Int.J.HIT.TRANSC:ECCN.Vol.9: Issue 1A (2023) Page 45-53 ISSN: 0973-6875

Available Online at www.hithaldia.in/locate/ECCN
All Rights Reserved

ORIGINAL CONTRIBUTION

Smart Movie Review Analysis: A Data-Driven Sentiment Classification System

[a]Dipankar Dey, [b]Sudip Diyasi, [c]Supriya Maity, [d]Prajna Bhunia and [d]Ankita Ghosh

[a,c]Department of Computer Application, Global Institute of Science & Technology, Haldia 721657, West Bengal, India

[d]Department of Computer Application, George College of Management and Science, Kolkata 700141, West Bengal, India

[b]*Department Of Chemistry, Shahid Matangini Hazra Government General Degree College For Women, Chakshrikrishnapur, Kulberia,Tamluk, Purba Medinipur 721649, West Bengal, India.*

ABSTRACT

The proposed movie review system leverages machine learning and natural language processing (NLP) techniques to evaluate and classify user feedback on films. By analysing textual reviews, the system categorizes sentiment into positive, neutral, and negative sentiments, offering a clear understanding of audience reactions. Key NLP techniques such as tokenization, stemming, and stop word removal prepare the data for further processing. Sentiment prediction is achieved using machine learning models like Naive Bayes and Random Forest. The system detects patterns in movie reviews, providing valuable insights into film performance and audience satisfaction. It supports prospective viewers by offering a summarized sentiment analysis to aid in their decision-making. Additionally, it assists filmmakers and industry professionals by offering data-driven insights into audience feedback, improving decision-making in areas like marketing, production, and development. In essence, the system enhances interaction between filmmakers and audiences, fostering growth in the entertainment industry.

KEY WORDS: Movie Review System; Sentiment Analysis; Natural Language Processing; Machine Learning; Audience Feedback.

1. INTRODUCTION

The film industry, with its diverse genres and massive audience, heavily relies on understanding audience sentiments. Movie review systems are essential in this process, as they collect and analyse user feedback to gauge public opinion. The explosion of movie reviews, driven by social media and online platforms, has made it increasingly difficult for filmmakers and viewers to manually extract meaningful insights. This study introduces an automated, scalable solution for real-time sentiment analysis through a movie review system that utilizes advanced machine learning (ML) and natural language processing (NLP) algorithms to classify sentiments within user-generated reviews.

One of the major benefits of this proposed system is its ability to efficiently summarize collective opinions about a film in a data-driven manner. By classifying reviews into positive, neutral, and negative sentiments, the system enables potential viewers to make informed decisions based on overall public sentiment. Additionally, filmmakers, producers, and industry stakeholders can gain valuable insights, allowing them to refine their strategies and better align with audience expectations. The effectiveness of this system not only improves the accuracy of movie performance predictions but also reduces time consumption and manual effort.

Recent advancements in sentiment analysis, particularly with respect to user-generated content like movie reviews, have led to the exploration of various techniques aimed at enhancing classification accuracy. This proposed system builds on these advancements by

*Corresponding Address:
Doi: 10.5281/zenodo.14740314

integrating ML models for sentiment classification with NLP techniques for processing textual data. By merging these technologies, the system not only categorizes sentiments but also identifies trends and patterns in reviews, offering a deeper understanding of audience behaviour and preferences.

2. Contribution of the Research

- Suggests an automated method for classifying movie reviews based on sentiment.
- Combines machine learning and natural language processing to analyse sentiment with high accuracy.
- Provides real-time sentiment insights to filmmakers and producers.
- Enhances the efficiency of classifying user-generated reviews.
- Offers a valuable tool for the entertainment industry to predict movie success.

3. Objectives of the Proposed System

The objective of this paper is to develop an automated system that accurately classifies movie reviews into sentiment categories (positive, neutral, and negative) using a combination of machine learning and natural language processing techniques [1]. The system aims to provide real-time sentiment insights to filmmakers and producers, enhance the efficiency of classifying user-generated reviews, and offer a tool for the entertainment industry to predict the success of films based on audience feedback.

4. Dataset Collection

The dataset for the proposed system was sourced from the Kaggle database, known for its extensive and diverse collection of datasets [2]. This dataset consists of 2,000 movie reviews, evenly split into 1,000 positive and 1,000 negative reviews, ensuring a balanced representation of sentiments. All reviews are written in English and have been pre-processed to maintain consistency. The preprocessing steps include converting all text to lowercase, adding spaces around punctuation for better tokenization, and formatting each review so that each sentence appears on a separate line.

The dataset is organized into two directories, positive and negative, with each review stored in an individual file following a structured naming convention. This structure provides a robust foundation for training and evaluating sentiment analysis models. The dataset's quality and balance make it ideal for developing a system capable of classifying sentiments into positive, neutral, and negative categories.

By utilizing this dataset, the proposed system aims to deliver accurate and meaningful insights into audience opinions through machine learning and natural language processing techniques. Its availability on Kaggle ensures easy access and suitability for modern sentiment analysis applications.

5. Background

The proposed method leverages various Natural Language Processing (NLP) and machine learning techniques to perform sentiment analysis on movie reviews. The approach follows a well-established, multi-stage pipeline for processing and classifying text to categorize reviews as positive, neutral, or negative in sentiment.

5.1 Text Preprocessing

The initial stage of sentiment analysis is text preprocessing, which is crucial for transforming raw textual data into a format suitable for analysis. In this approach, the RegexpTokenizer from the NLTK library is employed to break down movie reviews into individual words, eliminating non-alphanumeric characters. Additionally, stop words (such as "the," "and," etc.) are removed using NLTK's stop words module. This helps in ensuring that only relevant words are considered in the analysis, minimizing the noise in the data and enhancing the accuracy of sentiment classification.

5.2 Stemming

Following tokenization and stop word removal, the next step is stemming, which reduces words to their base or root form [3]. The

PorterStemmer from NLTK is applied here to consolidate various forms of a word (e.g., "running," "runner") into a single root form ("run"). This process helps the model recognize underlying patterns in the data more effectively by reducing the impact of minor variations in word forms.

5.3 Feature Extraction

Once the text has been pre-processed, feature extraction comes next. This step converts the processed text into numerical features that can be used by machine learning algorithms [4]. The CountVectorizer from Scikit-learn is employed to create a matrix of token counts [5]. This sparse matrix represents the frequency and occurrence of words in the reviews. To manage the dimensionality, the number of features is limited to the top 1500 most frequent words, focusing the analysis on the most important terms. The resulting feature matrix is then used as input for the machine learning model.

5.4 Sentiment Classification

With the features extracted, the next step is sentiment classification. The algorithm initially uses a Naive Bayes classifier, which is known for its simplicity and effectiveness in text classification tasks, particularly with large datasets [6]. After training the classifier on a set of labelled movie reviews, it predicts the sentiment of unseen reviews.

Additionally, other classifiers such as Random Forest and Decision Tree are also implemented to compare their performance. These models are trained using the same feature matrix, and their accuracy is evaluated on a separate test set. The accuracy score function from Scikit-learn is used to measure the classification accuracy.

5.5 Evaluation

In the final step, the performance of the sentiment classification models is assessed using metrics like accuracy and the confusion matrix. While accuracy provides an overall evaluation of the model's correct predictions, the confusion matrix offers a detailed view of the distribution between predicted and actual sentiment labels, highlighting the strengths and weaknesses of each model [7, 8].

6. Literature Review

The amalgamation of ML and NLP is in fact an important constituent of modern examination of textual data, especially within entertainment, for sentiment analysis purposes. This section reviews the key contributions and approaches in sentiment classification systems, emphasizing text preprocessing, feature extraction, and machine learning algorithms as applied to movie review analysis.

6.1 Sentiment Analysis in Textual Data

Sentiment analysis, or opinion mining, involves identifying and categorizing opinions expressed in a text to determine their sentiment as positive, negative, or neutral. In recent years, it has been widely applied to user-generated content, such as movie reviews, to understand audience reactions. Studies have shown that robust preprocessing and feature extraction methods significantly impact the accuracy and reliability of sentiment classification models [9].

Key preprocessing steps include tokenization, stop word removal, and stemming. These steps transform raw text into structured input suitable for machine learning algorithms, ensuring that irrelevant noise is minimized. Additionally, feature extraction techniques, such as the CountVectorizer, are crucial for representing textual data numerically and identifying the most informative features for sentiment prediction.

6.2 Machine Learning in Sentiment Classification

Traditional machine learning algorithms, such as Naive Bayes, Decision Trees, and Random Forests, have proven effective in sentiment classification tasks. These models excel at handling structured, pre-processed data, offering a balance between interpretability and computational efficiency.

- **Naive Bayes Classifier**: This probabilistic model is particularly effective for large datasets with independent features, making

it a popular choice for initial sentiment classification.

- **Decision Tree Classifier**: Decision Trees recursively partition data based on feature importance, providing intuitive visualizations and explanations for classification decisions.

- **Random Forest Classifier**: As an ensemble method, Random Forest combines multiple decision trees to improve generalization and reduce overfitting, often outperforming single-model approaches in sentiment analysis.

The proposed system leverages these algorithms to classify movie reviews sourced from a balanced dataset of 2,000 reviews, achieving high classification accuracy. Out of all these models, the Random Forest model had the highest accuracy, which made it suitable for sentiment analysis in this domain.

6.3 Text Preprocessing and Feature Selection

Effective preprocessing is critical for enhancing model performance in sentiment analysis. The proposed system employs a multi-stage pipeline, including tokenization, stop word removal, and stemming, to prepare movie reviews for analysis. Feature extraction is performed using the CountVectorizer with a feature cap of the top 1,500 words, ensuring computational efficiency while retaining the most relevant terms.

Feature selection further refines the dataset by focusing on words with high predictive power for sentiment classification. Positive features like "awesome" and "best" and negative features like "worst" and "boring" are identified as significant contributors to the sentiment determination process [9].

6.4 Applications in the Film Industry

Sentiment analysis of movie reviews provides actionable insights for various stakeholders in the film industry. Filmmakers and producers can leverage these insights to gauge audience reactions, improve marketing strategies, and refine production decisions. Additionally, prospective viewers benefit from summarized sentiment analyses, aiding their decision-making process.

7. Feature Selection in Sentiment Analysis

Feature selection is a critical step in improving the efficiency and accuracy of sentiment analysis models. It involves identifying and retaining the most relevant features (such as words, phrases, or tokens) from the text data that are most influential in sentiment classification. The primary objective is to eliminate irrelevant or redundant information that could introduce noise and degrade the model's performance. Feature selection techniques often rely on frequency-based methods, where the most commonly occurring words in the dataset are selected, or domain-specific knowledge is used to identify words that have high predictive power. In our case, we utilized a `CountVectorizer` to extract the top 1500 most frequent words from the movie reviews, which were then analysed for sentiment.

7.1 Sentiment Analysis

Sentiment analysis refers to the process of determining the emotional tone or sentiment expressed in a piece of text and classifying it into categories such as positive, negative, or neutral. This involves several stages, including text preprocessing, feature extraction, and sentiment classification. Preprocessing steps like tokenization, stop word removal, and stemming are employed to clean and standardize the raw text data. After preprocessing, feature extraction converts the text into a numerical format, typically a vector of word counts, suitable for input into machine learning models. In our approach, sentiment analysis was performed using SentiWordNet to assign sentiment scores to words, and the polarity of each review was determined based on these scores.

7.2 Positive Features

In sentiment analysis, positive features refer to words or phrases that convey favourable sentiment or emotions such as approval, satisfaction, or joy. These features are essential for identifying content that reflects positive opinions, which are indicative of a good experience or sentiment. For instance, words like "best," "better," "awesome," and "awe" are all positive features, as they suggest that the reviewer had a favourable opinion. In our dataset, the most common positive sentiment words included terms like "best," "adrenalin," "avatar," and "bang," which are often associated with positive experiences in movie reviews. The presence of these positive features typically correlates with high ratings and favourable reviews, helping the sentiment analysis model classify the reviews as positive.

7.3 Negative Features

Negative features are words or phrases that express unfavourable sentiment, dissatisfaction, or criticism. These features are crucial for identifying text that reflects negative emotions such as disappointment, frustration, or discontent. Common negative sentiment words include "bad," "awful," "boring," and "worst," as these terms typically indicate disapproval or negative reactions. In our analysis, words such as "bad," "blast," "black," and "beast" were identified as negative sentiment features, indicating negative reviews. These words are usually associated with lower ratings or critical feedback. By identifying and analysing these negative features, the sentiment analysis model can accurately classify negative reviews and detect unfavourable opinions.

By analysing both positive and negative features in text, sentiment analysis algorithms are able to classify reviews accurately and provide meaningful insights into the emotional tone of the content. This ability to distinguish between positive and negative sentiments is particularly useful in applications like movie review aggregation and customer feedback analysis, where understanding public opinion is key.

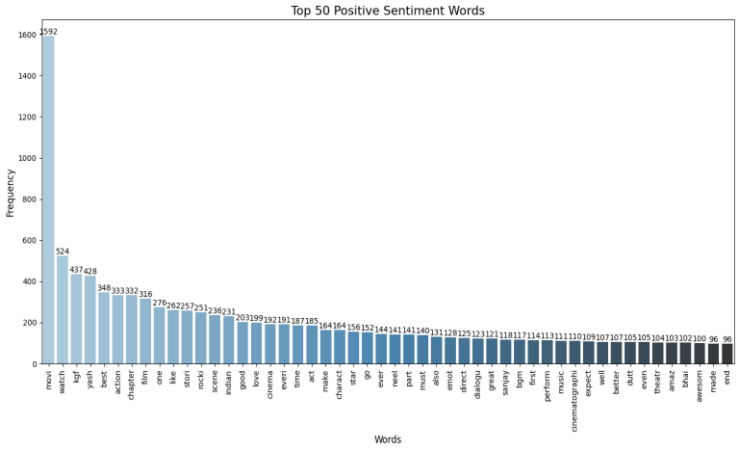

Figure 1: Represents top 50 Positive Sentiment Words.

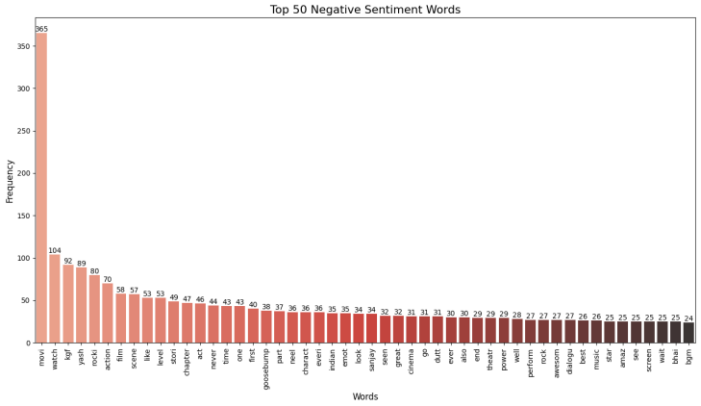

Figure 2: Represents top 50 Negative Sentiment Words.

8. Training Details

In the proposed model, multiple machine learning classifiers are trained to categorize movie reviews into three sentiment categories: positive, neutral, and negative. Each classifier follows a well-defined training process, utilizing the pre-processed text data and feature extraction pipeline outlined previously. Below are the details of the training process for each machine learning model in the proposed system.

8.1 Naive Bayes Classifier Training

The Naive Bayes classifier (See Figure 3) is the first model utilized in the sentiment analysis process. Known for its simplicity and probabilistic nature, it is highly effective for text classification, particularly when working with large datasets. During training, the feature

matrix—constructed from the word frequencies in the movie reviews—is provided as input, and the sentiment labels (positive, neutral, negative) serve as the target variable. The model computes the probabilities of words appearing in each sentiment category, which are then used to classify the sentiment of unseen movie reviews. After training, the model's performance is evaluated using a separate validation set to ensure its ability to generalize effectively (See Table 1).

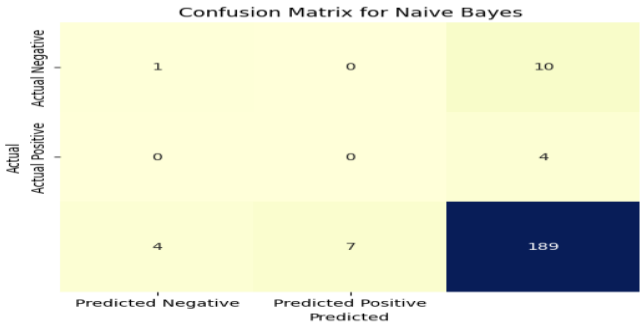

Figure 3: Confusion Matrix for Naïve Bayes

Table 1: Naive Bayes Classification Report

Metric	Class 0 (0.0)	Class 1 (1.0)	Class 2 (2.0)	Accuracy	Macro Avg	Weighted Avg
Precision	0.2	0	0.931034	0.883721	0.377011	0.876311
Recall	0.090909	0	0.945	0.883721	0.345303	0.883721
F1-Score	0.125	0	0.937965	0.883721	0.354322	0.878921
Support	11	4	200		215	215

8.2 Decision Tree Classifier Training

The Decision Tree classifier (See Figure 4) is the second model used in the proposed system. It works by recursively splitting the feature space into subgroups based on the most informative features. During the training phase, the Decision Tree model uses the same feature matrix derived from the pre-processed reviews. The model constructs a tree structure where each node represents a feature, and the tree splits into branches that classify the reviews into sentiment categories. The entropy criterion is used to guide the splits, ensuring that the partitions are as pure as possible. After training, the model's classification accuracy is tested using a separate test set to determine its effectiveness (See Table 2).

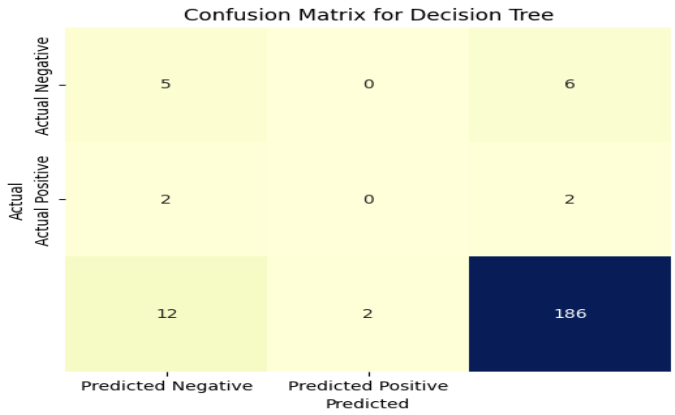

Figure 4: Confusion Matrix for Decision Tree

Table 2: Decision Tree Classification Report

Metric	Class 0 (0.0)	Class 1 (1.0)	Class 2 (2.0)	Accuracy	Macro Avg	Weighted Avg
Precision	0.263158	0	0.958763	0.888372	0.407307	0.905336
Recall	0.454545	0	0.93	0.888372	0.461515	0.888372
F1-Score	0.333333	0	0.944162	0.888372	0.425832	0.895345
Support	11	4	200		215	215

8.3 Random Forest Classifier Training

It is, in fact an ensemble learning methodology; the Random Forest classifier constructs many decision trees and synthesizes their results [10]. In this approach, a set of decision trees is trained on random subsets of the data set; and the final sentiment prediction is aggregated from the outputs of each individual tree. Although the feature matrix employed is more or less the same as in the Decision Tree model, the Random Forest method introduces additional randomness in the process due to the random choice of feature subsets at every split. This approach improves the model's robustness, reduces overfitting, and enhances generalization. In the proposed system, the Random Forest model is trained with 10 trees, and its performance is evaluated on an unseen test set to assess its accuracy (See Table 3) [11] .

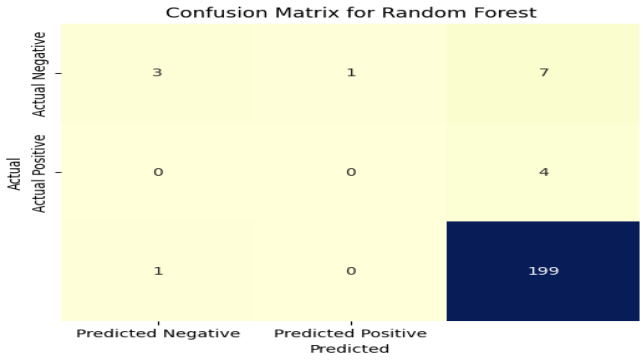

Figure 5: Confusion Matrix for Random Forest

Table 3 Random Forest Classification Report

Metric	Class 0 (0.0)	Class 1 (1.0)	Class 2 (2.0)	Accuracy	Macro Avg	Weighted Avg
Precision	0.263158	0	0.958763	0.888372	0.407307	0.905336
Recall	0.454545	0	0.93	0.888372	0.461515	0.888372
F1-Score	0.333333	0	0.944162	0.888372	0.425832	0.895345
Support	11	4	200		215	215

9. Model Evaluation and Performance Comparison

After training each classifier, the system evaluates their performance using accuracy and confusion matrices. Accuracy is computed by comparing the sentiment labels it predicts to the actual labels of the test set, giving an overall measure of how good a model is. The confusion

matrix provides additional insights into how well the models handle each sentiment category by showing the number of correct and incorrect predictions for each class. This allows a deeper understanding of the strengths and limitations of each model [12].

The performance of the Naive Bayes, Decision Tree, and Random Forest classifiers is compared to determine which model performs best in classifying movie review sentiments. By analysing these metrics, the proposed model can select the most effective approach for delivering accurate sentiment predictions on new, unseen movie reviews (See Table 4).

Table 4: Accuracy Scores Report

Model	**Accuracy**
Naive Bayes	0.883721
Decision Tree	0.888372
Random Forest	0.939535

10.Conclusion

This research successfully developed a sentiment analysis system for movie reviews using machine learning and natural language processing techniques. The system effectively categorizes movie reviews into three sentiment categories: positive, neutral, and negative. By training the model on 2,000 pre-processed reviews sourced from Kaggle, the study ensures a balanced and robust foundation for sentiment prediction. Key text preprocessing techniques such as tokenization, stop word removal, stemming, and feature extraction were applied to enhance the model's classification accuracy. Several machine learning models, including Naive Bayes, Decision Tree, and Random Forest, were tested for sentiment prediction, with Random Forest yielding the most promising results in terms of accuracy and generalization. This system provides valuable insights for filmmakers, producers, and potential viewers, offering a real-time summary of public opinion on movies.

10.1 Future Objectives

The future direction of this research involves improving classification accuracy and expanding the system by incorporating more advanced sentiment analysis techniques. A potential enhancement could be the integration of deep learning models, such as recurrent neural networks (RNNs) or long short-term memory (LSTM) networks, which can better capture the context and nuances of movie reviews. Furthermore, expanding the system to handle reviews in multiple languages and from various movie genres could be achieved through the use of multilingual sentiment analysis models. Enhancing the system's sentiment granularity by distinguishing different levels of positivity or negativity will provide deeper insights into audience perceptions. Another goal is to incorporate real-time sentiment analysis from social media platforms to ensure that the system remains current and relevant. Lastly, the development of recommendation systems based on sentiment trends could further personalize the movie-watching experience for users.

REFERENCES

[1] R. K. Shah, S. Kumar, and N. Shashank, "Multilabel News Category Classification using Machine Learning," Conference: 2023 8th International Conference on Communication and Electronics Systems (ICCES), Jun. 2023, doi: 10.1109/icces57224.2023.10192826.

[2] IMDB Dataset of 50K Movie Reviews," Kaggle, Mar. 09, 2019. [Online]. Available: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews.

[3] L. Abu-Ghoush and Industrial and Systems Engineering, "An Integrated Framework Using Variable Encoding-TF-IDF-PCA-Classification for Predicting Adverse Event Action," thesis, Aug. 2024. [Online]. Available:

https://www.binghamton.edu/ssie/about/thesis_abstract_layan_abu-ghoush.pdf.

[4] R. Agrawal, M. Paprzycki, and N. Gupta, Big Data, IoT, and Machine Learning. 2020. doi: 10.1201/9780429322990.

[5] A.-A. Lina, H. Khadidja, J. Imene, and M. Neila, "Analyzing US Airline Customer Sentiment on Twitter using Multinomial Logistic Regression and Feature Reduction," Conference: 2023 7th IEEE Congress on Information Science and Technology (CiSt), vol. 3, pp. 265–270, Dec. 2023, doi: 10.1109/cist56084.2023.10409979.

[6] C. Kamusoko, Explainable Machine Learning for Geospatial Data Analysis. 2024. doi: 10.1201/9781003398257.

[7] C. S. Lima and A. Tavares, "Computer assisted diagnosis (CAD) with enhanced cognitive architectures (BorisCAD)," Jul. 26, 2023. https://hdl.handle.net/1822/87522.

[8] A. Shachar, "Introduction to Algogens," arXiv (Cornell University), Mar. 2024, doi: 10.48550/arxiv.2403.01426.

[9] B. Liu, Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press, 2020.

[10] C. Tekinbaş, "RISK MANAGEMENT BASED ON MACHINE LEARNING," thesis, Jan. 2024. [Online]. Available: https://open.metu.edu.tr/bitstream/handle/11511/108414/index.pdf.

[11] S. Satpathy, B. K. Paikaray, M. Yang, and A. Balakrishnan, Sustainable Farming through Machine Learning. 2024. doi: 10.1201/9781003484608.

[12] M. Messaoudi and H. Khouidmi, "Comparative Analysis of Machine Learning and Autoregressive Models for Forecasting Economic Growth: A Case Study," International Journal of Sustainable Development and Planning, vol. 19, no. 8, pp. 3049–3061, Aug. 2024, doi: 10.18280/ijsdp.190820.