Sentiment Analysis on Amazon Polarity Dataset: A Comparative Study of Feature Extraction Techniques and Machine Learning Methods

Morteza Layegh Mirhosseini
220003166 - PG
MSc Data Science
Project Code: https://drive.google.com/drive/folders/1fSNvt5djPDQO9Dj3X3H02rn4aIys5o4L?usp=share_link
Morteza.layegh-Mirhosseini@city.ac.uk

1 Problem statement and Motivation

Sentiment analysis, also known as opinion mining, is a common task in natural language processing that involves extracting emotions, attitudes, and sentiments from text data in different contexts. It can be a valuable tool for companies to automatically analyze customer feedback and improve their services[1]. In this study, our goal is to create a sentiment analysis pipeline for Amazon reviews.

We will explore various methods for data pre-processing, feature extraction, and machine learning-based sentiment analysis techniques. We will then critically evaluate and compare the outcomes obtained. Our research will be guided by the following question:

Which feature extraction method and machine learning model combination yields the highest performance for sentiment analysis of Amazon reviews, and how does it compare to a fine-tuned DistilBERT model?

In this study, various feature extraction techniques are explored, including Bag of Words with N-grams, TF-IDF, and two pre-trained word embedding models, namely Google News Word2Vec and Wikipedia GloVe. Two machine learning models, Multi-layer Perceptron (MLP) and Support Vector Machine (SVM), are also trained using the aforementioned feature extraction methods. Additionally, a fine-tuned DistilBERT (Bidirectional Encoder Representations from Transformers) model is developed. The results are then critically evaluated to identify the best-performing models. The selected models are used to create an optimal sentiment analysis pipeline.

2 Research hypothesis

In this study, we have performed some experiments with different feature extraction to convert text into a numerical representation. These methods include Bag-of-Words (BOW) with N-grams, TF-IDF, word2vec, and GloVe. Once the text is converted to a numerical representation, it is fed into two machine learning models, namely SVM and MLP, to train a sentiment analysis classifier. The effectiveness of different feature extraction methods can be measured using different classification metrics such as accuracy, precision, recall, and f1 score. Additionally, we fine-tune a DistilBERT model for our task.

Our hypothesis is that the fine-tuned BERT model will outperform the feature extraction methods and machine learning models for sentiment analysis of Amazon reviews. This is due to its ability to capture contextual relationships between words and phrases, which can lead to more accurate predictions.

3 Related work and background

This section examines several studies being conducted in the field of NLP and sentiment analysis, which are focused on aspects such as pre-processing and methodology.

The text data can be too noisy, containing lots of repeating characters, white spaces, spelling errors, and so on. To reduce the dimensionality of the data and make classification easier for sentiment analysis tasks, the data needs to be cleaned[2]. There are many pre-processing techniques that can be explored and check their effects on the results of a classifier.

Effrosynidis et al. [3] used fifteen different pre-processing techniques on two datasets to evaluate the performance of three machine learning models on a sentiment analysis task. Their results demonstrated that overall stemming, and removal of punctuation and numbers enhance the performance of machine learning models. Moreover, their study revealed that handling capitalized words, replacing slang, replacing negations with antonyms, and spelling correction can reduce the accuracy of sentiment analysis on Twitter data. Additional text preprocessing techniques and their impact on sentiment analysis can be explored in the work of Angiani et al. [4]. In this study, we adopted a similar approach to pre-processing as that of Katić et al., which is elaborated in Section 6 [2].

After text pre-processing, the next step involves choosing a feature extraction method to convert the text into a numerical representation. Feature extraction is a crucial step [1] because it can affect the results of the sentiment analysis directly[5].

In the study conducted by Katić et al.[2], various feature extraction methods were compared in the sentiment analysis of Amazon reviews. The authors experimented with techniques such as Bag of Words, TF-IDF, and GloVe word embedding to train multiple classifiers, including Support Vector Machines, Logistic Regression, and deep learning models such as Convolutional Neural Networks and Long Short-Term Memory models. The results of their analysis revealed that the deep learning models produced the best results.

In this research, we adopted a similar approach to that of Katić et al. [2] by comparing various feature extraction methods for sentiment analysis of Amazon reviews. However, we extended the methodology of Katić et al. by incorporating an additional classifier, i.e. MLP. Additionally, we developed a fine-tuned DistillBERT model[6]. It is worth noting that pre-trained language models such as BERT[7] have demonstrated their ability in reducing the need for feature extraction and achieving remarkable outcomes in sentiment analysis tasks [8, 9].

4 Accomplishments

The following provides an overview of the project's objectives as outlined in the proposal.

Table 1. Accomplishments

Task	Assessment
Literature Review	Completed
Text Pre-processing	Completed
Experimenting with BOW and TF-IDF	Completed
Experimenting with different word embedding models (word2vec and Glove)	Completed
Training a word2vec model	Completed (supplementary materials)
Build and train two different machine learning models using different feature extraction methods and measure their performance.	Completed
Fine-tuning a DistilBERT Model for the sentiment analysis	Completed (not mentioned in the proposal)
Perform error analysis	Completed (not mentioned in the proposal)
Measure the effect of different feature extraction models on the machine learning model performance	Completed

5 Approach and Methodology

The process of analyzing sentiment involves several steps, including data extraction, preprocessing, feature extraction, classification and evaluation. In this section, we will briefly explain how we approached each step in our task.

Concerning the first step, i.e., data extraction, the data was retrieved from the Hugging Face library and used for subsequent analysis. The next step, data preprocessing, involves text processing and feature extraction. Specifically, in this project, the text is processed using techniques such as lowercasing, white-space characters and punctuation removal. This methodology aims to preserve important words while simultaneously reducing the input data's dimensionality for better analysis.

As for feature extraction methods, we concentrate on the following techniques: Bag of Words with N-grams, TF-IDF, word2vec, and GloVe word embeddings. These methods provide a vector representation of each review within the dataset. Specifically, in the case of pre-trained Word2vec and Glove models, the mean of word vectors for all words in a review is computed via mean pooling, resulting in a vector representation for that review.

In the next step, after feature extraction, we feed the resulting samples into two different sentiment classifiers: SVM and MLP. We also fine-tuned a DistilBERT model for our task which does not require feature extraction.

Finally, in the last step, we evaluated the performance of each model using different classification metrics, such as accuracy, precision, recall, and F1 score.

5.1 Feature Extraction Methods:

In this section, we briefly explain each feature extraction technics used in this study.

Bag of words: The bag-of-words (BoW) method is a simple and effective technique for feature extraction. It represents text as a set of words and their frequency of occurrence, ignoring the word order and grammar. The BoW is easy to implement and interpret, but it suffers from a lack of context[10].

TF-IDF: Term Frequency-Inverse Document Frequency is a method for text representation that assigns weight to each word based on how frequently it appears in the document and how rare it is in the larger collection of documents (the corpus). Its pros include capturing important terms, giving more weight to rare terms, and being simple to implement. However, it may not work well with short texts[10].

Word2Vec: Word2vec is a neural network-based feature extraction technique that learns vector representations of words from a large corpus of text. One of its key advantages is that it can capture the semantic and syntactic relationships between words. However, its main drawback is that it requires a large amount of training data and can be computationally expensive[10]. For this reason, a pre-trained word2vec model from Google [11] is used in this study. This model was trained on a large corpus of Google News articles using the word2vec algorithm. Each word is represented as a 300-dimensional vector, and the model contains embeddings for over 3 million words and phrases. For this study, only the 100,000 most frequent words were used.

GloVe: (Global Vectors) is a type of word embedding that is similar to the skip-gram version of Word2Vec. It represents words as high-dimensional vectors based on the co-occurrence statistics of words in a corpus[12]. The advantage of the GloVe model is that it can be trained quickly on large amounts of data, and its implementation can be easily parallelized[10]. The GloVe model used in this study was trained on the Wikipedia 2014 dump and the Gigaword 5 corpus. It contains 300-dimensional vectors for 400,000 words[12].

5.2 Classification Algorithms:

In this study, three different classifiers were employed, namely the Support Vector Machine (SVM), the Multi-Layer Perceptron (MLP), and a fine-tuned DistilBERT model.

SVM: This algorithm is a type of machine learning model that is capable of performing both classification and regression tasks. SVMs are designed to classify data by transforming it into a higher-dimensional space using a kernel trick and then identifying a hyperplane that can effectively separate the classes[10]. In this study, we experimented with a linear kernel function and a penalty value of 1 for this classifier.

MLP: Multi-Layer Perceptrons are highly effective algorithms for detecting non-linear connections within data. They can be customized to suit classification and regression tasks [9], and have the capacity to handle diverse forms of information, such as images and text.

DistilBERT: Fine-tuned DistilBERT is a smaller and faster version of the popular BERT (Bidirectional Encoder Representations from Transformers) model[6]. For a sentiment analysis task, Fine-tuned DistilBERT uses a series of transformer layers to analyze text hierarchically, from single words to documents. In other words, this model encodes the text into a numerical representation that captures the sequence

information and semantic meaning. Then, this information goes through a classification layer that predicts the sentiment label.

6 Dataset

6.1 Introduction to the Dataset

The Amazon Polarity dataset is used in this research, which is readily available on the Hugging Face website[13, 14]. This dataset contains customer reviews on various Amazon products over a period of 18 years and consists of more than 3.5 million reviews. For the purpose of this study, we will use only a subset of this dataset, which will be limited to 20,000 samples. Each sample in the dataset is labeled as either positive (review scores 4 and 5) or negative (review scores 1 and 2). The review score of 3 is ignored.

It is imperative to examine the distribution of labels within the dataset and address any issues arising from class imbalance to prevent models from being biased towards the majority class. Upon investigation of the dataset, it was observed that the distribution of labels is fairly balanced, with the negative class accounting for 51% and the positive class accounting for 49%. As such, there is no need to implement any measures for dealing with class imbalance in this dataset.

6.2 Examples of the Dataset

In this section, two samples of the dataset are provided. The first one is a negative review:

“I am a big JVC fan, but I do not like this model, I was suspiscious when I saw several units in the return section of the store. I bought one anyway (new) and must say I am not happy. The unit sends out clicks to the receiver once in a while, the transition between scenes is not always smooth, (like a little pause) and while it is still fairly new I can't get any DVD,CD or even a DVD headcleaner to work. All I get is a "incorrect disc" message.”

The other example is a positive review:
“Very happy with purchase. Item was shipped fast, nicer than expected and the price was very reasonable. Ping to be a nice Christmas gift.”

It was observed that there are some misspelled words, such as " suspiscious" and "headcleaner," in the negative review. The word "suspicious" can be considered as a decisive word for the classifier in sentiment analysis. The misspelling emphasizes the importance of preprocessing steps, which will be explained in the next section.

6.3 Dataset preprocessing

Preprocessing text in sentiment analysis is necessary to remove noise, reduce dimensionality and ensure data consistency. By cleaning and transforming the text data, preprocessing enables sentiment analysis models to focus on meaningful information and learn accurate patterns.

In this report, a similar approach used by Katic et al [2] for text preprocessing was followed, as they had previously worked on a similar problem and data. Firstly, the text was converted to lowercase and URLs, emails, and repeated vowels were removed from words using regular expressions. The word "not" was substituted for negative contractions such as "can't," "don't," "isn't," "never," etc. This approach helps the classifier model to include more negation bigram constructs that would otherwise be excluded due to their low frequency. Finally, punctuation and white-space character were removed.

7 Baselines

Random guessing is considered as the primary baseline in this report, as simple baselines can serve as a benchmark to assess the performance of more advanced models. They can also offer insights into the minimum achievable performance for the task, and highlight potential shortcomings of the proposed models.

8 Description of the choice of training and evaluation methodology

In this study, a holdout method was employed to assess the performance of each trained model. The dataset was divided into three subsets: training (78%), validation (12%), and testing (20%). The training and validation sets were used during the training phase to examine the impact of different hyperparameters on the classifier's performance.

As the dataset used in this study is balanced, accuracy is an appropriate metric for evaluating the performance of the models. Based on the validation accuracy, the best model (which includes feature

extraction and a machine learning algorithm) was selected.

9 Results and Error Analysis

9.1 Choice of Parameters and Experimental Results

In this section, an overview of the chosen parameters for different feature extraction methods, as well as classifiers, will be provided. Furthermore, the results of each proposed model on the validation and test dataset will be presented.

For the SVM model, a linear kernel function with penalty value (c) equal to one was experimented with different feature extraction methods including BOW with N-grams, TFIDF, word2vec, and Glove. The results on the validation set and test set are presented in Table 2.

Table 2. The results for SVM classifier with different feature extraction methods

Models	Validation Accuracy	Test Accuracy	Precision	Recall	F1
SVM + BOW	0.83	0.82	0.82	0.83	0.83
SVM + BOW-BI	0.86	0.86	0.86	0.87	0.87
SVM + BOW-TRI	0.87	0.86	0.86	0.87	0.87
SVM + TFIDF-BI	0.88	0.88	0.88	0.87	0.88
SVM + Glove	0.78	0.80	0.82	0.78	0.80
SVM + Word2Vec	0.82	0.81	0.84	0.79	0.81

In regards to the MLP classifier, it consists of two hidden layers, with 100 and 50 neurons respectively, and use the ReLU activation function. For the output neuron, the sigmoid activation function is employed. The optimization algorithm used is ADAM, with a learning rate of 0.001. All other parameters are set to their default values in the MLPClassifier of the sklearn library. The models were trained for 5 epochs. The results on the validation set and test set are presented in Table 3.

Table 3. The results for MLP classifier with different feature extraction methods

Models	Validation Accuracy	Test Accuracy	Precision	Recall	F1
MLP + BOW	0.83	0.83	0.85	0.82	0.83
MLP + BOW-BI	0.87	0.87	0.88	0,87	0.88
MLP + BOW-TRI	0.88	0.87	0.89	0.86	0.87
MLP + TFIDF-BI	0.88	0.88	0.88	0.88	0.88
MLP + Glove	0.78	0.79	0.78	0.82	0.80
MLP + Word2Vec	0.81	0.81	0.84	0.77	0.80

Regarding the fine-tuned DistilBERT model, a linear layer is added on top of the pooled output to make it suitable for classification tasks. The Transformers library is used to load the pre-trained model. The Adam optimizer is used with a learning rate of 5e-5. The batch size is set to 24, and the model is fine-tuned over three epochs. The results of the fine-tuned model are presented in Table 4. It is clear that the fine-tuned DistilBERT model outperformed all other models.

Table 4. The results for fine-tuned DistilBERT classifier

Models	Validation Accuracy	Test Accuracy	Precision	Recall	F1
Fine-tuned DistilBERT	0.92	0.91	0.93	0.90	0.91

The results presented in Tables 2, 3, and 4 show that all models performed better than our baseline, which was set at 50% (random guess). The models with the highest performance are highlighted in green, while the ones with the lowest performance are highlighted in orange.

9.2 Analysis and Critical Evaluation of Results

9.2.1 Comparing SVM and MLP Models:

After analyzing the results of the SVM and MLP models (see Table 2 & 3), it is clear that both models with the bi-gram TF-IDF feature extraction performed the best, achieving the highest validation and test accuracy, as well as precision, recall, and F1 score on the test dataset. However, both SVM and MLP models using GloVe and Word2Vec had the lowest performance regarding these metrics. This could be because the GloVe and Word2Vec models used in this study were trained on different datasets - Wikipedia and Google News articles, respectively. As a result, the relationships between words, that these models capture, may not accurately reflect the characteristics of the Amazon review dataset we are analyzing.

It is noteworthy that increasing the number of n-grams in the bag-of-words (BOW) feature extraction method has led to an improvement in the performance of the SVM and MLP models. The results depicted in Figure 1 indicate that the unigram and bigram BOW with the SVM model exhibited lower validation accuracy, 83% and 86%, respectively, compared to trigram BOW with 87%. A similar pattern was observed in the

MLP model which is shown in Figure 2. This may be due to the fact that the increase in n-grams results in capturing more sequential information in the text. As a result, the SVM and MLP classifiers can better discern patterns in the text.

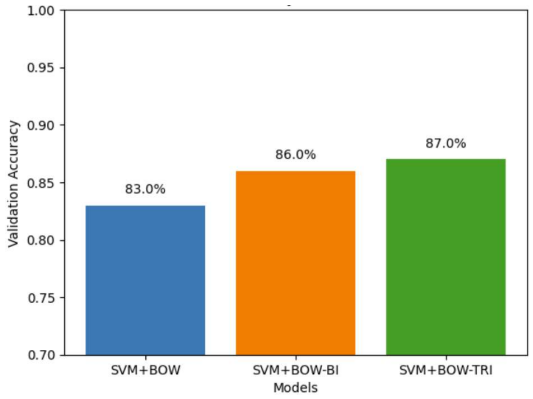

Figure 1. Validation accuracy of SVM models with different BOW n-grams

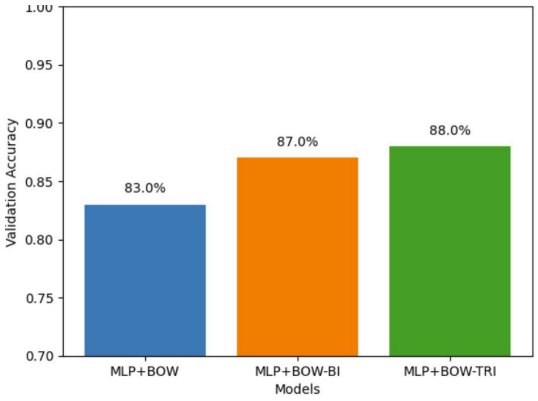

Figure 2. Validation accuracy of MLP models with different BOW n-grams

However, it is noteworthy that in general, SVM and MLP models with BOW had lower accuracy compared to the TF-IDF method. The bi-gram TF-IDF method with both SVM and MLP classifiers yielded 88% in both validation and text accuracy, which is higher than BOW models. This can be attributed to the fact that while the BOW method essentially counts the words, the TF-IDF method assigns weights to the words based on their frequency and importance in the context of the entire corpus. This makes the TF-IDF method more effective in capturing the important words in the text and consequently improving the accuracy of the classifiers.

9.2.2 Comparing the Best Models:

In this study, various feature extraction techniques were tested using two distinct machine learning models. Additionally, a DistilBERT model was fine-tuned for our specific task. In this section, three of the best models were selected for further comparison, including SVM with TF-IDF, MLP with TF-IDF, and the fine-tuned DistilBERT model.

Figure 3 depicts the test accuracy of the models, and it is evident that the fine-tuned BERT model outperformed the other two models by three percent, yielding 92% accuracy.

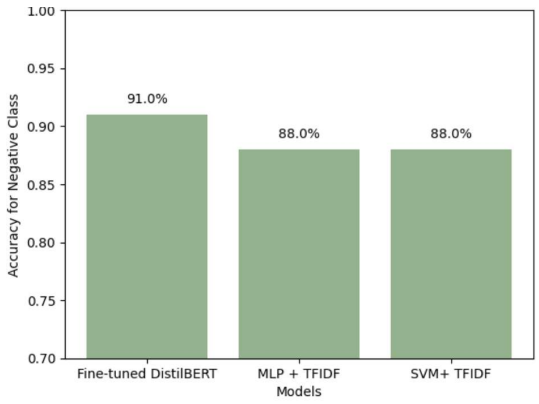

Figure 3. Test accuracy of top 3 models

Our models aim to predict whether a product review is positive or negative. However, it is crucial to better predict negative reviews as it enables sellers to identify the weak spots in their products and improve them. Hence, precision and recall for the negative class are appropriate metrics to evaluate the models' performance.

Figure 4 illustrates the precision score for the negative class for each model, and it indicates that the fine-tuned BERT model outperformed the other two models with a score of 0.89. This means 89% of the negative predictions were correct.

Figure 5 displays the recall score for the negative class for each model. This metric demonstrates the number of negative instances that correctly identified by the classifier in the dataset. Once more, the fine-tuned BERT model has an advantage with a score of 0.93 compared to 0.87 and 0.88 for the MLP and SVM classifiers, respectively.

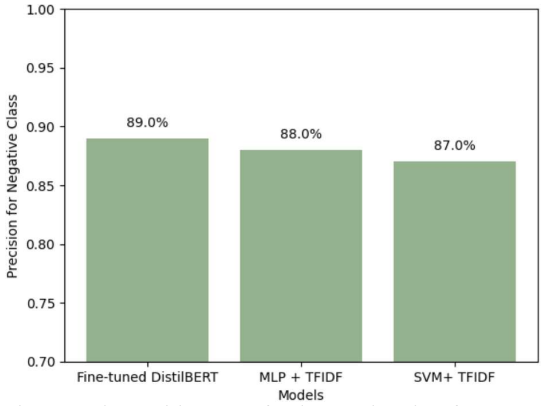

Figure 4. The precision score for the negative class for top 3 models

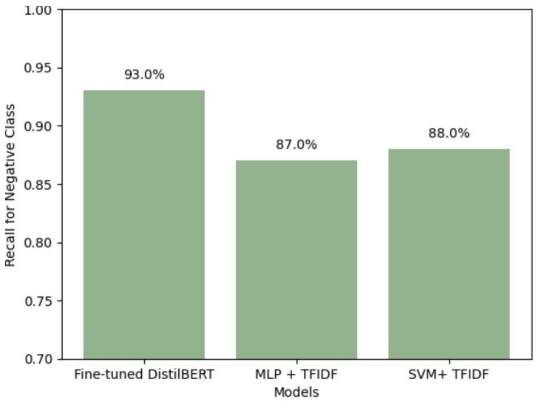

Figure 5. The recall score for the negative class for top 3 models

9.3 Error Analysis:

This section aims to conduct an analysis of some of the misclassifications made by the best model ,which is a fine-tuned DistilBERT with 91% accuracy. Following is a similar approach taken by AlQahtani et al.[8]. The misclassified reviews were sorted based on the highest loss. The objective is to gain an understanding of why the model made these errors. Upon investigating the top 50 instances, the most interesting cases have been identified and presented in Table 5.

Regarding review number 1 and 2, it is evident that the customers expressed a positive opinion. In fact, the classifier correctly identified these reviews as positive, while the true label was mistakenly given as negative in the dataset.

After analyzing other instances that were misclassified as positive reviews (review number 3,4,5), it was discovered that while these reviews were negative, customers had expressed some positive feedback regarding their previous experiences or other products. It is likely that these positive remarks caused the model to make an error, as they were highlighted in the text of the reviews.

The same pattern happens for some reviews that were misclassified as negative (reviews 6, 7, and 8). As highlighted in the text, some of the words used in the reviews may convey a negative sentiment, while the overall review is, in fact, positive.

Table 5. Error Analysis

#	Review	True label	Predicted label
1	A fascinating look at Franklin's life in his own voice. It is unfortunate that he did not narrate the last 30 years of his life as well.	0	1
2	Utterly amazing and unique sound. Her voice is so powerful.	0	1
3	If you like early Mark Harmon, then it is worth the time. Otherwise, it is a made for TV style situation comedy. There are some embarrassing moments and a heart-warming ending.	0	1
4	These pants were washed one time, following the washing instructions, they were not put in the dryer, and they shrunk not only around, but lengthwise. I had purchased these same pants a year or so ago and didn't have any problems. I look forward to your reply	0	1
5	While this album isn't bad, several of the songs also exist on the Everything I Need album. I own both and in my opinion, the Everything I Need versions of these songs are much better	0	1
6	I didn't like some of the plot twist, of a former leader of the Autobots but other than that it was pretty good.	1	0
7	It costs a lot because its an IMPORT, not a bootleg! If you're going to enjoy the CD, might as well enjoy it knowing that you didn't pay too much money for a crook to burn the CD for you. Honestly. Things cost money. To a real fan, this is worth the money. Costs a crook a penny to create a mediocre bootleg. Do you really think you're saving money	1	0
8	I purchased this book because I was writing a novel about the Native Americans who lived in the Finger Lake region f New York State. The book was helpful as far as gaining some insights into the life styles of these people. I really was not reading this book for its entertainment value	1	0

10 Lessons learned and Future Work

This study has revealed several critical points. Firstly, it is essential to use rules of thumb and refer to previous studies for pre-processing and tuning hyperparameters. This approach can save

time and lead to better results. Additionally, conducting a manual investigation of the predictions made by sentiment analysis models is important in understanding the reasons behind the model's mistakes.

As a suggestion for future work, a Word2Vec model could be trained specifically for Amazon reviews dataset to measure the classifiers performance based on an embedding that is specifically trained on the Amazon dataset. Additionally, the fine-tuned model trained in this study can be tested on other datasets, such as IMDB reviews, to assess whether the model generalizes well on the other datasets. The MLP and SVM models trained in this study can be further tuned by exploring different parameters through random grid search or Bayesian optimization.

11 Conclusion

Sentiment analysis is a popular approach used to investigate information from text data in various industries and businesses. With the vast amount of data available through customer reviews and comments on different platforms, developing sentiment analysis algorithms can allow businesses to quickly and efficiently analyze why their customers are satisfied or dissatisfied with their services or products. This information can then be used to focus resources on improving areas of weakness in their business. In this study, sentiment analysis was conducted on the Amazon Polarity dataset. Different feature extraction techniques were experimented with, including BOW with n-grams, TF-IDF, and word embeddings, namely Word2Vec and GloVe, to train two classifiers, SVM and MLP. Additionally, a DistilBERT model was fine-tuned for this task. Overall, 13 models were trained using different combinations of feature extraction methods and machine learning algorithms. The top three performing models were SVM with bi-gram TF-IDF, achieving an accuracy of 88%; MLP with bi-gram TF-IDF, also achieving an accuracy of 88%; and the fine-tuned DistilBERT model with the highest accuracy of 91%.

References

1. Ahuja, R., et al., *The impact of features extraction on the sentiment analysis.* Procedia Computer Science, 2019. **152**: p. 341-348.
2. Katić, T. and N. Milićević. *Comparing sentiment analysis and document representation methods of amazon reviews*. in *2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*. 2018. IEEE.
3. Effrosynidis, D., S. Symeonidis, and A. Arampatzis. *A comparison of pre-processing techniques for twitter sentiment analysis*. in *Research and Advanced Technology for Digital Libraries: 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017, Proceedings 21*. 2017. Springer.
4. Angiani, G., et al. *A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter*. in *KDWeb*. 2016.
5. Avinash, M. 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*. 2019. Springer.
6. Sanh, V., et al., *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.* arXiv preprint arXiv:1910.01108, 2019.
7. Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018.
8. AlQahtani, A.S., *Product sentiment analysis for amazon reviews.* International Journal of Computer Science & Information Technology (IJCSIT) Vol, 2021. **13**.
9. Selvakumar, B. and B. Lakshmanan, *Sentimental analysis on user's reviews using BERT.* Materials Today: Proceedings, 2022. **62**: p. 4931-4935.
10. Birjali, M., M. Kasri, and A. Beni-Hssane, *A comprehensive survey on sentiment analysis: Approaches, challenges and trends.* Knowledge-Based Systems, 2021. **226**: p. 107134.
11. Mikolov, T., et al., *Distributed representations of words and phrases and their compositionality.* Advances in neural information processing systems, 2013. **26**.

12. Pennington, J., R. Socher, and C.D. Manning. *Glove: Global vectors for word representation*. in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.
13. Face, H. *amazon_polarity*. Available from: https://huggingface.co/datasets/amazon_polarity#citation-information.
14. McAuley, J. and J. Leskovec. *Hidden factors and hidden topics: understanding rating dimensions with review text*. in *Proceedings of the 7th ACM conference on Recommender systems*. 2013.