EAI Endorsed Transactions Preprint

Research Article/Editorial

Sentiment Analysis using Machine Learning and Deep Learning Models

Thuy Vu Phuong[1]

[1] 20010751, k14-AIR, Faculty of Electrical Electronic Engineering, Phenikaa University

Abstract

INTRODUCTION: Sentiment analysis, a crucial facet of natural language processing, is explored in this research using two distinct datasets: the IMDb Movie Reviews corpus and a unique Financial Sentiment dataset. The study adopts a holistic approach, integrating traditional Machine Learning algorithms and Deep Learning models to analyze and interpret sentiments.

OBJECTIVES: Aiming to investigate sentiment analysis methodologies, evaluate model performance across diverse datasets, and contribute nuanced insights into sentiment classification in various contexts.

METHODS: The preprocessing phase involves a combination of techniques, including Punctuation Removal, Text Lowercasing, and specialized considerations for financial symbols like '$' and '%'. Additionally, float numbers are rounded to integers for consistency.

RESULTS: Machine Learning models, including Logistic Regression and Random Forest, are applied to the IMDb dataset, with Naive Bayes exhibiting superior performance with 89%. The Deep Learning model achieves an accuracy of 82% on the testing set, demonstrating proficiency in sentiment analysis. In the Financial Sentiment dataset, Naive Bayes emerges as the optimal model, despite a decline in overall model performance due to dataset limitations and imbalance.

CONCLUSION: This study contributes valuable insights into sentiment analysis methodologies, offering a nuanced understanding of model performance across diverse datasets. The exploration of both traditional Machine Learning and Deep Learning approaches enhances our comprehension of sentiment classification in various contexts.

Received on XXXX; accepted on XXXX; published on XXXX

Keywords: Sentiment analyze, LaTeX 2_ε, *EAI Endorsed Transactions*

doi:10.4108/XX.X.X.XX

1. Introduction

Sentiment analysis, a pivotal component of natural language processing, plays a crucial role in extracting valuable insights from textual data by discerning public opinions and emotions. In a study by Sun et al. [13], various fine-tuning methods of BERT were explored in the context of text classification tasks, including sentiment analysis, question classification, and topic classification. The study identified a general solution for BERT fine-tuning, utilizing the binary film review IMDb dataset [1] for the sentiment analysis task. However, Yang et al. [15] pointed out limitations in BERT related to the neglect of dependencies among masked positions and disparities between pretraining and fine-tuning models. Addressing these

★This is an example of a footnote to the title.

concerns, XLNet [15] was proposed, inheriting the strengths of Transformer-XL [4], the state-of-the-art autoregressive model. XLNet's pretraining approach overcomes BERT's limitations, demonstrating superior performance across 20 tasks including sentiment analysis.

This study delves into sentiment analysis, thoroughly examining the raw data and integrating pre-processing techniques to extract meaningful features from the text. The research employs a combination of traditional Machine Learning and Deep Learning models to analyze and interpret sentiments. The focus is on dissecting sentiments within two distinct datasets: the well-established IMDb Movie Reviews corpus, offering a binary sentiment classification task, and a Financial Sentiment dataset amalgamated from FiQA and Financial PhraseBank. The IMDb dataset, comprising 50,000 reviews, serves as a robust foundation for exploring natural language processing techniques in the context of film critiques. Additionally, the Financial Sentiment dataset, with its fusion of financial sentences and sentiment labels, presents a unique challenge for sentiment classification in the financial domain. Through comprehensive experimentation and evaluation, this study aims to unveil the performance nuances of different models on these diverse datasets, shedding light on their strengths and limitations.

2. Related works

2.1. Data Engineering and Machine Learning algorithms

Feature extraction [5], [2] plays a crucial role in Natural Language Processing in general, and Sentiment analysis in particular. In vectorization, there are a vast of research was carried out to survey the impact of these methods on the models. TF-IDF (Term Frequency Inverse Document Frequency), which was first proposed in [5], is an omnipresent algorithm to represent text in meaningful numbers, suitable for fitting on machine algorithms and a significant increase in accuracy compared to earlier feature extraction methods. The authors [2] have surveyed the effect of feature extraction methods on the Sentiment analysis tasks with TF-IDF word level and N-Gram, and using six machine learning algorithms (Decision Tree, Support Vector Machine, K-Nearest Neighbour, Random Forest, Logistic Regression, Naive Bayes) to objectively compare with the performance of the models without feature extraction and show their outstanding. Machine Learning is a favorable approach because of its efficiency and performance on the data after being extracted [8], [2], [3]

2.2. Deep Learning

The application of word embedding and Long Short-Term Memory (LSTM) networks in sentiment analysis has been extensively explored in various research works. [14] and [12] explore the synergy of word embedding and LSTM for effective sentiment classification. Word embedding captures nuanced word usages, and LSTM learns long-distance dependencies in sentence structures. AEC-LSTM [7], a novel model for sentiment analysis, integrates emotional intelligence (EI) and attention mechanisms into LSTM networks. The emotion-enhanced LSTM (ELSTM) leverages EI, including an emotion modulator and estimator, and combines operations like convolution and pooling to capture diverse structural patterns in text. citewang2016attention addresses aspect-level sentiment classification, a nuanced task in sentiment analysis receiving increased attention. The method recognizes that a sentence's sentiment polarity is not solely determined by its content but is also closely tied to the specific aspect in focus. To explore this connection, the paper proposes an Attention-based Long Short-Term Memory (LSTM) Network. The attention mechanism enables the model to focus on different parts of a sentence based on the considered aspects.

3. Methodology

3.1. Pre-processing data

In the preprocessing phase aimed at training the Machine Learning model on both datasets, a series of techniques are employed to optimize the data. This involves the amalgamation of various preprocessing methods, including but not limited to punctuation removal, text lowercasing, etc., as visually depicted in Fig. 3. It is noteworthy that the Financial Sentiment dataset, being intricately tied to financial contexts, entails specialized considerations. Specifically, certain symbols such as '$' and '%' bear substantive meaning within the financial domain. Consequently, a meticulous substitution process is executed, wherein these symbols are replaced with their corresponding textual representations. Additionally, to ensure consistency and mitigate potential alterations during subsequent preprocessing stages, all floating-point numbers are rounded to integers. This preemptive measure safeguards against unintended modifications, particularly in the aftermath of procedures such as punctuation removal.

With the Deep Learning approach, a simpler preprocessing procedure was utilized as Fig. 2. To undertake the classification task, it is imperative to prepare the data by utilizing comments as the primary text input and their associated sentiments as the corresponding labels for training the model during

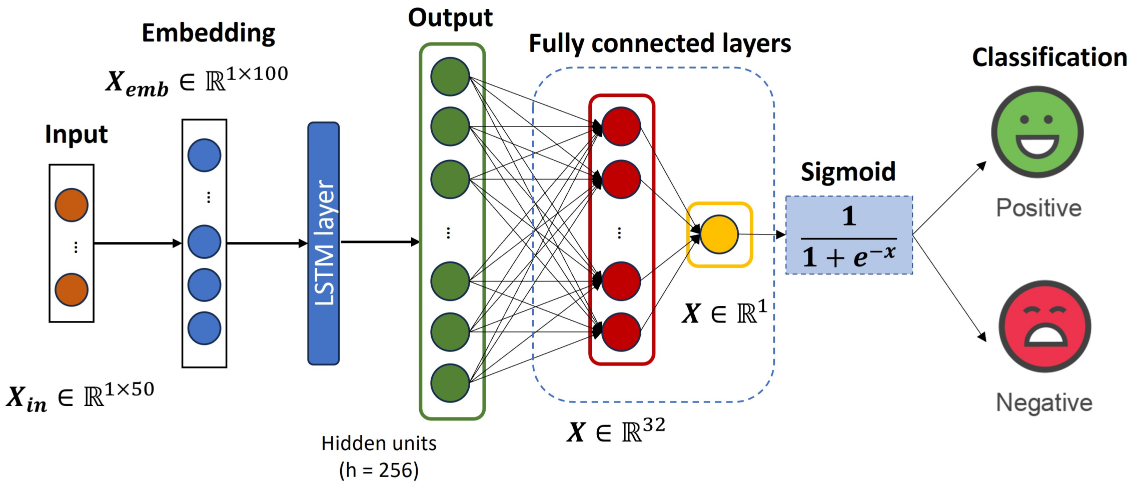

Figure 1. Model architecture

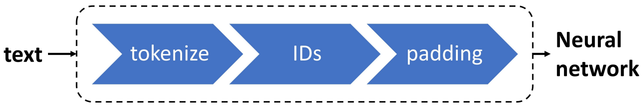

Figure 2. Pre-processing data for Deep Learning model

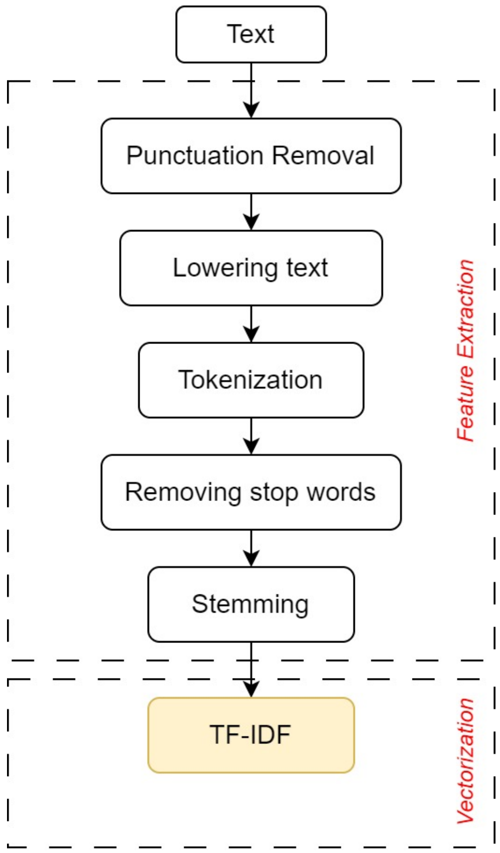

Figure 3. Pre-processing data for Machine Learning model

the pre-processing phase. Initially, the comments undergo tokenization, a crucial step in breaking down the textual content into individual tokens, effectively forming the basic units for analysis. This tokenized representation is further utilized to construct a comprehensive vocabulary. The Keras' Tokenizer is employed for this task instead of TF-IDF[5], facilitating the assignment of unique IDs to each token within the vocabulary. Subsequently, in the interest of ensuring uniformity in the input data, short sequences of utterances are padded to adhere to a predefined maximum length. This step is essential for maintaining consistency in the input dimensions, thus optimizing the compatibility of the data for subsequent stages of model training and evaluation. The obtained dataset was divided into 2 training and testing sets, accounting for one-fifth of the total. In the training process, a 20% training set was used to validate the model's performance.

3.2. Training models

Machine Learning Algorithms. Machine Learning algorithms are widely used in NLP because of their simplicity and inexpensive computation. In this study, there are some algorithms with default hyperparameters used to analyze the sentiment.

Logistic Regression is a widely used and versatile statistical method for binary and multi-class classification tasks. Despite its name, it is primarily employed in classification rather than regression. The algorithm models the relationship between a binary or multiclass dependent variable and one or more independent variables by estimating probabilities using the logistic function. In the binary case, Logistic Regression predicts the probability that an instance belongs to a particular class, allowing for efficient decision-making based on thresholding these probabilities. For multi-class scenarios, it can be extended through various strategies, such as One-vs-Rest or softmax regression.

Random Forest [9] operates by constructing an ensemble of decision trees during training, where each tree is built on a different subset of the training data and a random subset of features. The name Random Forest stems from the inherent randomness introduced in both the data sampling and feature selection processes. This randomness helps mitigate overfitting issues and enhances the model's robustness. The final prediction in a Random Forest is determined through a voting mechanism for classification tasks, where the most frequent class among the individual trees is chosen, or through averaging for regression tasks.

Naïve Bayes is a probabilistic machine learning algorithm based on Bayes' theorem, designed for classification tasks. The algorithm makes an assumption of conditional independence among the features, given the class label, which is where the term "naive" originates. It calculates the probability of each class for a given set of features and assigns the class with the highest probability as the final prediction.

Support Vector Machine Support Vector Machines (SVM) [6] operates by finding the optimal hyperplane that best separates data points of different classes in a high-dimensional space. The key objective is to maximize the margin between the classes, defined as the distance between the hyperplane and the nearest data points from each class. SVM can handle linear and non-linear decision boundaries through the use of different kernel functions, such as polynomial, radial basis function (RBF), or sigmoid kernels. This flexibility allows SVM to effectively capture complex relationships in the data. SVM is known for its ability to handle high-dimensional feature spaces, resistance to overfitting, and generalization to various domains, including image recognition, text classification, and bioinformatics.

Deep Learning method. Incorporated within the model architecture are fundamental layers including *Embedding*, Long Short-Term Memory (*LSTM*), and *Dense* layers as Fig. 1. The sequential output from these layers undergoes processing through the *Sigmoid* function as Eq.1, resulting in a transformation of values within the range of 0 to 1. A binary interpretation is then applied, where an output of 0 designates an affiliation with Negative sentiment, and conversely, an output of 1 signifies a positive sentiment classification. This systematic approach within the model's structure facilitates the classification of sentiment based on the continuous range of transformed values, providing a discernible indication of the sentiment polarity associated with the given input. The details of the model configuration are shown in Table 1.

$$S = \frac{1}{1 + e^{-x}} \tag{1}$$

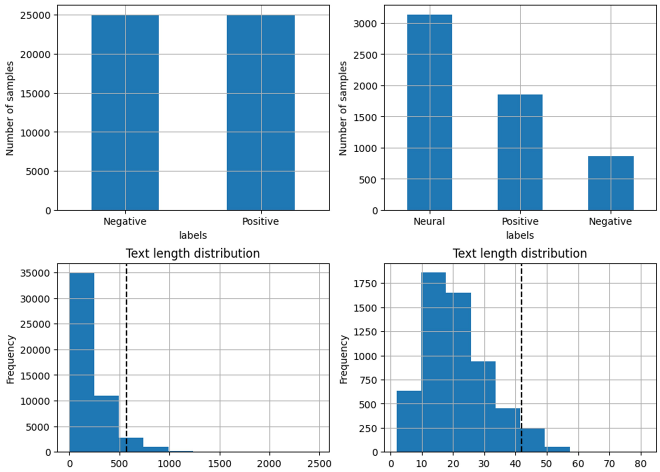

Figure 4. Number of samples per each class (upper row) and text length distribution (lower row) of IMDb (left column) and Financial Sentiment (right column) dataset

Table 1. DETAILED CONFIGURATION OF THE ARCHITECTURE

Layer	Output size	Params
Input	0	50
Embedding	100	14,579,000
LSTM	256	365,568
FC	32	8,224
FC	1	33
Total		14,970,825

4. Experiments

We experiment on the Internet Movie Database (IMDb) [11] and Financial Sentiment dataset [10]. The IMDb Movie Reviews dataset serves as a binary sentiment analysis corpus, encompassing 50,000 reviews meticulously curated from the Internet Movie Database (IMDb). Each review is distinctly categorized as either positive or negative, contributing to a balanced dataset with an equal distribution of 25,000 positive and 25,000 negative sentiments as Fig. 4. Notably, the dataset is selectively composed of highly polarized reviews, where a negative sentiment corresponds to a score of ≤ 4 out of 10, while a positive sentiment is indicated by a score of ≥ 7 out of 10. Moreover, the dataset adheres to a constraint of no more than 30 reviews per movie, a measure implemented to maintain diversity and representation across the entire collection. This dataset offers a robust foundation for sentiment analysis research and facilitates the exploration of natural language processing techniques in the context of film reviews.

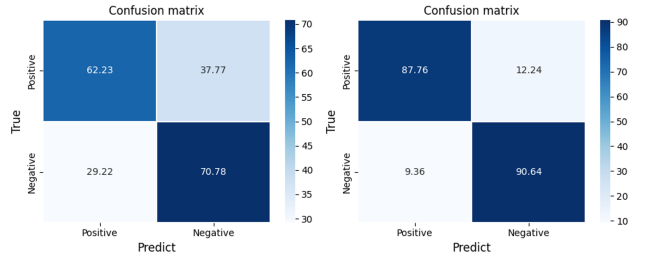

Figure 5. Confusion matrix of Random Forest (the worst-left) and Naive Bayes (the best-right)

The Financial Sentiment dataset presented is a consolidated resource derived from the combination of two distinct datasets, FiQA (Financial QA) and Financial PhraseBank. Each entry in the dataset comprises financial sentences accompanied by sentiment labels, providing a valuable foundation for training and evaluating natural language processing and machine learning models specializing in sentiment classification within financial texts. The dataset includes 5000 samples. Each piece of text within the dataset is labeled with sentiment indicators, categorizing the expressed sentiment as positive, negative, or neutral.

Additionally, 80% of both datasets is allocated for training, while the remaining 20% is earmarked for testing. Each dataset undergoes experimentation involving a pre-processing procedure, as elucidated in Sec. 3 of the study. This involves the application of Machine Learning algorithms and Deep Learning models. The Machine Learning algorithms are employed to train the pre-processed dataset and subsequently compare its performance against alternative configurations. The optimal model is ascertained through Cross-validation, a rigorous evaluation method aimed at enhancing the overall performance of the chosen model.

5. Results and discussion

5.1. IMDb dataset

Considering the efficacy of the Machine Learning models applied to the dataset, the results are presented in Table 3 and Fig. 5. Notably, the Naive Bayes model demonstrates superior performance compared to the Random Forest algorithm following the prescribed pre-processing procedure. Subsequently, the Naive Bayes model undergoes training with Cross-Validation, thereby enhancing its overall performance, as illustrated in the corresponding Fig. 6.

The Deep Learning model demonstrates an accuracy of 82% on the testing set. A visual inspection of Fig. 7 reveals apparent evidence of proficient training, as both the training and validation sets exhibit a converging trend in the loss function, accompanied by an increase in accuracy. Notably, after 2 epochs, the loss function

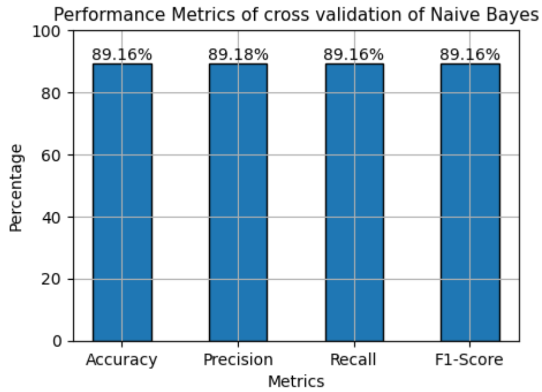

Figure 6. Evaluation of Cross-Validation of Naive Bayes

Model	P(%)	R(%)	F1(%)	Acc(%)
Logistic Regression	89	88	88	88
Random Forest	67	66	65	66
Naïve Bayes	**89**	**89**	**89**	**89**

Table 2. Performance Metrics of Different Machine Learning Models on IMDb dataset

and accuracy for the validation set remain relatively stable, whereas positive changes are observed in the testing set metrics.

Fig. 8 illustrates that the model excels in accurately classifying positive sentiment, achieving a correctness rate of 85.02%, with only 14.98% of positive comments being erroneously predicted as negative. Conversely, a notable observation is that approximately one-fifth of negative comments are misclassified as positive, indicating a comparatively lower performance in discerning negative sentiment.

Besides, some external comments were used to test the model and we got the results:

```
Comment: So amazing
Label: Positive
Model predicts: Positive

Comment: This film is similar to my
life, so touching
Label: Positive
Model predicts: Positive

Comment: The plot is so so bad, I was
asleep when I watched this film, and I
will never watch it again
Label: Negative
Model predicts: Positive
```

Generally, the model can analyze the sentiment of

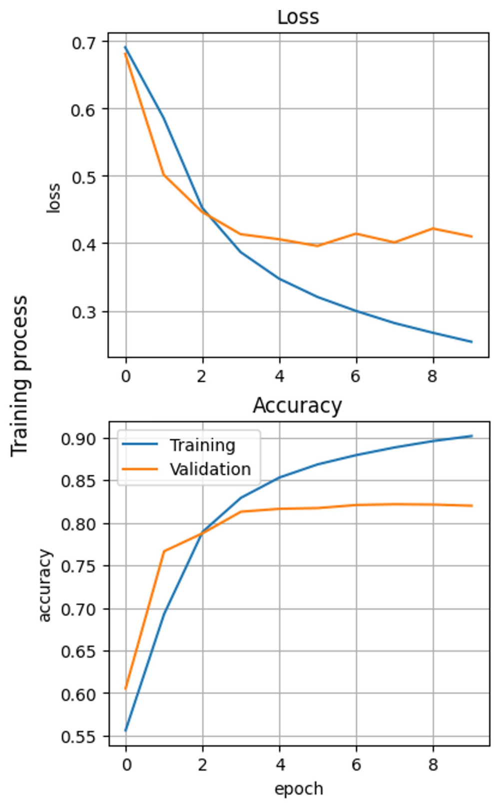

Figure 7. Training process of the Deep Learning model on IMDb dataset

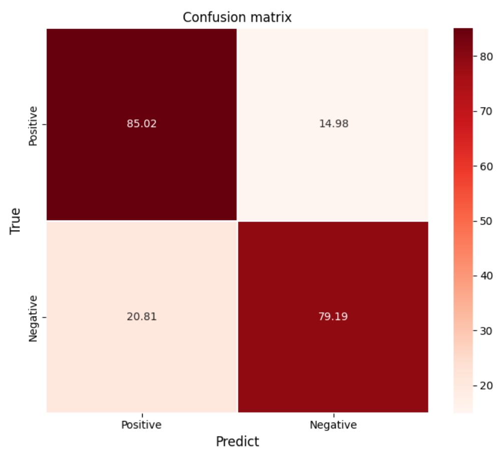

Figure 8. Confusion matrix of Deep Learning model

the comment. However, with the last case which is a negative comment, the predicted label is positive. It can be easily understood since the model showed better performance when predicting positive class.

5.2. Financial Sentiment dataset

The performance of the models exhibits a decline compared to the previous dataset, attributable to the limited number of samples in the current

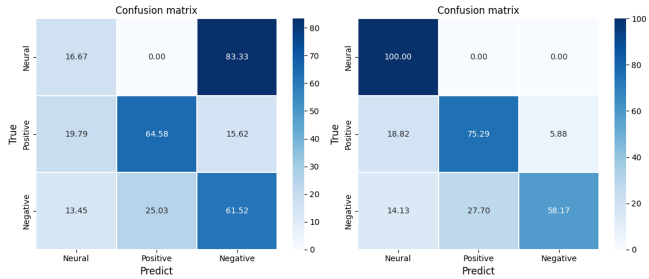

Figure 9. Confusion matrix of Logistic Regression (the worst-left) and Naive Bayes (the best-right) on Financial dataset

Model	**P(%)**	**R(%)**	**F1(%)**	**Acc(%)**
SVM	54	58	49	58
Naive Bayes	70	59	49	59
Logistic Regression	56	61	52	61

Table 3. Performance of Different Machine Learning Models on Financial Sentiment dataset

dataset. Notably, the optimal model, Naive Bayes, demonstrates precise predictions for the Neural class, while experiencing a reduction in accuracy for the Positive and Negative classes, achieving 75.2% and 58.17% as 9, respectively. This discernible deviation can be attributed to the imbalanced nature of the dataset, as visually depicted in Figure 4.

6. Conclusion

The study presented delves into sentiment analysis, exploring the intricacies of natural language processing to dissect public opinions and emotions. The focus is on two distinct datasets: the well-established IMDb Movie Reviews corpus, offering a binary sentiment classification task, and a Financial Sentiment dataset amalgamated from FiQA and Financial PhraseBank, presenting a unique challenge in sentiment classification within the financial domain.

The study employs a comprehensive approach, integrating traditional Machine Learning algorithms and Deep Learning models to analyze and interpret sentiments. The preprocessing techniques involve a combination of methods, including punctuation removal, text lowercasing, and specialized considerations for financial symbols such as '$' and '%'. Additionally, float numbers are rounded to integers to maintain consistency through subsequent preprocessing stages. In Machine Learning algorithms, logistic regression, and random forest models are employed on the IMDb dataset, with Naive Bayes demonstrating superior performance of 89%. The Deep Learning model achieves an accuracy of 82% on the testing set, showcasing its proficiency in sentiment analysis. With the Financial Sentiment dataset, Machine Learning models, including SVM, Naive Bayes, and logistic regression, exhibit a decline in performance, attributed to the dataset's limited size and imbalanced nature. Naive Bayes emerges as the optimal model, showcasing precise predictions for the Neutral class.

References

[1] Andrew L. "Maas et al. “"Learning Word Vectors for Sentiment Analysis"”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011.

[2] Ravinder Ahuja et al. “The Impact of Features Extraction on the Sentiment Analysis”. In: *Procedia Computer Science* 152 (2019). International Conference on Pervasive Computing Advances and Applications-PerCAA 2019, pp. 341–348. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2019.05.008. URL: https://www.sciencedirect.com/science/article/pii/S1877050919306593.

[3] M Avinash and E Sivasankar. “A Study of Feature Extraction Techniques for Sentiment Analysis”. In: *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 3*. Springer. 2019, pp. 475–486.

[4] Zihang Dai et al. “Transformer-XL: Attentive Language Models beyond a Fixed-Length Context”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 2978–2988. DOI: 10.18653/v1/P19-1285.

[5] Bijoyan Das and Sarit Chakraborty. “An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation”. In: 2018. arXiv: 1806.06407 [cs.CL].

[6] M.A. Hearst et al. “Support vector machines”. In: *IEEE Intelligent Systems and their Applications* 13.4 (1998), pp. 18–28. DOI: 10.1109/5254.708428.

[7] Faliang Huang et al. “Attention-Emotion-Enhanced Convolutional LSTM for Sentiment Analysis”. In: *IEEE Transactions on Neural Networks and Learning Systems* 33.9 (2022), pp. 4332–4345. DOI: 10.1109/TNNLS.2021.3056664.

[8] Bac Le and Huy Nguyen. “Twitter Sentiment Analysis using Machine Learning Techniques”. In: *Advanced Computational Methods for Knowledge Engineering: Proceedings of 3rd International Conference on Computer Science, Applied Mathematics and Applications-ICCSAMA 2015*. Springer. 2015, pp. 279–289.

[9] Gilles Louppe. *Understanding Random Forests: From Theory to Practice*. 2015. arXiv: 1407.7502 [stat.ML].

[10] Pekka Malo et al. “Good debt or bad debt: Detecting semantic orientations in economic texts”. In: *Journal of the Association for Information Science and Technology* 65 (2013). URL: https://api.semanticscholar.org/CorpusID:7700237.

[11] *Movie Review Analytics, Sentiment & Ratings*. Accessed 21-01-2024.

[12] GSN Murthy et al. “Text based sentiment analysis using LSTM”. In: *Int. J. Eng. Res. Tech. Res* 9.05 (2020).

[13] Chi Sun et al. “How to Fine-Tune BERT for Text Classification?” In: (2020). arXiv: 1905.05583 [cs.CL].

[14] Jenq-Haur Wang et al. “An LSTM approach to short text sentiment classification with word embeddings”. In: *Proceedings of the 30th conference on computational linguistics and speech processing (ROCLING 2018)*. 2018, pp. 214–223.

[15] Zhilin Yang et al. “XLNet: Generalized Autoregressive Pretraining for Language Understanding”. In: (2020). arXiv: 1906.08237 [cs.CL].