Sentiment Analysis of Movie Reviews

Yuna Lee

1 Problem statement and Motivation

Online movie review platforms offer an effective channel for users to share their opinions on movies. As movies become more accessible, these platforms play a crucial role in guiding decision-making by offering insights aligned with viewers' preferences. Moreover, companies can leverage these reviews to understand customer sentiments towards their projects. However, the sheer volume of reviews available on the social web poses a challenge, as they are not easily processable by machines (Tetteh and Thushara, 2023).

To overcome this challenge, our study focuses on employing sentiment analysis on movie reviews using the Rotten Tomatoes dataset from Hugging Face. By determining whether reviews express positive or negative sentiments, we aim to assist both individual users and movie businesses in gauging public opinion accurately.

This initiative shapes our research question: How do various feature extraction techniques and classification models impact the accuracy and efficiency of supervised learning models in determining sentiments conveyed in movie reviews? Through this exploration, we aim to showcase the effectiveness of different techniques in achieving high performance in analysing sentiments conveyed in movie reviews.

2 Research hypothesis

Our research hypothesis assumes that a transformer-based model, specifically the Bidirectional Encoder Representations from Transformers (BERT) will achieve the highest accuracy in sentiment analysis of movie reviews. Developed by Google, BERT has significantly enhanced natural language processing tasks by employing a deep neural network architecture trained on extensive unannotated text data to create contextually aware word representations (Putrada, Alamsyah and Fauzan, 2023). We believe that BERT's transformer architecture, coupled with its ability to capture intricate relationships within text, enables it to comprehend word context effectively, thus delivering superior results for this task.

To validate our hypothesis, we conduct experiments with various feature extraction techniques and classification models to compare their performance and identify the most effective approach. We compare the accuracy of our proposed model with those achieved in studies using alternative learning approaches, such as Support Vector Machines, Logistic Regression and Naïve Bayes, employing different feature vectorizers like Bag of Words, N-grams, Term Frequency – Inverse Document Frequency, and Word2Vec. Additionally, we plan to construct sentiment analysis using a lightweight pre-trained model called DistilBERT, a transformer-based model, and compare the results with those obtained from standard machine learning models. Through a series of experiments, we test that a better selection of variants often outperforms the recently published state-of-the-art. By doing so, we aim to gain insights that address our original research question.

3 Related work and background

Sentiment analysis methodologies fall into two main categories: lexical-based and machine learning-based approaches. While lexical methods use predefined sentiment dictionaries to discern positive and negative sentiments, they struggle with polarity identification beyond specific domains due to contextual limitations (Taboada et al., 2011). Consequently, studies often turn to traditional machine learning classification models. A comprehensive survey (Teja et al., 2018) evaluating various sentiment analysis approaches highlighted NB and SVM as effective benchmarks. Drawing inspiration from these established methodologies, we adopt NB and SVM models as benchmarks in our study.

Previous studies have underscored the significance of exploring alternative algorithms and pre-processing methods in sentiment analysis. For example, Samsir et al. (2022) revealed that data without lemmatisation outperformed lemmatised data in NB classification. Consequently, we prioritise comparing pre-processing methods such as stop word removal and lemmatisation to assess their impact. Additionally, Das and Chakraborty (2018) proposed an alternative approach employing TF-IDF with Next Word Negation (NWN) for text classification. With TF-IDF and NWN, they reversed the polarity of the next word following a negation word. Subsequently, they evaluated these techniques with three algorithms and concluded that SVM yielded the most favourable results. Their study revealed a notable enhancement in accuracy levels with the incorporation of NWN and TF-IDF, providing valuable insights for our TF-IDF extraction technique.

In a recent study, Başarslan and Kayaalp (2023) combined W2V embeddings with NB for sentiment analysis. They utilised a skip-gram model to construct a W2V model based on 40,000 samples from Rotten Tomatoes, exploring various scenarios, including stemming impact, word2vec dimensionality, and NB variants. The study found that MNB achieved high accuracy without stemming and with 300 dimensions. The study suggests further research on feature combinations and alternative classification models, highlighting the importance of leveraging W2V embeddings effectively. Consequently, our study will explore multiple scenarios, including additional pre-processing steps and alternative models.

Jain et al. (2017) compared various ML methods for sentiment analysis on the Rotten Tomatoes dataset. Combinations of BoW and N-grams (Unigram, Bigram, Unigram + Bigram, Unigram + Bigram + Trigram) with LR, MNB and SVM. LR outperformed MNB and SVM for binary sentiment classification, especially with the combination of Unigram, Bigram, and Trigram features. However, for multi-class tasks and longer documents, Convolutional Neural Networks (CNNs) are preferred. Since our study focuses on binary sentiment analysis with short text reviews, we do not utilise CNNs.

In a separate study (Angelard-Gontier, 2022), the accuracy of two algorithms in classifying movie review sentiment was compared: a Random Forest (RF) classifier with BoW and W2Vs features, and a Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN). While the LSTM network marginally outperformed the RF classifier, the improvement was minimal, particularly for short reviews. This suggests that complex time-series models like RNNs may not offer significant advantages for sentiment analysis tasks with brief texts. Therefore, we opt not to employ such complex time-series models in our study.

We also recognised the potential of transformer-based models through various studies that compared their performances against those of traditional models. In a study proposing pre-trained advanced language models XLNet and BERT (Danyal et al., 2024), they compared them with conventional ML models using TF-IDF. The results indicate that XLNet achieved the highest accuracy on both datasets, demonstrating the capability of deep learning models to extract complex emotional information from textual data. Another study (Putrada, Alamsyah and Fauzan, 2023) used the lightweight pre-trained DistilBERT model and compared the performance with benchmark models: SVM, NB, with TF-IDF. They found that DistilBERT performed best, highlighting its advantages in terms of accuracy and efficiency. From the studies comparing machine learning models using TF-IDF with BERT methods, we consider TF-IDF as a good baseline.

The literature mentioned showcases various effective techniques in sentiment analysis, from pre-processing to classification. Our study aims to explore and combine these methods to enhance sentiment analysis of movie reviews.

4 Accomplishments

- Task 1: Pre-process the dataset to a suitable format – Completed.
- Task 2: Experiment with additional pre-processing of the dataset: removing stop

words and applying lemmatisation - Completed.

- Task 3: Tokenise the dataset – Completed.
- Task 4: Vectorise test data using TF-IDF and TF-IDF with NWN – Completed.
- Task 5: Build and train MNB, SVM and LR models using TF-IDF and TF-IDF with NWN and examine their performance (with and without additional pre-processing) – Completed.
- Task 6: Train a W2V model – Completed.
- Task 7: Build and train MNB, LR and SVM models on a dataset using the trained W2V model and examine their performance (with and without additional pre-processing) – Completed.
- Task 8: Experiment with the use of BoW and different combinations of N-grams – Completed.
- Task 9: Build and train LR and SVM models on a dataset using N-grams and examine their performance (with and without additional pre-processing) – Completed.
- Task 10: Implement a variation of a DistilBERT transformer – Completed.
- Task 11: Assess whether the DistilBERT is the best classifier out there for this sentiment analysis task – Completed.
- Task 12: Perform an in-depth analysis to figure out which kinds of examples each approach struggles with – Completed.

5 Approach and Methodology

Our methodology for sentiment analysis integrates diverse approaches, drawing from existing research to construct a comprehensive pipeline.

Our baseline model employs basic TF-IDF with NB yielding an accuracy of 0.7758. However, we acknowledge the limitations inherent in this approach, particularly its inability to capture word order and context, and the independence assumption made by Naïve Bayes. To mitigate these limitations, we experiment with additional pre-processing steps, such as stop word removal and lemmatisation. Studies (Danyal et al., 2024) have debated the necessity of stop word removal, considering the presence of negation words within stop words, impacting sentiment conveyance. In addition, we are exploring lemmatisation as a preferred technique over stemming, as highlighted in studies (Putrada, Alamsyan and Fauzan, 2023). Unlike stemming, lemmatisation considers the context and grammatical structure of words, resulting in more accurate transformations that retain the semantic value of the original word.

In our approach, we systematically explore various machine learning approaches and feature combinations using 10-fold cross-validation with grid search for hyperparameter optimisation. This ensures robust model selection for each feature extraction method. Subsequent sections outline our machine learning methodologies.

TF-IDF - In a previous study by Das and Chakraborty (2018), combining next word negation (NWN) with TF-IDF was found to improve model performance. We begin by applying basic TF-IDF, then introduce NWN and evaluate the outcomes. Additionally, we examine the impact of stop word removal on negation sentiments like 'no' and 'not'. While our baseline model employs basic TF-IDF with NB, we compare this baseline with other models using TF-IDF and TF-IDF with NWN, employing SVM, LR, and NB with various pre-processing methods.

Word2Vec - Building on the insights from Başarslan and Kayaalp (2023), we explore various scenarios using W2W embeddings. We investigate how W2W feature extraction affects model performance and explore additional pre-processing methods, including stop word removal and lemmatisation. Additionally, we aim to create our own W2W corpus. Our analysis compares different models, NB, SVM and LR, to evaluate their performance against TF-IDF feature extraction methods. Given W2W's ability to grasp the contextual meaning of words and capture semantic relationships effectively, it holds promise for enhancing sentiment analysis tasks.

N-grams - Additionally, we explore N-grams by employing various combinations of feature combinations, such as BoW and different combinations of N-grams (Unigram, Bigram, Unigram + Bigram, Unigram + Bigram + Trigram) with LR, MNB, and SVM, while also investigating further pre-processing methods. N-grams are adept at capturing local word order effectively representing the syntactic structure, which is advantageous for our analysis.

BERT - For our BERT model, we're using the DistilBERT architecture, a lightweight version of BERT that maintains much of its power while being more computationally efficient. By leveraging DistilBERT, we benefit from BERT's advanced capabilities in understanding context and representing semantics without the computational overhead. We optimise hyperparameters using the validation set and evaluate performance on the test set. Additionally, we fine-tune the DistilBERT model on our dataset to tailor it to the specifics of movie review sentiment analysis.

For implementation, we utilise a range of libraries including pandas, matplotlib, nltk, scikit-learn, gensim, transformers, and PyTorch. The uploaded files contain the best-performing TF-IDF and BERT models, showcasing the effectiveness of our methodology.

6 Dataset

The study uses the ‘Movie Review Dataset’ (Pang and Lee, 2005) obtained from Hugging Face, comprising 10,662 text reviews extracted from Rotten Tomatoes. Each review is labelled with '1' for positive and '0' for negative sentiment, with equal distribution. The dataset is split into 80% training, 10% testing, and 10% validation subsets, totalling 8,530, 1,066, and 1,066 rows, respectively. Figure 1 illustrates the balanced distribution of the dataset across these subsets. There are no missing values.

Figure 2 displays boxplots of word counts per review based on their labels. On average, both positive and negative reviews contain around 21 words, with some exceeding 50. The distribution of word counts is comparable between labels, indicating similar review lengths. While most reviews are manageable, longer ones exist. Examples of such reviews are shown in the top two lines of Figure 3.

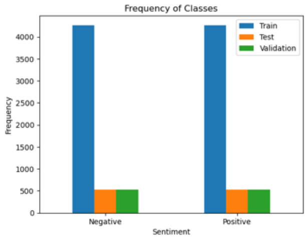


Figure 1 – Bar plot displaying the frequency across different sentiments.

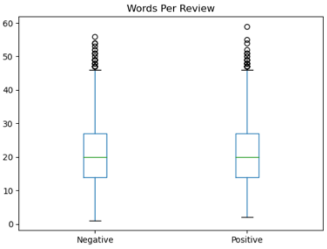
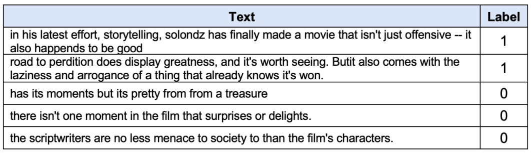


Figure 2 – Box plot displaying the distribution of the number of words per review across different sentiments.

Additionally, some reviews exhibit sarcasm, irony, or ambiguity, posing challenges to classification accuracy. Thus, ensuring models can correctly classify such examples is crucial for reliable performance evaluation.



Text	Label
in his latest effort, storytelling, solondz has finally made a movie that isn't just offensive -- it also happends to be good	1
road to perdition does display greatness, and it's worth seeing. Butit also comes with the laziness and arrogance of a thing that already knows it's won.	1
has its moments but its pretty from from a treasure	0
there isn't one moment in the film that surprises or delights.	0
the scriptwriters are no less menace to society to than the film's characters.	0

Figure 3 – Examples of text reviews with their corresponding labels.

We analysed review content by creating word clouds for Positive and Negative sentiments (Figure 4). Common words like 'film', 'movie', 'one', and 'make' appear in both. Positive reviews feature words like 'good', 'comedy', 'funny', while negative reviews include terms such as 'bad', 'doesn't', and 'never'. It is essential to recognise that negative reviews often include negation words, which can reverse sentiment polarity and affect accuracy. Failing to address these negation words can lead to inaccuracies in sentiment classification.

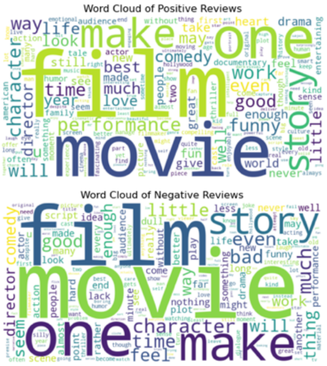


Figure 4 – Word clouds for positive (above) and negative (below) sentiments.

6.1 Dataset preprocessing

To construct machine learning models with different feature extraction methods, we combined validation and test data for an 80:20 split for training and testing, employing 10-fold cross-validation. A general function was developed to pre-process the dataset by removing HTML content, punctuation, and special characters, and converting text to lowercase, aiming to standardise the data and reduce noise. Additionally, another function was created to explore the effects of further pre-processing by removing stop words and performing lemmatisation.

For the DistilBERT model, no pre-processing was applied due to its ability to learn effectively from unaltered data. However, the text data was tokenized using the DistilBERT tokenizer, with sequences padded to a maximum length of 500 tokens and mapped to PyTorch datasets for model compatibility.

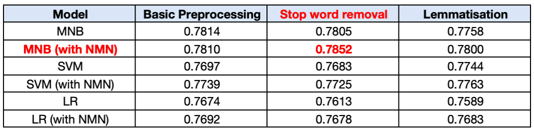
7 Baselines

We considered several baselines to guide our study, aiming to provide benchmarks for evaluating more complex models. As mentioned earlier, it has been established that NB and SVM are commonly used for text categorisation tasks (Teja et al., 2018). Additionally, we observed from studies comparing BERT models with traditional machine learning models that features extracted using TF-IDF have been effective. Hence, we chose TF-IDF representation combined with the NB classifier as our baseline. This baseline incorporates a widely adopted pre-processing technique, where stop words are removed, and lemmatisation is applied.

This choice was influenced by the computational efficiency of the approach compared to SVM. Despite its simplicity, this baseline achieved reasonable performance metrics (accuracy and recall rate both at 0.7758 – further details in Supplementary material). It provides a starting point for evaluating more advanced models and allows us to identify and address limitations to improve natural language processing system performance.

8 Results, error analysis

1. We expanded our model experiments using a TF-IDF vectorizer, incorporating NWN to enhance performance. MNB, SVM, and LR models were employed, with variations in pre-processing.



Model	Basic Preprocessing	Stop word removal	Lemmatisation
MNB	0.7814	0.7805	0.7758
MNB (with NMN)	0.7810	**0.7852**	0.7800
SVM	0.7697	0.7683	0.7744
SVM (with NMN)	0.7739	0.7725	0.7763
LR	0.7674	0.7613	0.7589
LR (with NMN)	0.7692	0.7678	0.7683

Table 1– Accuracy scores for MNB, SVM and LR models with and without NMN, with variations in pre-processing.

Table 1 presents the accuracy scores of all combinations. Notably, the MNB model exhibited the highest performance, improving from the baseline accuracy (0.7758). However, for SVM and LR models, the scores were either similar or lower than the baseline, suggesting that MNB is the most suitable for this task.

Comparing TF-IDF with and without NWN, promising improvements were observed with NWN across models. Particularly, MNB with NWN, after stop word removal, outperformed the basic TF-IDF. Incorporating NWN with TF-IDF notably boosted performance for the MNB model, resulting in the highest accuracy of 0.7852 after stop word removal.

Consequently, we identified the MNB model employing TF-IDF with NWN, after basic pre-processing and stop word removal, as the best-performing model. The hyperparameters of this model were optimised through 10-fold cross-validation with grid search, resulting in a smoothing parameter ‘alpha’ of 1 and ‘fit_prior’ set as True. Figure 5 illustrates the confusion matrix of this selected model.

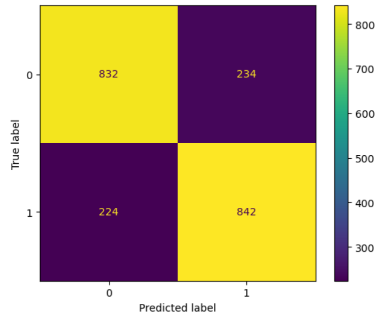


Figure 5– Confusion matrix of the MNB model employing TF-IDF with NMN pre-processed with stop word removal.

Examples illustrating Baseline Failures and Model Success

- “too much power, not enough puff.” – Correctly classified as Negative.
- “the uneven movie does have its charms and its funny moments but not quite enough of them.” – Correctly classified as Negative.

Our model effectively handles texts containing negation words, benefiting from TF-IDF with NWN. This approach not only detects negation words but also evaluates the subsequent word's impact on sentiment, crucial for nuanced sentiment analysis. The model captures negation contexts that would otherwise be overlooked if stop words were removed. Baseline models struggle to capture negation, as evident in these examples. Further details and a processed text example are available in Supplementary Materials.

Misclassified Examples of this Model

- “it's like rocky and bullwinkle on speed, but that's neither completely enlightening, nor does it catch the intensity of the movie's strangeness.” – Misclassified as Positive.
- “by the time it ends in a rush of sequins, flashbulbs, blaring brass and back-stabbing babes, it has said plenty about how show business has infiltrated every corner of society and not always for the better – Misclassified as Positive.”

However, the model does not always successfully classify models correctly. Despite its capability to recognise nuances such as negation, such as 'not always better', it still misclassifies some text. This could be due to the integration of NWN, which requires adjustments to existing sentiment analysis models to comprehend the connection between negation words and subsequent words, increasing model complexity and potentially leading to challenges when classifying longer texts.

2. Next, we explored the Word2Vec on models like MNB, SVM, and LR, focusing on stop word removal and lemmatisation instead of stemming. Additionally, we sought to compare these models with TF-IDF-based approaches.

We developed our own W2V model by training on the entire available Rotten Tomatoes training dataset, rather than using a pre-trained model like Google's, which was too memory intensive. Training our model on the dataset helped reduce computational demands and allowed us to customise parameters like vector dimensions, window size, and training epochs to better fit our task. Furthermore, as the model was trained on the same domain as the test data, we hoped it would generalise well compared to other domains.

Model	Basic Preprocessing	Stop word removal	Lemmatisation
MNB	0.6918	0.7054	0.7045
SVM	0.7294	0.7322	0.7303
LR	0.7190	0.7219	0.7233

Table 2– Accuracy scores for MNB, SVM and LR models employing W2V with variations in pre-processing.

Table 2 presents accuracy scores for all model and pre-processing method combinations. Surprisingly, we observed performance improvements across all models with additional pre-processing, especially after stop word removal, contrary to previous findings where stemming reduced performance.

Among models, MNB demonstrated the highest performance, particularly with basic pre-processing and stop word removal, achieving an accuracy and recall score of 0.7322, which was still lower than the baseline model's (0.7758) performance. Additionally, TF-IDF outperformed W2V when comparing extractors. The lower performance of W2V compared to the baseline could be attributed to training our own model on the Rotten Tomatoes dataset, which may have lacked the necessary diversity to capture the full range of language patterns and contexts present in the data. Utilising a larger pre-trained W2V model may improve results.

Furthermore, it's worth mentioning that the hyperparameters of the best-performing model were optimised through 10-fold cross-validation with grid search, resulting in the selection of C = 1, gamma of 0.1, with a 'rbf' kernel. The confidence matrix can be found in Figure 6.

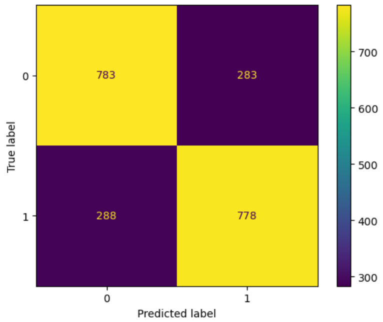


Figure 6– Confusion matrix of the SVM model employing W2V.

While the baseline performed better overall, we examine examples where our model successfully classified sentiment while the baseline failed.

Examples Illustrating Baseline Failures and Model Success

- "the uneven movie does have its charms and its funny moments but not quite enough of them" – Correctly classified as Negative.
- "an energetic and engaging film that never pretends to be something it isn't." – Correctly classified as Positive.
- "a thriller without a lot of thrills." - Correctly classified as Negative.

The model demonstrates proficiency in correctly classifying sentiment in text segments that contain clear and unambiguous language. It excels in identifying positive or negative connotations based on straightforward expressions and context, such as "not quite enough of them" or "never pretends to be something it isn't". However, the model misclassifies examples which have layers of ambiguity, irony, or sarcasm.

Misclassified Examples of this Model

- "anyone who suffers through this film deserves, at the very least, a big box of consolation candy." – Misclassified as Positive.
- "an enormously entertaining movie, like nothing we've ever seen before, and yet completely familiar." – Misclassified as Negative.
- "the film truly does rescue [the funk brothers] from Motown's shadows. It's about time." – Misclassified as Negative.

The model faces challenges in interpreting irony and sarcasm due to its reliance on word embeddings without a broader context. Nuanced expressions like "it's about time" are difficult for W2V to capture accurately, especially without sufficient examples of sarcastic language in the training data.

W2V excels in unambiguous contexts but may struggle with ambiguity or nuanced language, leading to misclassifications. This emphasises the need to consider context and understand language intricacies when analysing sentiment with W2V models.

3. We examined various BoW and N-grams methods on LR and SVM models. Unlike a previous study, we omitted MNB to focus on the effects of different models. Additionally, we explored additional pre-processing techniques such as stop word removal and lemmatisation.

We initially employed LR for its computational efficiency and assessed the accuracy scores of various N-gram combinations with different pre-processing methods. The results are provided in the Supplementary material, revealing that Unigram + Bigram consistently outperformed other combinations across all pre-processing methods, particularly with basic pre-processing yielding superior results.

As LR's performance fell below the baseline, we conducted experiments with SVM using Unigram + Bigram. Table 3 illustrates improved performance compared to LR, especially with basic pre-processing yielding the best results (0.7716). Despite the improvement, the accuracy remained lower than the baseline but comparable (0.7758). We further refined the SVM model's hyperparameters through 10-fold cross-validation with grid search, selecting C = 0.1, gamma of 0.1, and employing a 'linear' kernel. The confidence matrix can be found in Figure 7.

While the Unigram + Bigram approach with SVM showed potential, it fell short of the baseline. Nevertheless, here are examples where our model succeeded while the baseline failed.

Examples Illustrating Baseline Failures and Model Success

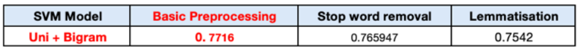
- "at its worst, the movie is pretty diverting." – Correctly classified as Positive.
- "in the end, Tuck Everlasting falls victim to that everlasting conundrum experienced by every human who ever lived: too much to do, too little time to do it in." - Correctly classified as Negative.
- "about as original as a gangster sweating bullets while worrying about a contract on his life." – Correctly classified as Negative.

The model accurately classifies these examples correctly by considering both individual words and pairs of adjacent words (bigrams) in the text, enabling it to capture more nuanced language patterns and contextual cues. The model identifies sentiments, such as "falls victim" and "sweating bullets," which would not have been captured by TF-IDF.

Misclassified Examples of this Model

- Director Hoffman, his writer, and Kline's agent should serve detention." – Misclassified as Negative.
- "this is a children's film in the truest sense. it's packed with adventure and a worthwhile environmental message, so it's great for the kids parents, on the other hand, will be ahead of the plot at all times , and there isn't enough clever innuendo to fil" – Misclassified as Positive.

The model struggle to accurately classify these examples due to its mixed sentiment. It may focus on individual words or pairs of consecutive words without fully capturing the overall context and nuanced language patterns. As a result, it could potentially misclassify this sentence. In addition, a previous study (Jain et al., 2017) identified that the models were not able to identify lengthy examples, which seems to be the case here.



SVM Model	Basic Preprocessing	Stop word removal	Lemmatisation
Uni + Bigram	0. 7716	0.765947	0.7542

Table 3 - Accuracy scores for SVM model employing Unigram + Bigram with variations in pre-processing.

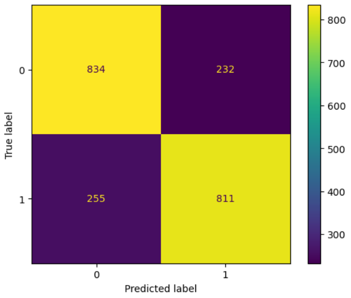


Figure 7 - Confusion matrix of SVM model employing Uni + Bigram.

4. Lastly, we implemented the DistilBERT – transformer-based neural network using PyTorch. DistilBERT is smaller and faster to implement and promises high performance for general applications.

We constructed a model architecture by incorporating a pre-trained DistilBERT model, with additional layers for binary classification, including a dropout layer and linear layers. Text data is tokenized using the DistilBERT tokenizer and mapped to PyTorch datasets for training and evaluation. The model is trained using the AdamW optimizer with a learning rate of 5e-4 and a linear scheduler. Cross-entropy is employed as the loss function, and the model is trained over multiple epochs, with validation accuracy computed to save the best model.

We opted not to preprocess the dataset, directly feeding the training and validation data to the pretrained DistilBERT tokenizer. The choice of the AdamW optimizer was due to its weight decay capabilities, aiding in preventing overfitting.

Upon testing with the test set, the model achieved a test accuracy of 0.827, marking the highest accuracy obtained from the models thus far. However, we should still examine examples for further insight.

Misclassified Examples of this Model

- "if we sometimes need comforting fantasies about mental illness , we also need movies like tim mccann's revolution no. 9." – Misclassified as Negative.
- “the acting in pauline and paulette is good all round , but what really sets the film apart is debrauwer's refusal to push the easy emotional buttons” - Misclassified as Positive.
- "rather than real figures, elling and kjell bjarne become symbolic characters whose actions are supposed to relate something about the naïf's encounter with the world.” - Misclassified as Positive.

We see that the model struggles to discern the underlying critical tone of the statement, highlighting the challenge of understanding subtle linguistic cues that convey nuanced opinions. While transformers, like the DistilBERT model used here, excel at capturing local context within a given window size, they may struggle with understanding broader contextual nuances or world knowledge required for tasks like common-sense reasoning or understanding humour.

9 Lessons learned and conclusions

In summary, our study delved into various machine learning models and techniques for sentiment analysis, encompassing both traditional methodologies and cutting-edge transformer-based models. As anticipated in our initial hypothesis, the top-performing model emerged as the DistilBERT model. We established a robust baseline accuracy using the TF-IDF Vectorizer in conjunction with MNB and subsequently advanced this baseline by incorporating next word negation, resulting in the second-best performing model.

Our focus was primarily on investigating different pre-processing methods, revealing their significant impact on performance. This underscores the importance of tailoring pre-processing techniques to suit specific tasks. Notably, the integration of negation word handling notably boosted model performance, particularly when data stop words were removed. Nevertheless, despite these enhancements, models still encountered challenges with longer texts and mixed sentiments.

Furthermore, all models struggled to some extent with ambiguity and sarcasm, indicating the necessity for further research in addressing these linguistic complexities. To enhance results, leveraging pretrained W2V models could be explored, given the availability of resources.

In conclusion, the methodologies explored in our study hold promise for enhancing sentiment analysis on movie reviews, which can be invaluable to businesses and consumers alike. While transformer-based models show potential for achieving high accuracy, they may require additional refinements to effectively capture nuanced language nuances and contextual cues. Future research efforts could concentrate on refining transformer models using larger and more varied datasets and crafting specialized preprocessing techniques to address complex linguistic phenomena more effectively.

References

Angelard-Gontier N. (2022). Sentiment Analysis on Movie Reviews.

Başarslan, M. S., & Kayaalp, F. (2023). Sentiment analysis with ensemble and machine learning methods in multi-domain datasets. Turkish Journal of Engineering, 7(2), 141-148.

Danyal, M. M., Khan, S. S., Khan, M., Ullah, S., Mehmood, F., & Ali, I. (2024). Proposing sentiment analysis model based on BERT and XLNet for movie reviews. Multimedia Tools and Applications, 1-25.

Das, B., & Chakraborty, S. (2018). An improved text sentiment classification model using TF-IDF and next word negation. arXiv preprint arXiv:1806.06407

Jain, S., Malviya, S., Mishra, R., & Tiwary, U. S. (2017, December). Sentiment analysis: An empirical comparative study of various machine learning approaches. In Proceedings of the 14th

International Conference on Natural Language Processing (ICON-2017) (pp. 112-121).

Nguyen, D. Q., Vu, T., & Pham, S. B. (2014, June). Sentiment classification on polarity reviews: an empirical study using rating-based features. In Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis (pp. 128-135).

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. arXiv preprint cs/0506075.

Putrada, A. G., Alamsyah, N., & Fauzan, M. N. (2023, August). BERT for Sentiment Analysis on Rotten Tomatoes Reviews. In 2023 International Conference on Data Science and Its Applications (ICoDSA) (pp. 111-116). IEEE.

Samsir, S., Kusmanto, K., Dalimunthe, A. H., Aditiya, R., & Watrianthos, R. (2022). Implementation naïve bayes classification for sentiment analysis on internet movie database. Building of Informatics, Technology and Science (BITS), 4(1), 1-6.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.

Teja, J. S., Sai, G. K., Kumar, M. D., & Manikandan, R. (2018). Sentiment analysis of movie reviews using machine learning algorithms-a survey. International Journal of Pure and Applied Mathematics, 118(20), 3277-3284.

Tetteh, M., & Thushara, M. (2023, May). Sentiment Analysis Tools for Movie Review Evaluation-A Survey. In 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 816-823). IEEE.