. Topics are then typically interpreted by examining the terms that are most strongly associated with each left singular vector in U.

Non-negative Matrix Factorization: Non-negative Matrix Factorization(NMF) is a group of algorithms in multivariate analysis and linear algebra where a matrix V is factorized into (usually) two matrices W and H, with the condition that all three matrices have no negative elements. This non-negativity makes the resulting matrices easier to inspect. NMF is useful for data analysis techniques such as feature extraction and classification.
The Non-negative Matrix Factorization (NMF) is represented by the equation:

$$V \approx WH$$

where V is the data matrix, W is the basis matrix, and H is the coefficient matrix, with all matrices having non-negative elements. NMF was applied for topic modeling, to extract the hidden topics from large volumes of text. For topic modeling, V represents a document-term matrix with documents as rows and terms as columns. Applying NMF involves decomposing V into W, which can be seen as the topic representation of the documents, and H, which represents how much each word is important to the topics.

Hierarchical Dirichlet Process: Hierarchical Dirichlet Process (HDP) is an advanced Bayesian nonparametric model, which can be considered an extension of the Dirichlet Process (DP). It provides a way to define a distribution over a potentially infinite number of latent (hidden) topics, where the number of topics does not need to be specified in advance.

$$\begin{aligned} G_0|\alpha_0, H &\sim \mathrm{DP}(\alpha_0, H) \\ G_j|\alpha_1, G_0 &\sim \mathrm{DP}(\alpha_1, G_0) \\ \theta_{ji}|G_j &\sim G_j \\ x_{ji}|\theta_{ji} &\sim F(\theta_{ji}) \end{aligned}$$

Here, G_0 is the global measure drawn from a Dirichlet Process with concentration parameter α_0 and base distribution H. Each G_j represents the distribution of topics for group j (e.g., a specific document), drawn from another Dirichlet Process with concentration parameter α_1 and based on the global distribution G_0. The θ_{ji} are the topic assignments for each element i in group j, and x_{ji} are the observed data generated from a distribution F parameterized by θ_{ji}.

In topic modeling, HDP allows for a flexible analysis of a collection of documents. Each document is modeled as a mixture of potentially infinite topics, where the topics themselves are drawn from a shared base distribution across the document corpus. This hierarchical approach enabled sharing of statistical strength through the common topics among all documents, while also allowing each document to adapt its specific mixture of these topics.

5 Results

5.1 Evaluation Metrics

5.1.1 Accuracy

Accuracy represents the simplest form of evaluation metric, measuring the proportion of correctly predicted instances out of the total number of instances. It is defined as:

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{Total Number of Predictions}}$$

In the context of sentiment analysis, accuracy is particularly useful here as the classes are evenly distributed.

5.1.2 Precision and Recall

Precision measures the accuracy of positive predictions and is defined as the ratio of true positive predictions to the total predicted positives. It is particularly critical in scenarios where the cost of a false positive is high. The formula for precision is:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

Recall, also known as sensitivity, measures the model's ability to identify all relevant instances and is defined as the ratio of true positive predictions to the actual positives in the dataset. Recall is calculated as:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

5.1.3 F1-score

The F1-score is the harmonic mean of precision and recall, offering a balance between these two metrics. It is particularly useful when you need to consider both false positives and false negatives equally, and is defined as:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

This metric is crucial for sentiment analysis where both types of errors (false positives and false negatives) can have significant implications, such as misinterpreting review sentiments, which could potentially lead to misguided strategic decisions.

5.1.4 ROC-AUC Curve

The ROC-AUC curve is a crucial evaluation metric for sentiment analysis models, particularly effective when applied to test data. It measures the model's ability to distinguish between classes—such as positive and negative sentiments—across various thresholds.

5.1.5 Word Cloud

Word clouds are an effective visual tool for evaluating topic modeling in movie reviews, offering immediate insights into the dominant themes and sentiments expressed by viewers. By creating separate word clouds for positive and negative reviews, analysts can visually identify the most frequent and impactful words associated with each sentiment.

5.2 Sentiment Analysis

5.2.1 Parameters used

For feature extraction, following parameters were used:

- CountVectorizer with max_features = 500, 1500, 2500
- TFIDFVectorizer with n_features = 500, 1500, 2500:
- Glove embeddings with dimensions of 50x50, 100x100, 200x200, 300x300

For models, following parameters were utilised:

- Logistic Regression with C=1, max_iter=500 and penalty='l2'
- Support Vector Machine with max_iter=500 and loss='hinge'
- Multinomial Naive Bayes for CV and TFIDF and Gaussian Naive Bayes for Glove embedding

5.2.2 Accuracy

The accuracy metric obtained from validation data is systematically displayed in a table to compare the performance of various models across different parameters.

	n_value	LR	SVM	NB	MLP
CountVectorizer	500	0.847	0.841	0.825	0.905
	1500	0.882	0.876	0.847	0.936
	2500	**0.892**	**0.885**	**0.850**	**0.943**
TFIDF	500	0.849	0.839	0.826	0.906
	1500	0.884	0.876	0.849	0.935
	2500	**0.894**	**0.883**	**0.883**	**0.942**
GloVe	50	0.765	0.762	0.713	0.807
	100	0.805	0.804	0.707	0.848
	200	0.831	0.829	0.708	0.875
	300	**0.841**	**0.840**	**0.708**	**0.886**

Table 2: Accuracy Comparison for different models and their parameters

5.2.3 ROC-AUC Curve

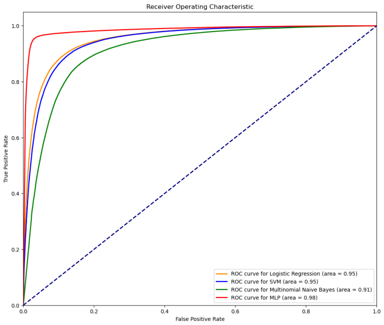


Figure 2: ROC-AUC Curve for Sentiment Analysis

The ROC-AUC curve results show that all four models(with parameters as CountVectorizer n=2500 as it performs the best among all model parameters)—Logistic Regression(Orange), Support Vector Machine(Blue), Naive Bayes(Green), and MLP(Red)—perform well in the sentiment analysis task, with AUC values all above 0.90, indicating strong predictive abilities.

5.2.4 Precision, Recall and F1-score

Model	Precision			Recall			F1 Score		
	CV	TFIDF	GloVe	CV	TFIDF	GloVe	CV	TFIDF	GloVe
LR	**0.90**	0.90	0.84	0.88	**0.89**	0.84	**0.89**	**0.89**	0.84
SVM	0.88	**0.89**	0.83	**0.89**	0.88	0.85	**0.89**	0.88	0.84
MLP	**0.94**	0.93	0.88	**0.95**	**0.95**	0.89	**0.94**	**0.94**	0.89
NB	0.84	**0.85**	0.67	**0.86**	0.85	0.81	**0.85**	**0.85**	0.74

Table 3: Sentiment Analysis Evaluation for Positive Reviews

The highest n-value yields the most favorable outcomes across all models, hence the table presents precision, recall, and F1-score metrics for n=2500 for both CountVectorizer and TFIDFVectorizer, and n=300*300 dimensions for GloVe embeddings.

Model	Precision			Recall			F1 Score		
	CV	TFIDF	GloVe	CV	TFIDF	GloVe	CV	TFIDF	GloVe
LR	0.88	**0.89**	0.84	**0.90**	**0.90**	0.84	**0.89**	**0.89**	0.84
SVM	**0.89**	0.88	0.85	0.88	**0.89**	0.83	**0.89**	0.88	0.84
MLP	**0.95**	**0.95**	0.89	**0.95**	0.93	0.88	**0.94**	**0.94**	0.89
NB	**0.86**	0.85	0.76	0.84	**0.85**	0.61	**0.85**	**0.85**	0.68

Table 4: Sentiment Analysis Evaluation for Negative Reviews

5.3 Topic Modeling

5.3.1 Parameters used

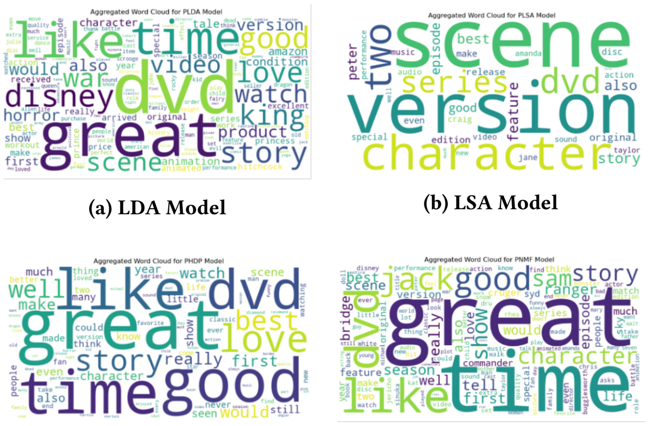
For models(for positive and negative reviews dataframe), following parameters were utilised:

- LDA with num_topics=10 and passes=10
- LSA with num_topics=10
- HDP with no explicit parameters
- NMF with num_topics=10 and passes=10

20 words per topic were used for LDA, LSA and NMF models' word cloud while 40 words per topic were used for HDP models' word cloud.

5.3.2 Word Cloud

Positive Reviews: From all models' word clouds, interpretations suggest that positive reviews are more related to love, disney, animation, king-queen type of movies.



(a) LDA Model

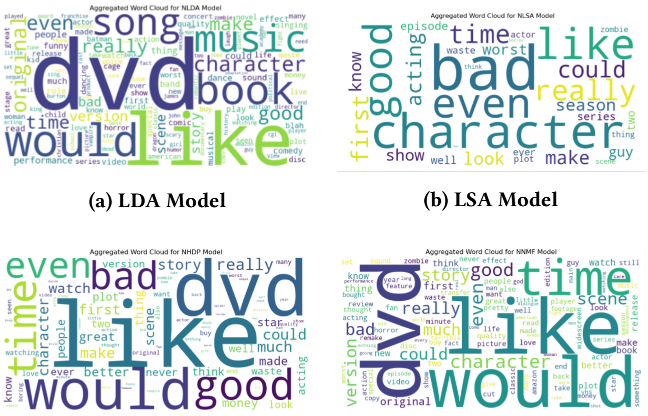
(b) LSA Model

(c) HDP Model

(d) NMF Model

Figure 3: Aggregated Word Clouds from Various Topic Modeling Techniques for Negative Reviews

Negative Reviews: From all models' word clouds, interpretations suggest that negative reviews are more related to horror, guy character, bad timing, remake types.



(a) LDA Model

(b) LSA Model

(c) HDP Model

(d) NMF Model

Figure 4: Aggregated Word Clouds from Various Topic Modeling Techniques for Negative Reviews

6 Discussion

6.1 Sentiment Analysis

In the discussion of our research findings, the comparative analysis of sentiment analysis models revealed significant trends and performances across varying parameters. Validation accuracy results demonstrated that all models, namely Logistic Regression (LR), Support Vector Machine (SVM), Naive Bayes (NB), and Multi-Layer Perceptron (MLP), exhibited enhanced accuracy with the increment of feature numbers from CountVectorizer and TFIDF, from 500 to 2500 features. This enhancement suggests that a more extensive feature set enables a more nuanced understanding of the data. Notably, the MLP model outperformed others consistently, achieving the highest accuracy with 2500 features for both CountVectorizer and TFIDF, showcasing its robustness in handling feature-rich data in sentiment classification tasks.

When examining the models trained with GloVe embeddings, we observed an upward trend in accuracy correlating with the increase in embedding dimensions from 50 to 300. However, the accuracy scores obtained with GloVe were marginally lower than those of CountVectorizer and TFIDF. This outcome could imply that within the context of this validation set, pre-trained GloVe embeddings may be less adept at capturing sentiment nuances compared to other feature extraction methodologies. Nevertheless, the MLP model maintained its leading performance, peaking in accuracy with 300-dimensional embeddings, further solidifying its superiority across diverse feature extraction techniques.

The Receiver Operating Characteristic-Area Under Curve (ROC-AUC) results corroborated the high predictive capabilities of the models, as all reported AUC values exceeded the 0.90 threshold. Both Logistic Regression and SVM distinguished themselves with an AUC of 0.95, indicative of their proficiency in sentiment classification. Although the Multinomial Naive Bayes model presented a slightly lower AUC of 0.91, it still held a commendable performance. The MLP model emerged as the most exceptional, boasting an AUC of 0.98, underscoring its superior accuracy in identifying positive and negative sentiments.

Overall, the findings from this study underscore the critical influence of feature selection on model performance in sentiment analysis tasks and highlight the MLP model's superior adaptability and predictive strength in evaluating sentiments within textual data.

6.2 Topic Modeling

In the discussion of our research results, the word clouds generated from topic modeling of both positive and negative reviews provide a nuanced understanding of the underlying sentiments in the dataset. For positive reviews, the word clouds from the HDP, LDA, LSA, and NMF (positive reviews) models prominently feature affirmative terms like "great," "love," "story," and "best." These terms not only reflect the positive sentiment of the reviews but also suggest a focus on narrative quality and enjoyment of media. Notably, terms related to home entertainment such as "dvd" and "watch," as well as references to family and animated films like "disney," indicate a preference for certain genres or viewing experiences.

Conversely, the word clouds for negative reviews from the HDP, LDA, LSA, and NMF (negative reviews) models reveal a mixture of positive and negative sentiments, indicating a complex response from reviewers. While positive words persist, the prominence of terms such as "bad," "waste," and "worse" suggests critical engagement with the content. The recurring appearance of "story" and "character" across both positive and negative reviews underscores the centrality of these aspects in shaping viewer reception and critique.

7 Conclusion

Multilayer Perceptron (MLP) neural network classifiers achieved high F1 scores of 0.94 for both positive and negative sentiment classes when employing CountVectorizer and TFIDFVectorizer in sentiment analysis tasks. Notably, these traditional vectorization methods outperformed the utilization of GloVe embeddings for document-level sentiment analysis. Additionally, it was observed that Latent Dirichlet Allocation (LDA) consistently yielded more interpretable results compared to alternative models in topic modeling endeavors. Despite encountering some common words in word clouds representing both positive and negative reviews, it was noted that further preprocessing techniques may be warranted to enhance interpretability in subsequent analyses.

References

[1] 2013. Deep learning of representations: Looking forward. *CoRR* abs/1305.0445 (2013). http://arxiv.org/abs/1305.0445

[2] Dimitar Angelov. 2020. Top2Vec: Distributed Representations of Topics. *arXiv* arXiv:2008.09470 (2020). https://doi.org/10.48550/arXiv.2008.09470

[3] Mohammad Ehsan Basiri, Shahla Nemati, Moloud Abdar, Erik Cambria, and U. Rajendra Acharya. 2021. ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis. *Future Generation Computer Systems* 115 (2021), 279–294. https://doi.org/10.1016/j.future.2020.08.005

[4] F. Chiavetta, G. Lo Bosco, and G. Pilato. 2016. A lexicon-based approach for sentiment classification of Amazon Books reviews in Italian language. 159–170. https://doi.org/10.5220/0005915301590170

[5] N.F. da Silva, E.R. Hruschka, and E.R. Hruschka. 2014. Tweet sentiment analysis with classifier ensembles. *Decis. Support Syst.* 66 (2014), 170–179. https://doi.org/10.1016/j.dss.2014.07.003

[6] L. E. George and L. Birla. 2018. A Study of Topic Modeling Methods. In *Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS '18).* https://doi.org/10.1109/ICCONS.2018.8663152

[7] A. Go, R. Bhayani, and L. Huang. 2009. Twitter sentiment classification using distant supervision. http://www.stanford.edu/alecmgo/papers/TwitterDistantSupervision09.pdf

[8] Z. Jianqiang and G. Xiaolin. 2018. Deep convolution neural networks for Twitter sentiment analysis. *IEEE Access* PP (2018), 1. https://doi.org/10.1109/ACCESS.2017.2776930

[9] V. N. Khuc, C. Shivade, R. Ramnath, and J. Ramanathan. 2012. Towards building large-scale distributed systems for twitter sentiment analysis. In *Proceedings of the 27th annual ACM symposium on applied computing.* 459–464.

[10] S. Kiritchenko, X. Zhu, C. Cherry, and S. Mohammad. 2014. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014).* 437–442.

[11] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (Eds.). Curran Associates, Inc., 3111–3119.

[12] Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013).* Association for Computational Linguistics, Atlanta, Georgia, USA, 321–327.

[13] U. Naseem, S.K. Khan, I. Razzak, and I.A. Hameed. 2019. Hybrid Words Representation for Airlines Sentiment Analysis. In *AI 2019: Advances in Artificial Intelligence. AI 2019. Lecture Notes in Computer Science*, Vol. 11919. Springer, Cham. https://doi.org/10.1007/978-3-030-35288-2_31

[14] U. Naseem and K. Musial. 2019. DICE: Deep Intelligent Contextual Embedding for Twitter Sentiment Analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR).* Sydney, NSW, Australia, 953–958. https://doi.org/10.1109/ICDAR.2019.00157

[15] Usman Naseem, Imran Razzak, Peter Eklund, and Katarzyna Musial. 2020. Towards improved deep contextual embedding for the identification of irony and sarcasm. In *2020 International joint conference on neural networks (IJCNN).* IEEE, 1–7.

[16] U. Naseem, I. Razzak, and I.A. Hameed. 2019. Deep Context-Aware Embedding for Abusive and Hate Speech detection on Twitter. *Aust. J. Intell. Inf. Process. Syst.* 15, 3 (2019), 69–76.

[17] U. Naseem, I. Razzak, K. Musial, and M. Imran. 2020. Transformer based Deep Intelligent Contextual Embedding for Twitter sentiment analysis. *Future Generation Computer Systems* 113 (2020), 58–69. https://doi.org/10.1016/j.future.2020.06.050

[18] B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *The Conference on Empirical Methods on Natural Language Processing, Association for Computational Linguistics.* 79–86. https://doi.org/10.3115/1118693.1118704

[19] Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. 2022. Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. *IEEE Transactions on Knowledge and Data Engineering* 34 (2022). https://doi.org/10.1109/TKDE.2020.2992485

[20] J. F. Raisa, M. Ulfat, A. Al Mueed, and S. M. S. Reza. 2021. A Review on Twitter Sentiment Analysis Approaches. In *Proceedings of the 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD).* Dhaka, Bangladesh, 375–379. https://doi.org/10.1109/ICICT4SD50815.2021.9396915

[21] Z. Saeed, R.A. Abbasi, A. Sadaf, M.I. Razzak, and G. Xu. 2018. Text Stream to Temporal Network - A Dynamic Heartbeat Graph to Detect Emerging Events on Twitter. In *Advances in Knowledge Discovery and Data Mining. PAKDD 2018. Lecture Notes in Computer Science*, Vol. 10938. Springer, Cham. https://doi.org/10.1007/978-3-319-93037-4_42

[22] Stanford Network Analysis Project (SNAP). Accessed: 2024. Amazon movie reviews. https://snap.stanford.edu/data/web-Movies.html.

[23] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Comput. Linguist.* 37, 2 (2011), 267–307. https://doi.org/10.1162/COLI_a_00049

[24] P.D. Turney. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.* ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA, 417–424. https://doi.org/10.3115/1073083.1073153

[25] S. Vanaja and M. Belwal. 2018. Aspect-Level Sentiment Analysis on E-Commerce Data. In *2018 International Conference on Inventive Research in Computing Applications (ICIRCA).* Coimbatore, India, 1275–1279. https://doi.org/10.1109/ICIRCA.2018.8597286

[26] Zhigang Yuan, Sixing Wu, Fangzhao Wu, Junxin Liu, and Yongfeng Huang. 2018. Domain attention model for multi-domain sentiment classification. *Knowledge-Based Systems* 155 (2018), 1–10. https://doi.org/10.1016/j.knosys.2018.05.004