Exploring the Comparative Effectiveness of a Rule Based Approach vs. Machine Learning to Assess Viewer Sentiment towards IMDB Movie Reviews

Ana Melchor Pérez
7600259
Seminar group 3, Dr. Jing Zeng

Index

I Introduction

Sentiment analysis, crucial for assessing public sentiment on specific topics, faces challenges in movie reviews due to variability, subjectivity, and domain-specific language. Platforms like IMDb offer valuable datasets, yet accurate machine analysis is hindered by these complexities. Sentiment analysis is categorised into rule-based and machine learning (ML) methods, using baseline models for approaches. The former employs explicit linguistic rules, offering simplicity with predefined rules. The latter utilizes data-driven models, providing adaptability through algorithms for nuanced sentiment classification (Ray, 2019). This study evaluates the effectiveness of sentiment analysis methods on the IMDb dataset through text classification experiments. It emphasizes clarity in execution to explore sentiment analysis nuances. The paper examines literature, outlines methodology, presents results, and concludes, offering insights for comparing both models and guiding future research.

II Literature review

Sentiment Analysis, a facet of Natural Language Processing (NLP) or Opinion Mining, quantifies people's feelings in text (Liachoudi, 2020). It involves the categorization of text into negative, neutral, or positive sentences, providing valuable insights into emotions, opinions, and attitudes. Diverse stakeholders, academia, industry, and government (Alqaryouti, 2019), are contributing to recent developments such as fine-grained analysis (Munikar, 2019), contextualized embeddings (Peters, 2018), and multimodal approaches (Poria, 2019).

II.I Rule-based

Rule-based sentiment analysis relies on predefined linguistic rules to analyze text sentiment. In the context of movie sentiment analysis, explicit criteria such as keyword matching, negation handling, and sentence structure rules are employed. This approach captures nuances and context, enhancing the accuracy of sentiment classification (Onalaja, 2021). Lexicon-based methods, a type rule-based approach, use pre-established sentiment dictionaries but struggle with negations, novel words, and nuanced comprehension compared to advanced ML models, despite being accessible and capable of identifying mixed emotions without labeled data (Taboada, 2011).

II.II Machine Learning

ML in sentiment analysis uses system-defined rules for effective negation detection and context understanding. ML models analyse language to computationally grasp emotions and attitudes.

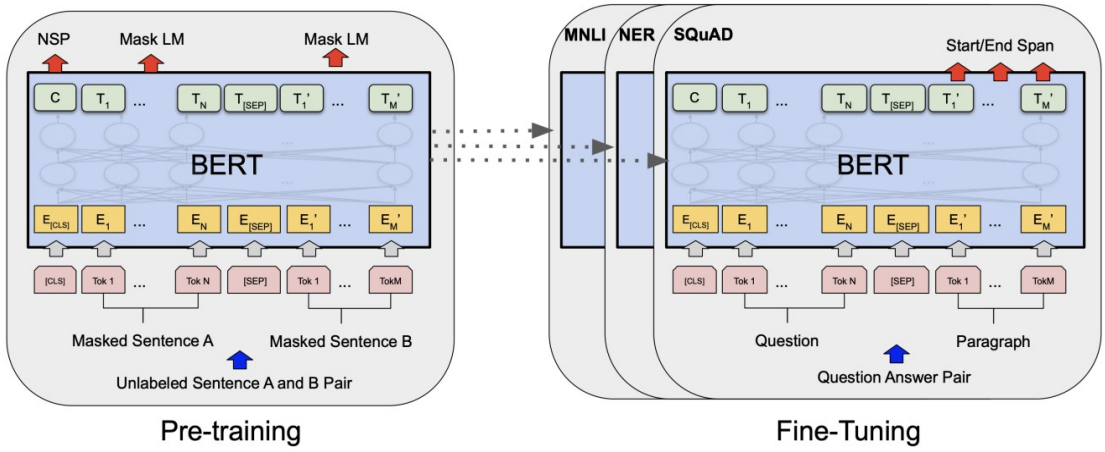
Two notable unsupervised learning models include the eXtreme Multi-Label Learning Network (XLNet), a pre-trained model focusing on complex relationship handling in textual data (Alduailej, 2022), and Implicit Target-Based Pre-training (ITPT), enhancing sentiment analysis through predicting implicit targets (Li, 2021).

Within supervised learning models, Bidirectional Encoder Representations from Transformers (BERT) excels in contextual understanding through bidirectional transformers, comprehending context in pre-trained sequential data (Figure 1) (Batra, 2021). Universal Sentence Encoder (USE) transforms text into fixed-length vectors, enhancing emotion detection and classification accuracy by capturing semantic meaning (Cer, 2018). Long Short-Term Memory (LSTM), a recurrent neural network, enhances emotion analysis (Murthy, 2020). Universal Language Model Fine-tuning (ULMFiT) utilizes pre-trained language representations in three stages (Figure 2) (Howard & Ruder, 2018). Furthermore, Multinomial Naive Bayes (MNB), Random Forest (RF), and Support Vector Machine (SVM) function as probabilistic, decision-tree, and linear classifiers respectively (Lee, 2018).

Pouransari and Ghili (Pouransari, 2014) addressed binary representation challenges through ML, achieving accuracies with models like RF, SVM, and Logistic Regression, while Palkar et al. (Palkar,

2016) tackled similar issues with IMDB movie reviews, showcasing competitive performances and suggesting exploration of Deep models. Dridi and Recupero (Dridi, 2017) employ MNB for Opinion Mining in IMDB, utilizing numerical vectors, "n_gram" tokenization, "frame semantics," and "lexical resources".

Figure 1 BERT Pre-Training and Fine-Tuning Procedures (Batra, 2021)



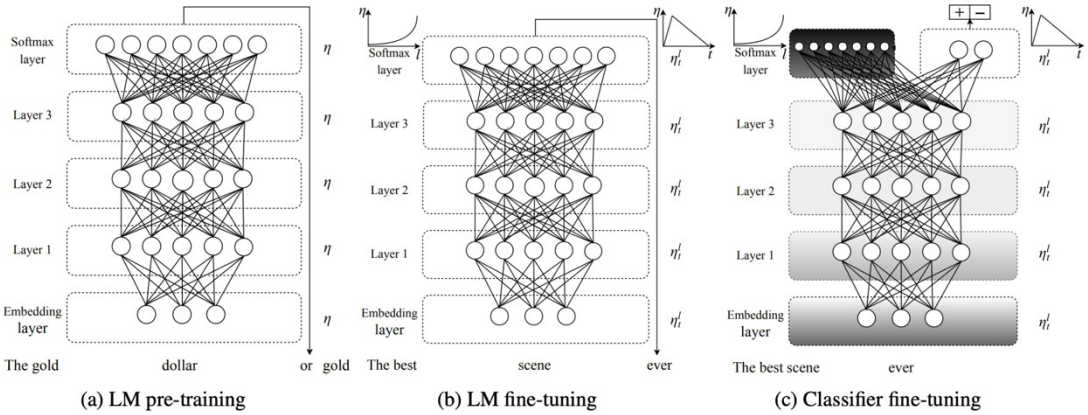


Figure 2 ULMFiT Model (Howard & Ruder, 2018)

Comparison of rule-based and ML methods for assessing movie stance has been explored across different modes. ML approaches, such as XLNet (96.21%) (Yang Z. D., 2020), BERT_large+ITPT (95.79%) (Sun, 2020), BERT_base+ITPT (95.63%) (Sun, 2020), ULMFiT (95.4%) (Howard J. &., 2018), Block-sparse LSTM (94.99%) (Yang Z. D., 2020), oh-LSTM (94.1%) (Johnson, 2016) and Virtual adversarial training (94.1) (Miyato, 2021) consistently outperforms in accurately classifying sentiments on IMDB movie dataset).

III Data & Methodology

Sentiment analysis in NLP evaluates subjective information in text, essential for gauging audience sentiments in movie reviews, as depicted in Figure 3 outlining the research methodology.

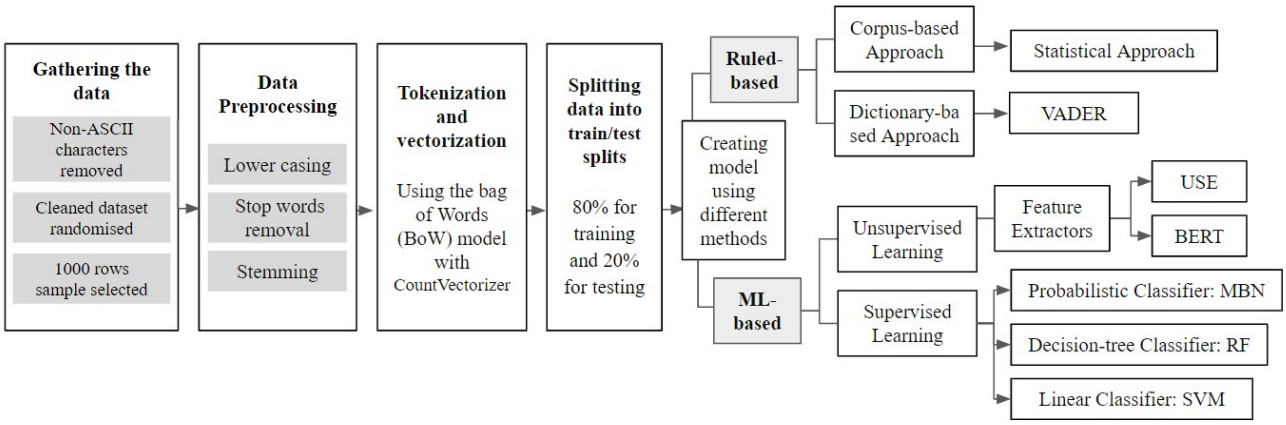


Figure 1 Process Flow for Sentiment analysis approaches

The IMDb Movie Review Dataset (50,000 entries) contains diverse lengths (min: 70, max: 1651) with a balanced sentiment analysis binary system (mean: 0.53). Figures 5 and 6 illustrate right-skewed lengths and a balanced positive/negative sentiment distribution.

Summary statistics	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
review lengths	70.0	707.7	994.5	1349.9	1651.5	1651.5
sentiments	0.000	0.000	1.000	0.503	1.000	1.000

Figure 2 Dataset summary statistics

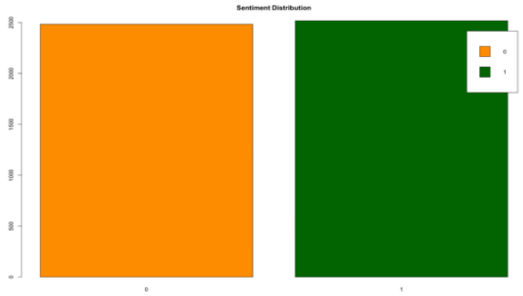


Figure 3 Sentiment Distribution

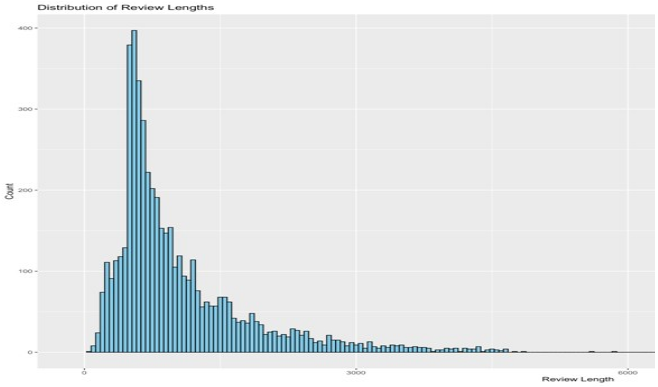


Figure 4 Distribution of Review Lenghts

Text columns are preprocessed by removing 5.042% of rows (2521) starting with non-ASCII characters (5.042%) due to parse errors in extracting data, removing stop words and lowercasing characters. The data is randomized, and 1000 rows are selected for representative sampling. Tokenization and vectorization are performed using the Bag of Words (BoW) model with CountVectorizer to prepare the data. Model performance is assessed with a 80-20 train-test split to ensure generalization to unseen data.

Figure 5 WordCloud before pre-processing

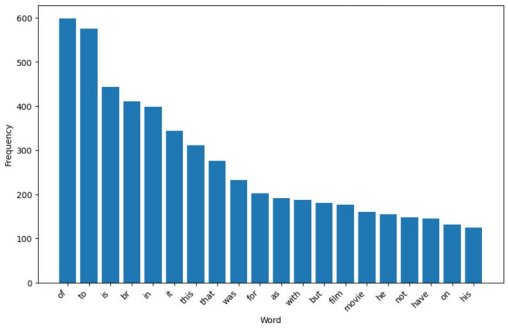


Figure 6 Top 20 Words in BoW before

Figure 7 WordCloud after pre-processing

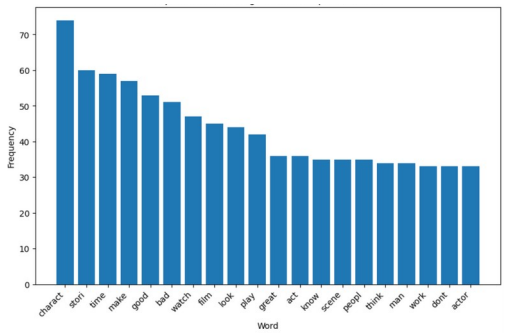


Figure 8 Top 20 Words in BoW after

III.I: Rule-Based Models

Various models are used, with subsequent explanations for the chosen selections.

The Statistical model within the corpus-based approach employs strategies like negation handling, sentiment-shifting recognition, emoticon analysis, and diverse linguistic features. It assigns sentiment scores, considering intensifiers, diminishers, sarcasm, word frequency, specific genres, etc.

VADER, a dictionary-based sentiment analysis tool, is chosen for computational efficiency and nuanced handling incorporated in two models. VADER Analysis with Clarity Emphasis uses a 0.2 binary classification threshold, prioritizing simplicity; Sensitivity Exploration enhances accuracy employing a 0.1 threshold for nuanced variations. The former focuses on simplicity, while the latter explores nuances through varied thresholds and model combinations for increased accuracy.

III.II: ML Models

Selected models cater to movie stance assessment strengths: probability (MNB), complexity (RF), intricate patterns (SVM), semantic meaning (USE) and contextual language understanding (BERT).

The BERT model undergoes review tokenization, encoding, and PyTorch DataLoader setup. Sentiment labels transform into tensors for datasets, and fine-tuning involves the use of the AdamW optimizer, linear scheduler, and PyTorch training loop.

The MNB uses a process to turn text into numbers, normalize importance, and classify text with MultinomialNB. GridSearchCV fine-tunes settings to handle challenges in movie reviews like different structures and subtle feelings.

The SVM model uses a linear kernel and TF-IDF vectorization to highlight important words. The linear kernel is chosen to efficiently understand complex text connections.

The USE combined with Logistic Regression is chosen for its strong semantic capturing ability, utilizing USE embeddings as effective features that enhance nuanced sentiment representations.

The RF employs TfidfVectorizer and RandomForestClassifier for capturing complex textual relationships crucial for sentiments in movie reviews. CountVectorizer considers the top 5000 features, incorporating English stop words to address movie review intricacies.

IV Evaluation metrics

This study evaluates rule-based and ML methods for movie sentiment analysis using accuracy, precision, recall, and F1-score metrics (Setiawan, 2022).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN} \quad \text{F-Measure} = 2 \times \frac{Precision \times Recall}{Precision \times Recall}$$

Figure 9 Evaluation metrics (Accuracy, Precision, Recall, F-measure)

Accuracy measures correct predictions over total instances; precision evaluates positive prediction accuracy; recall assesses the model's ability to identify all positive instances and F1-score provides a

balanced evaluation by considering false positives and negatives. True Positive (TP) and True Negative (TN) represent correct identifications, while False Negative (FN) and False Positive (FP) indicate missed positives and false alarms. These metrics are crucial for evaluating a model's effectiveness, especially in imbalanced class distributions. Due to the setting random samples, model evaluation leads to different convergence in each run. Besides, the computational cost is also accounted for to assess models' scalability and cost-effectiveness.

V Results

Rule-based and ML models are contrasted in Figure 10 and 11, detailing accuracy and classification. BERT (90%) and VADER2 (72%) are used as baseline models for ML and rule-based approaches respectively. ML models, such as BERT (90%), SVM (87.05%), and USE (86.68%), demonstrate effectiveness in understanding movie sentiments. Rule-based methods offer a straightforward analysis, but improving accuracy can be challenging due to the simplicity of the approach. Nevertheless, in terms of time-wise computational cost, BERT (34' to run) is 80.95 times more costly than VADER2 (0.42'). The choice between methods should consider research objectives, computational efficiency, interpretability, and sensitivity. Besides, achieving higher accuracy or reducing error rates does not automatically indicate superior performance on the task (Hand, 1997).

		Rule-based approaches			Machine Learning approaches				
		Corpus-based	Dictionary-based		Supervised Learning				
		STAT	VADER	VADER2	BERT	USE	MNB	SVM	RF
		M1	M2	M3	M4	M7	M5	M6	M8
Accuracy		0.67	0.69	0.72	0.9	0.86	0.83	0.87	0.83
Precision	negative	0.63	0.77	0.76	0.83	0.87	0.83	0.89	0.83
	positive	0.73	0.66	0.69	1.00	0.87	0.84	0.85	0.85
	macro avg	0.46	0.71	0.73	0.93	0.87	0.83	0.87	0.84
	weighted avg	0.7	0.71	0.73	0.91	0.87	0.84	0.87	0.84
Recall	negative	0.81	0.56	0.6	1.00	0.87	0.82	0.84	0.86
	positive	0.54	0.84	0.83	0.71	0.87	0.83	0.9	0.82
	macro avg	0.45	0.7	0.7	0.86	0.87	0.83	0.87	0.84
	weighted avg	0.7	0.7	0.7	0.90	0.87	0.83	0.87	0.84
F1-score	negative	0.71	0.65	0.67	0.93	0.87	0.84	0.87	0.84
	positive	0.63	0.74	0.75	0.83	0.87	0.83	0.87	0.84
	macro avg	0.45	0.69	0.71	0.88	0.87	0.83	0.87	0.84
	weighted avg	0.67	0.69	0.72	0.9	0.87	0.83	0.87	0.84

Figure 10 Comparison of model effectiveness

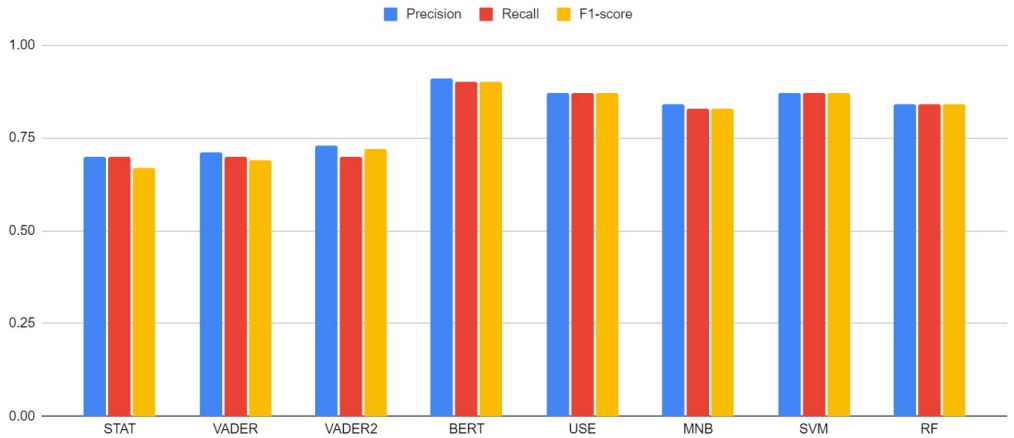


Figure 11 Visual comparison of model effectiveness

VI Discussion and Conclusion

This study compares sentiment analysis in movie reviews, contrasting rule-based simplicity with nuanced ML models. ML models such as BERT provided better accuracy (90%). Each model has distinct strengths, emphasizing the importance of aligning methodologies with research goals. Future opportunities for refinement include addressing limited training time and further research in rule-based methods. The study suggests exploring broader use cases, like Posner et al.'s Circumplex model (Figure 12) (Posner, 2005), to enhance sentiment analysis in movie reviews beyond the binary "positive-negative" perspective. Sentiment analysis extends beyond movies, serving to predict success and understand customer opinions.

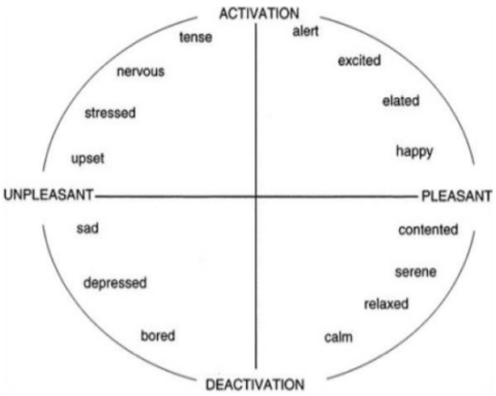


Figure 12 Circumplex model

Bibliography

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. https://arxiv.org/pdf/1801.06146.pdf.

Alduailej, A. A. (2022). *AraXLNet: pre-trained language model.* Retrieved from https://www.researchgate.net/journal/Journal-of-Big-Data-2196-1115/publication/360992903_AraXLNet_pre-trained_language_model_for_sentiment_analysis_of_Arabic/links/6296cf111117461e03af7bd5/AraXLNet-pre-trained-language-model-for-sentiment-analysis-of-Arab

Alqaryouti, O. S. (2019). *Aspect-based sentiment analysis using smart government review data.* Retrieved from https://www.emerald.com/insight/content/doi/10.1016/j.aci.2019.11.003/full/pdf?title=aspect-based-sentiment-analysis-using-smart-government-review-data

Batra, H. P. (2021). *BERT-Based Sentimental Analysis: A Software Engineering Perspective. Allahabad, India : Indian Institute of Information Tehnology, 2021.* Retrieved from https://arxiv.org/pdf/2106.02581.pdf

Cer, D. Y.-y.-C.-H. (2018). *Universal Sentence Encoder. arXiv preprint arXiv:1803.11175.* Retrieved from https://arxiv.org/pdf/1803.11175.pdf

Dridi, A. a. (2017). *More sense: Movie reviews sentiment. https://ceur-ws.org/Vol-1874/paper_3.pdf.*

Hand, D. (1997). *Construction and Assessment of Classification Rules (p. 98). Wiley. https://books.google.nl/books/about/Construction_and_Assessment_of_Classific.html?id=NRfvAAAAMAAJ&redir_esc=y.*

Howard, J. &. (2018). *Universal Language Model Fine-tuning for Text Classification. fast.ai. University of San Francisco. Insight Centre, NUI Galway. Aylien Ltd., Dublin. Retrieved from https://arxiv.org/pdf/1801.06146.pdf.*

Johnson, R. &. (2016). *Supervised and Semi-Supervised Text Categorization using LSTM for Region Embeddings. RJ Research Consulting, Tarrytown, NY, USA; Big Data Lab, Baidu Inc, Beijing, China. Retrieved from https://arxiv.org/pdf/1602.02373.pdf.*

Lee, S. M. (2018). *A Comparison of Machine Learning Algorithms for the Surveillance of Autism Spectrum Disorder. Centers for Disease Control and Prevention, Atlanta, GA. Retrieved from https://arxiv.org/ftp/arxiv/papers/1804/1804.0.*

Li, Z. Z. (2021). *Title of the Paper. Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University. Retrieved from https://arxiv.org/pdf/2111.02194.pdf.*

Liachoudi, G. (2020). *Sentiment Analysis of Movie Reviews by Merging Comments from Two Well-Known Platforms. Tilburg University, School of Humanities and Digital Sciences, Department of Cognitive Science & Artificial Intelligence. Retrieved from http://a.*

Miyato, T. D. (2021). *Adversarial Training Methods for Semi-Supervised Text Classification. Preferred Networks, Inc., ATR Cognitive Mechanisms Laboratories, Kyoto University, Google Brain, OpenAI. Retrieved from https://arxiv.or.*

Munikar, M. S. (2019). *Fine-grained Sentiment Classification using BERT. *Department of Electronics and Computer Engineering, Pulchowk Campus, Institute of Engineering. https://arxiv.org/pdf/1910.03474.pdf.*

Murthy, D. A. (2020). Text based Sentiment Analysis using LSTM. International Journal of Engineering Research, V9(05). https://doi.org/10.17577/IJERTV9IS050290.

Onalaja, S. R. (2021). *Aspect-based Sentiment Analysis of Movie Reviews. Data Science Review, volume number(issue number). Retrieved from https://scholar.smu.edu/cgi/viewcontent.cgi?article=1205&context=datasciencereview.*

Palkar, R. K. (2016). *Comparative evaluation of supervised learning algorithms for sentiment analysis of movie reviews. International Journal of Computer Applications. https://arno.uvt.nl/show.cgi?fid=156527.*

Peters, M. E. (2018). *Deep contextualized word representations. University of Washington. Retrieved from https://arxiv.org/pdf/1802.05365.pdf.*

Poria, S. M. (2019). *Multimodal Sentiment Analysis: Addressing Key Issues and Setting up the Baselines. Retrieved from https://arxiv.org/pdf/1803.07427.pdf.*

Posner, J. R. (2005). *The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and Psychopathology, 17 (3), 715-734. DOI: 10.1017/S095457940505.*

Pouransari, H. a. (2014). *Deep learning for sentiment analysis of movie reviews. https://cs224d.stanford.edu/reports/PouransariHadi.pdf.*

Ray, P. &. (2019). *A Mixed approach of Deep Learning method and Rule-Based method to improve Aspect Level Sentiment Analysis.* https://www.emerald.com/insight/content/doi/10.1016/j.aci.2019.02.002/full/pdf?title=a-mixed-approach-of-deep-learning-method-and-rule-based-method-to-improve-aspect-level-sentiment-analysis.

Setiawan, I. W. (2022). *Utilizing Random Forest Algorithm for Sentiment Prediction Based on Twitter Data. Retrieved from https://www.atlantis-press.com/art.* Retrieved from https://www.atlantis-press.com/proceedings/mimse-i-c-22/125980161

Sun, C. Q. (2020). *How to Fine-Tune BERT for Text Classification? Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, School of Computer Science. Retrieved from https://arxiv.org/pdf/1905.05583.pdf.*

Taboada, M. B. (2011). *Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics, 37(2), 267. https://doi.org/10.1162/COLI_a_00049.*

Yang, Z. D. (2020). *XLNet: Generalized Autoregressive Pretraining for Language Understanding. Carnegie Mellon University, Google AI Brain Team. https://arxiv.org/pdf/1906.08237.pdf.*

[Code used: https://colab.research.google.com/drive/1xkD-FE1_ul-7gDQTsyj8iprnyw3I2hzf?usp=sharing]

Appendix

Rule-based Approaches

Accuracy: 0.6700
Classification Report statistical approach:

	precision	recall	f1-score
negative	0.63	0.81	0.71
neutral	0.00	0.00	0.00
positive	0.76	0.54	0.63
accuracy			0.67
macro avg	0.46	0.45	0.45
weighted avg	0.70	0.67	0.67

Figure 13 Corpus-based Statistical Approach

Accuracy of Vader Model: 0.6993
Classification Report for Vader Model:

	precision	recall	f1-score
negative	0.77	0.56	0.65
positive	0.66	0.84	0.74
accuracy			0.70
macro avg	0.71	0.70	0.69
weighted avg	0.71	0.70	0.69

Figure 14 Dictionary-based VADER

Accuracy: 0.72
Classification Report:

	precision	recall	f1-score
negative	0.76	0.60	0.67
positive	0.69	0.83	0.75
accuracy			0.72
macro avg	0.73	0.72	0.71
weighted avg	0.73	0.72	0.72

Figure 15 Dictionary-based VADER2

Machine Learning Approaches

Accuracy: 0.9
Classification Report:

	precision	recall	f1-score
0	0.87	1.00	0.93
1	1.00	0.71	0.83
accuracy			0.90
macro avg	0.93	0.86	0.88
weighted avg	0.91	0.90	0.90

Figure 16 BERT

Accuracy of SVM Model: 0.8705
Classification Report for SVM Model:

	precision	recall	f1-score
negative	0.89	0.84	0.87
positive	0.85	0.90	0.87
accuracy			0.87
macro avg	0.87	0.87	0.87
weighted avg	0.87	0.87	0.87

Figure 17 SVM

Accuracy (USE embeddings): 0.8685
Classification Report:

	precision	recall	f1-score
0	0.87	0.87	0.87
1	0.87	0.87	0.87
accuracy			0.87
macro avg	0.87	0.87	0.87
weighted avg	0.87	0.87	0.87

Figure 18 USE

MNB Model Accuracy: 0.8339169584792396
MNB Model Classification Report:

	precision	recall	f1-score
negative	0.83	0.84	0.84
positive	0.84	0.82	0.83
accuracy			0.83
macro avg	0.83	0.83	0.83
weighted avg	0.83	0.83	0.83

Figure 19 MNB

Random Forest Classifier:
Accuracy: 0.839
Classification Report:

	precision	recall	f1-score
negative	0.83	0.86	0.84
positive	0.85	0.82	0.84
accuracy			0.84
macro avg	0.84	0.84	0.84
weighted avg	0.84	0.84	0.84

Figure 20 Random Forest