An Opinion Mining-Based Pretrained Model Analysis Depending on Multiple Thinking Patterns

Joydhriti Choudhury, Abdullah Al Farabe, Adria Binte Habib, Muhammad Iqbal Hossain

Department of Computer Science and Engineering, BRAC University, 66 Mohakhali, Dhaka 1212, Bangladesh

Corresponding author: Abdullah Al Farabe, Email: abdullah.farabe@gmail.com

People love to express their opinions online, and the use of social media for this purpose has skyrocketed in the Internet age. Customers also frequently share their first-hand knowledge of the goods or services they have used. Reviews can be either favorable or bad, and both have an impact on people's decisions and businesses. Therefore, it is essential to predict people's opinions in order to preserve and understand the reliability of online review systems. Our study concentrates on a certain kind of model, namely an opinion mining-based pre-trained model, and its analysis takes different types of thought patterns into account. Pretrained models based on transformers, such as BERT, RoBERTa, and DistilBERT may be effective at classifying people's opinions. Other classifiers, such as DISTILBERT and RoBERTa, offer more accuracy compared to the rest of the classifiers employed by researchers but in our model, RoBERTa offers the best accuracy at 90%.

Keywords: EDA, tokenization, BERT, RoBERTa, DistilBERT

2023. In Satyasai Jagannath Nanda & Rajendra Prasad Yadav (eds.), *Data Science and Intelligent Computing Techniques*, 317–327. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.56155/978-81-955020-2-8-28

1 Introduction

Today, millions of people share their daily thoughts and emotions on social networking sites such as Instagram, YouTube, Facebook, and others. There is a wealth of sentiment-rich data generated through blog posts, status updates, comments, reviews, and other forms of the social media industry. The process of looking at text data to ascertain its purpose is known as sentiment analysis [1]. Contextual text mining is what distinguishes and draws subjective data from the original content. It helps companies comprehend how the public feels about their names, goods, and services while monitoring Internet conversations. The online content Sentiment Analysis (SA) is generally used to represent the core elements of sentiment analysis, including views, sentiments, evaluations, attitudes, and emotions. Figure 1 shows the whole diagram of our proposed model. We have taken a total of five datasets on which we have used models to identify the performance based on opinion mining. Whenever people work on datasets there is a very high chance to have a few unnecessary data which causes a bad impact on the results. So, in starting we have to preprocess our datasets through EDA (Exploratory Data Analysis) which helps to remove unnecessary values from the datasets.

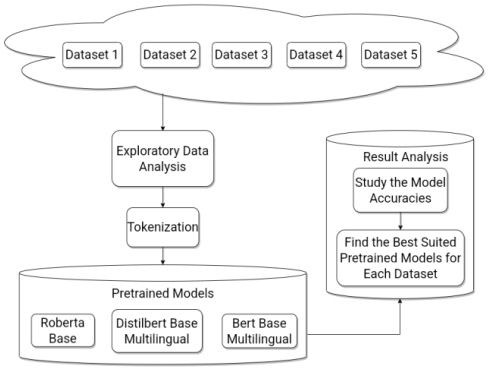

Figure 1: System Block Diagram

By doing EDA we try to remove the null value and our main concern is to count five-star ratings, count words and average use of words in each tag, and see comment length frequency. Unnecessary values refer to null values and avoid those words which are not related to emotions. After checking dropping null values we have shown the counting ratings as percentages through a pie chart that indicates from 1-star to 5-star reviews individually. Then we look after the type of our dataset to see if they are- biased, semi-biased or unbiased. Then We have tallied the words from phrases that are either unnecessary or crucial to comprehending the review from the review attribute. From the databases, we have determined the total number of words and total words in each comment's tag. Moreover, we have measured the highest using words and counting word values for calculating the word frequency. Our main purpose is to ignore the prepositions or pronoun words that have been used casually so many times which does not indicate any emotions. For this reason, we have ignored these preposition and pronoun words and have found the length of text based on the frequency that indicates a relation pattern between sentence length and word of reviews. After data analysis, data can be tokenized by dividing up different types of text input into tokens. As our dataset includes punctuation and other special characters they are deleted from the input data during tokenization because they don't affect the analysis's correctness. We are using tokenization because these tokens aid in context comprehension for NLP (Natural language processing) [2,3]. Then we are using three models named - Roberta, Distilbert, and BERT. In the end, we have analyzed with graphs and accuracy which model gives the perfect opinion mining to review.

Our motivation is that future customers' purchasing decisions are influenced by the reviews posted on internet platforms, and this has a financial impact on firms. People enjoy leaving reviews and like to express their opinions on social media. Building a trustworthy online review system can often be challenging when it comes to text classification. Finding the behavior of the models based on the types of datasets can help us to identify the proper and correct model which we can use to analyze the opinions of the customers more accurately and this is our target.

2 Literature Review

Rain, C. [4] aims to extend and apply recent advances in sentiment analysis and natural language processing to data obtained from Amazon. The goal of the experiment was to compare how well each participant could tag reviews accurately. Classification mistakes and general issues with feature selection are investigated and explored.

Kim, J., [5] has focused on age, gender, and negative reviews and is a woman's review data to understand the propensity for Amazon clothing purchases. A transformer model, TF-IDF, has been applied. Attempts were made to vectorize text in each approach in a variety of ways, and the accuracy of the methods was measured using MAE and the same model as random forest. The accuracy has significantly increased for vectors coupled with Embedding and Sentiment, though. TF-IDF, which has a 78% accuracy rate.

Pramudya.,W.B.N [6] has worked on the e-commerce site Flipkart, and the Apple iPhone SE customer reviews. The researcher initially concentrated on setting text to lowercase and then stemmed text in the text processing. The tf-idf transformer has also been used to calculate tf-idf scores on documents within their "training" dataset when it needs the word frequency vectors for various purposes. The Naive Bayes Model, which yields a 68% accuracy, has been utilized.

Singh, H., [7] research is based on sentiment analysis, one of the most well-liked NLP tools that are crucial to data analytics. Experimenting EDA process the classifications such as Random Forest - Count Vectors, Random Forest - TF-IDF, MultinomialNB - Count Vectors, and MultinomialNB - TF-IDF are used from where the RandomForest Classifier Count Vectors gives the best accuracy of 77.8%.

Dictionary classifications based on semantic orientation are used in unsupervised semantics-based techniques to categorize various word kinds [8]. There has been research on the particular behavior of biased, unbiased, and semi-biased datasets. We attempted to work on every type of behavior dataset at once to show which behavior of the dataset is appropriate for our classifier and to measure accuracy. Our main goal is to work on the various datasets' behaviors independently, to apply our classifiers to each one of them to determine how well people's opinions can be classified, and to show the accuracy comparisons between the literature review model and our model.

3 Methodology And System Implementation

1.1 Details of Dataset

We have taken a few datasets based on reviews from consumers. In our daily lives, we buy clothes, electronic devices or apps that have become a regular part of our lives and consumers give reviews based on their satisfaction or disappointment. The datasets we have chosen are named - amazon preprocessed kindle book reviews [9], amazon women's dresses reviews [5], apple iPhone SE reviews [6], and Spotify reviews [7]. Each of these datasets includes features that measure peoples' ratings, reviews, comments, and other feedback.

1.2 Exploratory Data Analysis (EDA)

EDA's major goal is to encourage data analysis before making any assumptions. It can assist in finding glaring errors, better understanding data patterns, spotting outliers or unusual occurrences, and discovering intriguing relationships between the variables. Exploratory Data Analysis (EDA) [10] is an approach for data analysis that employs a variety of techniques -

1. Null value drop from datasets
2. Counting of five-star ratings and Pie Chart based on ratings;
3. Counting words and Average use of words in each review tag;
4. Comment length frequency

Null value drop: From every dataset, we have planned to ignore data consisting of null values. In the EDA process first, we have dropped out those columns which consist of a null value because they are not useful for further procedures. For null values; the model sometimes does not give good accuracy.

Table 1: Changes in the datasets after dropping the null values

Dataset	Before checking null values	After dropping null values
amazon preprocessed kindle book review	12000	12000
amazon women dresses reviews	23486	19662
apple iPhone	9713	9713
spotify reviews	9713	9713

After using the dropna() method to drop null values from the datasets, from Table 1 we could see that every dataset remained the same except “amazon women dresses reviews” which consisted of 19662 after deleting null values.

Counting of five ratings: For good service sellers put a rating system on which they could understand their service and product. People who use iPhone, kindle, Spotify, and Women's Dress would like to give reviews through five stars rating system based on using them. Next step, we counted the rating values from 1 star to 5 stars individually out of 100%. We have also shown the pie chart in Figure 2 which consists of counting one to five stars out of 100% from the reviewers. For example, in Figure 2 we have shown the dataset of “amazon women dresses reviews”. The rest of the datasets have proceeded in the same way.

The behavior of datasets: We checked our datasets if they were biased,semi-biased, or unbiased. This is one of the main parts of our model and motivation to work in every type of dataset. Figure 2 makes anyone understand the behavior of datasets. From Figure 2, if we look at datasets such as - “Apple iPhone SE” and “amazon women dresses reviews”, they both are fully biased on only one attribute 5 stars which are 69.9% and 55.2% whereas “Spotify review’ is unbiased compared with these two datasets. In “Amazon preprocessed kindle book review” all attributes are similarly biased to each other and percentages are almost similar which we call semi-biased. Table 2 also shows the behavior of all datasets we have taken for this model.

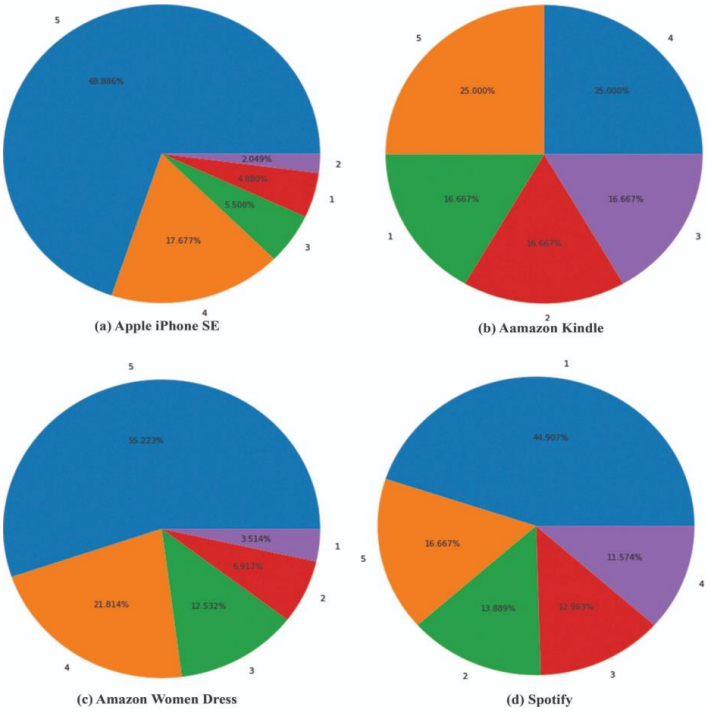

Figure 2: Behavior of dataset

Table 2: Type of dataset

Dataset	Dataset type
amazon preprocessed kindle book review	Semi Biased
amazon women dresses reviews	Biased
apple iPhone	Biased
spotify reviews	Unbiased

Table 2 also refers to the main idea/concept of using our models in different types of datasets to check the workflow rather than working on just one type of dataset.

Counting words and Average use of words in each review tag: We have calculated the total words from the datasets and in Table 3 we have shown for the dataset "amazon women dresses reviews". The total number of words in the dataset is 1219308 and the total words in each comment's tag are given in Table 4 based on ratings.

Table 3: Calculating total words from the dataset

Element	Word Count
5	650737
4	279321
3	162778
2	86462
1	42010

Table 4: Average Words in Each Comment's Tag

Ratings	Average words
5	60
4	65
3	66
2	64
1	61

For example, in the “amazon women dresses reviews” dataset a review is - “Love it; the bit of stretch in the denim makes it less stiff than traditional denim.” From this sentence, words like - Loved, stretch, makes, less, stiff, than, traditional, denim have been counted. So, in Table 4 for five-star ratings people have given reviews that consist of 60 words on average after cutting out other unnecessary words which do not relate to emotions. Moreover, from the rest of the reviews, the same words in a line we have also reduced to calculate the average word from each tag. Our goal is to shrink the sentence as much as we can to achieve success to get good accuracy.

Calculating word frequency: Under sentiment analysis, we have also calculated word frequency to avoid those words that are unnecessary or not related to showing emotions.

Table 5: Frequent Word in Dataset

Highest using Words	Counting word value
i	67392
and	44531
a	43431
it	38174
is	33411
to	21750

this	18556
in	18149
but	14338
on	12332
for	12197
of	12036
with	11345
was	10873
so	10207
my	9770
not	8297
that	8247
I	7913

From Table 5, it indicates that those words are either prepositions or pronoun words that do not mean emotions. As we are working based on emotions, we ought to ignore these words. Because of this, we ignored these emotionless terms before the use of classifiers.

1.3 Sentiment Analysis

We suggest a new family of algorithms that utilize relevant patterns in the input data to automatically discover patterns that may be used to enhance frequency predictions [11]. Through twating the maximum review length is 99 and the highest frequency is near 400. This plot shows the relation pattern between words and sentence length of reviews. Sentiment analysis' main goal is to determine whether customers like or dislike a product based on the review data they have provided [12].

1.4 Model Specification

Pretrained Bert Base: Its pre-training data comes from a sizable, unlabeled text corpus that includes both Wikipedia and a corpus of books. BERT is used in two different ways [12]. One is Masked Language Prediction, which involves making a small number of the input text's words before feeding them into the BERT model to predict the masked words. Next Sentence Prediction is the next strategy. This is possible with effective fine-tuning training.

Pretrained Distilbert: Distilled Bert is superior to BERT in terms of weight, price, size, and speed. DistilBERT employs better performance than the BERT base model [13]. DistilBERT was chosen as one of the models to be used in the classifier because it has a minimum resource need, which is in line with the goal of this research. These large-scale models improve performance significantly, but they frequently have millions of parameters.

Pretrained Roberta Base: The collaborative character of AI research is further highlighted by BERT. Because of Google's open release, it was possible to replicate BERT's performance and identify

areas for improvement. Modern results on the widely recognized NLP benchmark, General Language Understanding Evaluation, are produced by this optimized approach, RoBERTa (GLUE) [14,15].

During every model, we checked for the four datasets. After tokenization and showing the dataset types we used these models to find our final step of accuracy. We have tried to convert the reviews/comments to positive or negative signs. We have referred to them as numbers such as -
Negative=0
Neutral=1
Positive =2

The main reason for taking negative, positive, and neutral reviews as numbers is to fit the reviews of customers converting from characters to numeric numbers. Numeric numbers will help to find the final result of accuracy. Then with the converted training dataset, we also trained the dataset to find the accuracy.

4 Result Analysis

We collected the accuracy for all three classifiers for all datasets and presented them in Table 6. From the tables, it can be seen that the accuracy depends largely both on the model used and on the datasets fed. Furthermore, it's also known that the datasets "apple iPhone SE reviews" and "amazon women dress" are highly biased towards the rating of 5 in Figure 2. Table 6, shows that the biased datasets give higher accuracy than the unbiased data. The "amazon kindle book review" dataset is accurately unbiased and as a result, the accuracy of the model is the lowest compared to others. Biased dataset works better for the classifiers comparable to unbiased and semi-biased datasets. For the biased dataset, all classifiers crossed 79% accuracy. Now our focus is to find out for each dataset which classifier gives the best accuracy. So, from Table 6, we have collected the best classifiers we have used on each dataset we display in Table 7. For example, in Table 6 for the “apple iPhone SE” dataset we have used DISTILBERT, RoBERTa, and BERT which give accuracy of 88%, 90%, and 79%. In this case, we can come to the result that for “apple iPhone SE” dataset, RoBERTa gives the best accuracy of 90% which is put in Table 7. In exact same way, for amazon preprocessed Kindle book review, amazon women dresses reviews, and spotify reviews datasets by analyzing Table 6 we have figured out the best possibility classifier with accuracy put into Table 7.

Table 6: Finding accuracy using classifiers

Models	Datasets	Accuracy (%)
BERT	amazon preprocessed kindle book review	60
	amazon women dresses reviews	80
	apple iPhone SE reviews	88
	spotify reviews	72
DISTILBERT	amazon preprocessed kindle book review	73
	amazon women dresses reviews	80

	apple iPhone SE reviews	79
	spotify reviews	76
RoBERTa	amazon preprocessed kindle book review	71
	amazon women dresses reviews	83
	apple iPhone SE reviews	90
	spotify reviews	79

Table 7: Best classifier with accuracy for each dataset

Dataset	Best classifier	Accuracy
amazon preprocessed kindle book review	DISTILBERT	73
amazon women dresses reviews	RoBERTa	83
apple iPhone SE reviews	RoBERTa	90
spotify reviews	RoBERTa	79

Table 7 shows that for most of the datasets, the RoBERTa classifier gives the best accuracy. Now Table 8 defines the research of authors from whom we have got idea comparing with our model.

Table 8: Comparative Analysis

Dataset	Literature Review citation	Accuracy of the researchers	Accuracy of my model's best classifier's
amazon preprocessed kindle book review	Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning [4]	80	73
amazon women dresses reviews	Amazon Reviews on Women Dresses [5]	78	83
apple iPhone SE reviews	Apple Iphone Reviews & Ratings [6]	69	90

spotify reviews	Spotify Reviews Sentiment Analysis [7]	77.8	79

Table 8 compares our findings to the researchers we have discussed in the literature review. Compared to the Opinion Mining-Based Pretrained Model Analysis, BERT doesn't seem to be a good substitute. In comparison to the rest of the classifiers used by researchers and our classifier the BERT as well, other classifiers like DISTILBERT and RoBERTa provide superior accuracy. For example, in Table 8 in the literature review "Apple iPhone Reviews & Ratings " [5] has worked on the dataset "apple iPhone SE reviews" and they have got an accuracy of 78% where our model RoBERTa, DISTILBERT, and BERT provide accuracy of 90%, 79% and 88% in Table 6. The best accuracy in our model is 90% given by RoBERTa.

So, from Table 2 and Table 6, we might discuss our overall project's goal, which was to determine how the models behaved whether based on biased or unbiased datasets. The EDA preprocess enables us to fit the datasets in such a manner that we were able to eliminate all those extraneous words that were not required or applicable to work with this model before inputting our results.

5 Conclusion

To conclude, future customers' purchasing decisions are influenced by the reviews posted on internet platforms, and this has a financial impact on firms. Recognizing reviews is essential because spam reviews that are generated with a specific goal might mislead customers. To create a review classifier, we used transfer learning and the transformer-based pre-trained models BERT, RoBERTa, and DistilBERT. In order to build a classifier that can identify fraudulent reviews broadly while utilizing the fewest amount of computational resources possible, these models are trained using 10% and 50% of the Yelp dataset. SA is carried out after pre-processing (EDA) to analyze and assess datasets. The pre-trained models are then adjusted for both data samples. Accuracy and assessment metrics are taken into consideration while evaluating the performance of all models and RoBERTa attains a 90% accuracy rate.

However, the datasets we have got from online it seems that there are few data which are previously modified. The data are not authentic enough to make this model perfect. As a result, our classifiers performed with under 70% accuracy because of modifying data. These are the limitations that we want to work on in the future.

Nowadays, unsupervised machine learning is more useful because clustering helps to remove unnecessary data from a dataset. Mean or median is also another process to handle those unnecessary nun values which are more effective than removing null values from the dataset. We want to use more classifiers like - ALBERT and ELECTRA which might give us better results. Our next primary goal will be to build a predictor model that will help organizations make a decision about making or expanding products or not in their business based on a few people's reviews.

References

[1] S. Gupta, 'Sentiment analysis: Concept, analysis and applications', Towards Data Science, 07-Jan-2018. [Online]. Available: https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17. [Accessed: 08-Dec-2022].

[2] Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. Journal of the American Medical Informatics Association, 18(5), 544-551.

[3] De Kok, S., Punt, L., Van Den Puttelaar, R., Ranta, K., Schouten, K., Frasincar, F.: Reviewlevel aspect-based sentiment analysis using an ontology. Proc. ACM Symp. Appl. Comput. 315–322 (2018). https://doi.org/10.1145/3167132.3167163.

[4] Rain, C. (2013). Sentiment analysis in amazon reviews using probabilistic machine learning. Swarthmore College.

[5] kim.J. (2022, October 13). Amazon Reviews on Women Dresses,Kaggle. Available:https://www.kaggle.com/code/jaewook704/amazon-reviews-on-women-dresses

[6] PRAMUDYA.W.B.N, (2021, September 16). Apple Iphone Reviews & Ratings. Kaggle. Available:https://www.kaggle.com/code/bayunova/apple-iphone-reviews-ratings

[7] SINGH.H., (2023, January 19), Spotify Reviews Sentiment Analysis. Kaggle. Available:https://www.kaggle.com/code/harshsingh2209/spotify-reviews-sentiment-analysis

[8] W. Medhat, A. Hassan and H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal 5(4) (2014), 1093–1113

[9] Amazon Kindle Book Review for Sentiment Analysis. (2021, September 3). Kaggle. Available:https://www.kaggle.com/datasets/meetnagadia/amazon-kindle-book-review-for-sentiment-analysis

[10] 1.1.1. What is EDA? (n.d.). https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm

[11] Hsu, C. Y., Indyk, P., Katabi, D., & Vakilian, A. (2019, January). Learning-Based Frequency Estimation Algorithms. In International Conference on Learning Representations.

[12] Geetha, M. P., & Renuka, D. K. (2021). Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model. International Journal of Intelligent Networks, 2, 64-69. https://www.sciencedirect.com/science/article/pii/S2666603021000129

[13] Gupta, P., Gandhi, S., & Chakravarthi, B. R. (2021, December). Leveraging transfer learning techniques-bert, roberta, albert and distilbert for fake review detection. In Forum for Information Retrieval Evaluation (pp. 75-82).

[14] Warstadt, A., Zhang, Y., Li, H. S., Liu, H., & Bowman, S. R. (2020). Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). arXiv preprint arXiv:2010.05358.

[15] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019, July 26). Roberta: A robustly optimized Bert pretraining approach. arXiv.org. Retrieved December 16, 2022, from https://arxiv.org/abs/1907.11692