Sentiment Analysis:

Optimizing Time Saving and Cost Efficiency in the Film Industry

August 2024

Summary

In today's digital era, sentiment analysis has become a crucial aspect of understanding public perception of film products. This project focuses on two main business metrics: **Time Saving** and **Cost Efficiency**. By applying machine learning analysis techniques to an IMDb film review dataset containing 50,000 entries, the goal of this project is to significantly optimize the film review analysis process.

The process begins with exploratory data analysis (EDA) to understand patterns and distributions within the dataset. Next, a pre-processing stage is conducted to prepare the data before entering the modeling phase. After these processes, the Naive Bayes model using TF-IDF techniques proved to be the most effective approach for classifying sentiment.

The analysis results indicate that **time savings can be optimized by 99%, and cost efficiency can also reach up to 99% compared to traditional analysis methods**. With these achievements, the project not only met but also exceeded the established targets.

Table of Contents

1. Background

In today's digital era, the entertainment industry, particularly filmmaking, has undergone a significant transformation in how films are produced, distributed, and received by audiences. One crucial aspect of this transformation is the role of viewer reviews and ratings in shaping a film's success. The Internet Movie Database (IMDb), founded in 1990, has become a leading platform providing comprehensive information about films, TV shows, and other video content (IMDb, 2024). With millions of active users worldwide, IMDb not only serves as a database but also as a forum where viewers can provide reviews and ratings for the films they have watched.

This phenomenon has created a vast and information-rich dataset that offers unique opportunities to understand the sentiments and preferences of film audiences. However, manually analyzing such a large volume of data is an almost impossible and inefficient task. This is where sentiment analysis, a branch of natural language processing (NLP), becomes highly relevant and important. Sentiment analysis, also known as opinion mining, is the process of identifying and extracting opinions, sentiments, and emotions from text (Liu, 2020). In the context of the film industry, sentiment analysis plays an increasingly vital role in various aspects, from predicting film success to developing content and more effective marketing strategies.

The importance of sentiment analysis in the film industry can be seen from several perspectives. First, audience sentiment has proven to be a strong indicator of a film's commercial success. Research conducted by Yu et al. (2020) shows a significant correlation between online review sentiment and box office revenue. This indicates that an accurate understanding of audience sentiment can provide valuable insights for film producers and distributors in predicting and maximizing revenue potential.

Second, sentiment analysis can provide valuable feedback for content development. By understanding audience preferences and emotional responses to various elements of a film, such as the plot, characters, or visual effects, filmmakers can develop content that aligns more closely with market tastes. Kim and Kim (2018) demonstrated how insights from sentiment analysis can be used to enhance the quality and relevance of film content, which in turn can improve audience satisfaction.

Third, in an era where online reputation is critical, sentiment analysis becomes an invaluable tool for reputation management. For film studios, actors, and other industry professionals, the ability to monitor and respond to public sentiment quickly and accurately is key to maintaining a positive image and managing potential crises. Chen et al. (2019) illustrated how real-time sentiment analysis can help entertainment companies identify and respond to negative feedback before it escalates into a larger issue.

Fourth, in an increasingly competitive marketing landscape, sentiment analysis enables more targeted and personalized marketing strategies. Wang et al. (2021) demonstrated how a deep understanding of audience preferences and sentiments can be used to design more effective and efficient marketing campaigns, enhancing the return on investment (ROI) in film marketing expenditures.

While the importance of sentiment analysis in the film industry is clear, the main challenge lies in how to analyze such large volumes of data with high accuracy and efficiency. Traditional approaches that rely on manual analysis or simple rule-based methods often struggle to handle the complexities and nuances of natural language in film reviews. Moreover, these approaches are also difficult to scale to accommodate the millions of reviews that continue to grow each day.

This is where machine learning approaches, particularly supervised learning, emerge as a promising solution. Unlike traditional methods, machine learning approaches have the ability to learn and adapt from data, enabling more accurate and efficient analysis at scale. Zhang et al.

(2018) demonstrated how machine learning models can significantly enhance the accuracy of sentiment analysis compared to lexicon-based or rule-based methods.

The advantages of machine learning approaches in sentiment analysis can be seen from several aspects. First, its superior scalability allows for the efficient handling of very large data volumes. For instance, implementing machine learning models offers significant business benefits in terms of time savings and cost efficiency. In manual sentiment analysis, the time required to analyze thousands of reviews can be substantial, averaging around 2 minutes per review (Devlin et al., 2019).

Second, the ability of machine learning models to adapt to changes in language and context over time is a significant advantage. Howard and Ruder (2018) demonstrated how transfer learning techniques can be used to fine-tune universal language models for specific tasks like sentiment classification, allowing models to remain relevant despite changes in language usage or trends. Third, consistently high accuracy across various benchmarks has established machine learning approaches as the de facto standard in modern sentiment analysis. Sun et al. (2019) showed how advanced models consistently outperform traditional methods in various sentiment analysis tasks.

In this context, the IMDb dataset emerges as a valuable resource for the development and testing of machine learning models for sentiment analysis. This dataset, which consists of 50,000 film reviews with binary sentiment labels (positive or negative), provides an ideal corpus for training and evaluating sentiment analysis models (Maas et al., 2011). The balance between positive and negative reviews, as well as the diversity in writing styles and language complexity, makes the IMDb dataset a highly relevant benchmark for testing the models' ability to handle sentiment nuances in the film domain.

Although there are many approaches to sentiment analysis using deep learning, such as BERT and RoBERTa, in the context of the current project, the use of simpler machine learning models like Logistic Regression and Multinomial Naive Bayes with Bag of Words and TF-IDF

techniques is sufficient. These methods allow for effective and efficient analysis, especially for large datasets like IMDb. With a supervised learning approach, models can be trained to identify patterns in the data and provide reasonably accurate results without the additional complexity often required by deep learning models. This approach not only reduces computational requirements but also enhances model interpretability, providing clearer insights into how audience sentiment can be analyzed.

In the context of the increasingly dominant streaming platforms, sentiment analysis can significantly contribute to enhancing recommendation systems and content personalization. Smith and Lee (2020) demonstrated how integrating sentiment analysis into recommendation algorithms can improve relevance and user satisfaction, which in turn can increase customer retention and platform revenue. Therefore, conducting sentiment analysis as an initial step in this data-driven approach is crucial, as it can provide valuable insights into audience preferences. By understanding audience sentiment toward films, platforms can develop more effective strategies to enhance user experience and optimize their content offerings in the future.

2. Goal

Sentiment Analysis: Optimizing Time Savings and Cost Efficiency in the Film Industry

The main objective of this project is to leverage machine learning technology to optimize sentiment analysis of film reviews. This research focuses on achieving significant improvements in terms of time savings and cost efficiency. Thus, this study aims to provide a more efficient solution for handling a large volume of reviews while offering a more economical and effective approach to sentiment analysis compared to traditional methods.

3. Objective

The film sentiment analysis project based on machine learning is designed to revolutionize the process of evaluating film reviews, with a focus on improving efficiency and accuracy. The main objectives of this project are:

1. **Optimize the speed of sentiment analysis** to enhance responsiveness to audience feedback.
2. **Increase cost efficiency** in the process of analyzing film reviews for better resource allocation.
3. **Maintain high accuracy and reliability** in sentiment classification to support sound decision-making in the film industry.

1. Matrix

Business Matrix

In machine learning-based film sentiment analysis, there are two main business metrics that are the focus: Time Savings and Cost Efficiency. These metrics are designed to measure the significant impact of implementing machine learning models in the analysis process compared to manual methods.

1. **Time Savings**

   - Definition: Time Savings refers to the reduction in time required to analyze film reviews using machine learning compared to manual methods.
   - Calculation: Time savings are calculated using the following formula:

     $$\textit{Time Savings} = (\textit{Manual Analysis Time per Review} - \textit{Machine Learning Analysis Time per Review}) \times \textit{Number of Reviews}$$

   - Target: In this project, the target is to achieve a 95% reduction in analysis time compared to manual methods.

2. **Cost Efficiency**
   - Definition: Cost Efficiency measures the reduction in operational costs achieved through the use of machine learning models in the sentiment analysis process.
   - Calculation: Cost efficiency is calculated using the following formula:

     $$(\textit{Business Cost per manual review} - \textit{Business Cost per machine learning review}) \times \textit{Number of reviews}$$

   - Target: The long-term goal for reducing operational costs is 80%.

Model Matrix

To support the achievement of time and cost savings targets, two evaluation metrics for the machine learning model used are Throughput and Accuracy.

- **Throughput/Runtime**: This metric measures the number of reviews that can be processed within a specific timeframe, providing a direct indication of the analysis speed. Throughput is essential for achieving significant time savings targets in the analysis.
- **Accuracy**: This metric is used to measure how well the model classifies reviews overall, both in the Bag of Words and TF-IDF approaches. Accuracy provides an overview of how reliable the model is in accurately predicting positive and negative sentiments.

With a combination of high throughput and good accuracy, the sentiment analysis process is expected to be faster and more efficient while still delivering relevant and high-quality results for the film industry.

5. About Dataset

The IMDb dataset consists of 50,000 rows of data that includes two main columns: "review," which contains the text of English-language movie reviews, and "sentiment," which indicates the sentiment label as positive (1) or negative (0). Here is a detailed explanation.

Table 1. Structure of the IMDB Dataset

No.	Column	Data Type	Description
1.	review	String/object	English-language movie review text
2.	Sentiment	Integer	The target column indicates sentiment labels as positive (1) or negative (0) based on the reviews.

6. *Exploratory Data Analysis & Pre-Processing*

Tabel 2. EDA & Pre-Processing

No.	Step	Description & Findings
1.	Check for Missing Values	There are no missing values in the dataset.
2.	Check for Duplicates	There are 418 duplicate rows, but they were not removed because removing them decreased the model's performance.
3.	Feature Engineering	
	1. review_length	This column contains the length of the text review in each row. The aim is to analyze whether the length of the review correlates with the sentiment given.
	2. review_length_binned	This column categorizes review lengths into several categories: 'Short', 'Medium', 'Long', 'Very Long', and 'Extreme' based on the length of the text review. These categories are used for further analysis.
4.	Check Sentiment Class	The distribution of classes between positive sentiment (1)

	Distribution	and negative sentiment (0) is balanced.
5.	Check Review Length Distribution	The data distribution is right-skewed, with most reviews having a length between 100-400 words. Outliers were detected but were not removed because their influence on the model is insignificant.
6.	Check Correlation Between Review Length and Sentiment	There is a significant difference in review lengths between positive and negative sentiments, but Cohen's d value is very small, indicating that the impact of this difference is not significant in practical terms.
7.	Removing HTML Strips and Noise Text	HTML elements and irrelevant or noise text were removed from the reviews.
8.	Removing Special Characters	Special characters were removed to clean the review texts.
9.	Text Stemming	Stemming was performed on the text to reduce words to their base forms, such as changing "running" to "run."
10.	Removing Stopwords	Stopwords were not removed because experiments showed that not removing stopwords improved model performance.
11.	Removed Features	Engineered features were removed because they did not significantly impact the model's performance in the experiments conducted.
12.	Labeling the Sentiment Text	Labeling the sentiment column as 1 for positive and 0 for negative (int).
13.	Split Data	Data was split into 70% for training and 30% for testing the model.
14.	Term Frequency-Inverse Document Frequency (TF-IDF)	TF-IDF was applied to convert the review texts into numerical vectors that consider the relative frequency of words in the reviews.
15.	Bag of Words Model	The Bag of Words model was used to model the text by counting the frequency of each word's occurrence without considering their order.

7. Fundamental Theories of Bag of Words and Term Frequency-Inverse Document Frequency (TF-IDF)

Bag of Words Model

Bag of Words, often abbreviated as BoW, is a simple feature extraction method in natural language processing (NLP). This technique converts text data into vectors that can be processed by computers. The concept of Bag of Words can be likened to a bag containing a collection of words from a text document. In this bag, the order or context of the words' appearance is not considered. The primary focus is on which words appear and how often each word occurs. That is why this technique is called "Bag of Words," because, like looking into a bag, what matters is what is inside, not the order of the objects (Rina, 2024).

Bag of Words (BoW) is used to extract features from text documents so that they can be applied to machine learning algorithms. In this approach, each word in the document is considered a feature, and its frequency of occurrence is counted. Each document is then represented as a vector, where each element of the vector indicates how often that word appears in the document. The BoW method is often applied in text classification using machine learning algorithms such as Naïve Bayes, Support Vector Machine, and Logistic Regression. If you want to see a simple practical example to understand it, click the notebook **here.** However, this method has several drawbacks, which are:

- **Does not consider word order.** BoW does not take into account the arrangement of words in a sentence or document. For example, the sentences "Ilham chases Rina" and "Rina chases Ilham" are considered the same by BoW because they contain the same words, even though their meanings are different.
- **Inefficiency in memory and computation.** BoW becomes inefficient when there are many unique words. Each document is represented as a very long vector, with most of its

elements being zero, creating a sparse matrix that consumes memory and computational power.

- **Does not differentiate between informative and common words.** BoW cannot distinguish between frequently occurring but less meaningful words (e.g., "and," "in," "that") and rarer but more informative words. To address this issue, methods like TF-IDF can be used, which will be discussed further.

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

As an alternative to address some of the weaknesses of the Bag of Words method, we can use the TF-IDF (Term Frequency — Inverse Document Frequency) method. This method is designed to assign greater weight to words that are more informative or significant within a document or corpus (Rina, 2024).

First, we will explain what Term Frequency, commonly known as TF, is. The way TF works is similar to the BoW method, as it calculates how often a word appears in a specific document. However, in the calculation of the TF value, we need to divide the number of occurrences of that word by the total number of words in the document. It is important to note that there are several approaches that can be used to calculate the TF-IDF value. However, in this context, we will use the simplest and most straightforward approach. Here is the formula used to calculate the TF (Term Frequency) value:

tf(t, d) = the number of times word t appears in document d / the total number of words in document d

Next, we will discuss what is meant by Inverse Document Frequency or IDF. As the name suggests, "inverse" means reversed. IDF serves to reduce the weight of words that appear frequently across documents and gives greater weight to words that appear rarely. For example, if in 1000 data points the word "interesting" appears in all movie reviews, then the word

"interesting" may not provide meaningful information due to its uniform occurrence. Here is the formula used to calculate the IDF value:

id(t) = log(Total number of documents in d / Number of documents in which word t appears)

Next, we will calculate the TF-IDF value. This value is obtained by multiplying TF and IDF using the formula below. To understand how this model works, you can access the notebook **here**.

$$tf - idf(\mathrm{t}, \mathrm{d}) = tf \times idf$$

Although the TF-IDF method is considered superior to the BoW method, it still has some weaknesses, namely:

- **More complex calculations**: The TF-IDF method requires more complicated calculations compared to the BoW method.
- **Ignore word order**: Like the BoW method, TF-IDF also does not take into account the order of words in the document.
- **Less effective for short documents**: The TF-IDF method is less suitable for very short documents, as the weight of the resulting words can become inaccurate.

8. *Modelling*

Table 3. Model Performance Evaluation Based on Accuracy and Runtime (Throughput)

No.	Model	Bag of Words Model	TF-IDF	Runtime Bag of Words	Runtime TF-IDF
1.	Logistic Regression	0.76 %	0.76 %	0.06 seconds	0.05 seconds
2.	SVM	0.56 %	0.50 %	0.06 seconds	0.06 seconds
3.	Multinomial Naive Bayes	0.76 %	0.76 %	0.10 seconds	0.10 seconds

Tabel 4. Confusion Matriks

No.	Model	Bag of Words Model	TF-IDF
1.	Logistic Regression	True Positive (TP): 5693 False Positive (FP): 1861 False Negative (FN): 1783 True Negative (TN): 5663	True Positive (TP): 5747 False Positive (FP): 1914 False Negative (FN): 1729 True Negative (TN): 5610
2.	SVM	True Positive (TP): 7428 False Positive (FP): 6610 False Negative (FN): 48 True Negative (TN): 914	True Positive (TP): 7476 False Positive (FP): 7524 False Negative (FN): 0 True Negative (TN): 0
3.	Multinomial Naive Bayes	True Positive (TP): 5618 False Positive (FP): 1772 False Negative (FN): 1858 True Negative (TN): 5752	True Positive (TP): 5665 False Positive (FP): 1819 False Negative (FN): 1811 True Negative (TN): 5705

- **Best Model:** Naive Bayes and Logistic Regression demonstrate balanced and better performance compared to SVM. However, Logistic Regression is slightly superior in terms of True Positives (TF-IDF) with 5747 TP compared to Naive Bayes in both BoW

and TF-IDF, and it has fewer False Negatives (FN). Additionally, the execution time of these three models is relatively fast, so there are no issues with runtime implementation.

- **Decision:** The primary focus is on balance and effectiveness in detecting positive classes; thus, Logistic Regression TF-IDF is the best choice among the three models.

9. Impact on Business

Table 5. Simulation of Model Impact on Business

Parameter	**Value**	**Total**	**Description**
Average time for analysis/review (Manual)	120 Seconds	50.000 x 120 = 6.000.000 6.000.000/3600 = 1.667.67 hour	Time saving can be optimized by 99%
Rata-rata waktu analysis/review (Model) :	1 microsecond	0.05 seconds	
Rata-rata biaya analisis review/jam (Manual)	3.3 $/hour	3.3 x 1.666.67 = 5.500.01 $	Cost Efficiency dapat dioptimalkan mencapai 99%
Rata-rata biaya analisis review/jam (Model)	3.3 $/hour	3.3/3600 = 0.00091 $/seconds 0.00091 x 0.5 = 0.000045 $	

10. Conclusion

The sentiment analysis project in the film industry has achieved significant success by leveraging machine learning techniques to understand public sentiment towards movie reviews. Focusing on two main business metrics, Time Saving and Cost Efficiency, has yielded highly satisfactory results. The findings indicate that time savings in the analysis process can be optimized by up to 99%, while cost efficiency also reaches 99%. These achievements demonstrate that the application of the right techniques can lead to outstanding results. Therefore, this project not only meets but also exceeds the established targets, making a meaningful contribution to the film industry in understanding and responding to the needs and preferences of audiences. This success opens up opportunities for further application of similar analysis methods in various other business contexts.

11. Limitation

One limitation of this project is the inadequate computational resources, which resulted in some techniques and experiments not being executed optimally. Additionally, the developed machine learning model has not yet been followed up with a deployment process, preventing it from being integrated into a broader recommendation system.

12. Langkah Selanjutnya

To enhance the results and performance of the model, the next steps will involve implementing more advanced deep learning approaches and gathering and processing higher-quality and more diverse data. Additionally, there will be an exploration of new techniques in natural language processing (NLP) and consideration for integrating the model into a more comprehensive recommendation system to provide more accurate and relevant recommendations for users.

13. Referensi

Chen, Y., Zhang, H., Liu, R., Ye, Z., & Lin, J. (2019). Experimental explorations on short text topic mining between LDA and NMF based Schemes. Knowledge-Based Systems, 163, 1-13.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171-4186.

Garcia, D., & Ráez, A. (2019). Coping with the long tail: Hybrid approaches to manage and analyze big text collections. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 3215-3216.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 328-339.

IMDb. (2024). About IMDb. Retrieved from https://www.imdb.com/about/

Kim, S. M., & Kim, H. J. (2018). Sentiment classification of movie reviews using feature selection based on dynamic λ-measure. Applied Intelligence, 48(5), 1268-1285.

Liu, B. (2020). Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, 142-150.

Rina. (2024). Mengenal Bag of Words pada Model NLP. Medium. Retrieved from https://esairina.medium.com/mengenal-bag-of-words-pada-model-nlp-4013ec879e26

Rina. (2024). Mengenal Term Frequency-Inverse Document Frequency (TF-IDF) pada Model NLP. Medium. https://esairina.medium.com/mengenal-term-frequency-inverse-document-frequency-tf-idf-pada-model-nlp-e0cc571f7e37

Smith, J., & Lee, K. (2020). Personalized content recommendation in streaming platforms using deep learning-based sentiment analysis. IEEE Transactions on Multimedia, 22(3), 625-637.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification?. In China National Conference on Chinese Computational Linguistics (pp. 194-206). Springer, Cham.

Wang, Y., Wang, M., & Xu, W. (2021). A sentiment-enhanced hybrid recommender system for movie recommendation: A big data analytics framework. Wireless Communications and Mobile Computing, 2021.

Yu, X., Liu, Y., Huang, X., & An, A. (2020). Mining online reviews for predicting sales performance: A case study in the movie domain. IEEE Transactions on Knowledge and Data Engineering, 24(4), 720-734.

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.