Sentiment Analysis on Woman's Clothing Reviews Using BERT

Christian Arifin Gouw - F11315020

Abstract

In this experiment, I will be using Bidirectional Encoder Representations from Transformers (BERT) to prepare the data taken from the Women's E-Commerce Clothing Reviews dataset by the user Nicopotatato. The BERT captures deep contextual meaning, improving sentiment prediction performance. The best performance for sentiment analysis task in this dataset is using Bag-of-Words (BoW) approach with 84.0% on macro average on f1-score using either logistic regression or naive bayes as the classifier done by Burhan Y. Kiyakoglu. This experiment shows that using word embeddings as feature extractor in sentiment analysis helps even traditional model such as logistic regression and SVM to achieve great performance.

1 Introduction

The purpose of sentiment analysis is to process a text and classifies it into positive, negative, or neutral sentiments. By having a model that can do sentiment analysis, companies are able to process that data to plan future investments. The text data itself is usually generated from reviews, comments, or even social media posts that are publicly available online.

Sentiment analysis is important due to the massive amounts of data that can be generated from online sources. A human might sometimes misclassify the reviews or they might miss some important keywords that changes the whole sentiment. By having a robust sentiment analysis model, it helps the company to monitor circulating reviews of their products online and in real-time.

In this experiment, I will be using Bidirectional Encoder Representations from Transformers (BERT) to prepare the data taken from the Women's E-Commerce Clothing Reviews dataset by the user Nicopotato.[1] The best performance for sentiment analysis task in this dataset is using Bag-of-Words (BoW) approach with 84.0% on macro average on f1-score using either logistic regression or naive bayes as the classifier done by Burhan Y. Kiyakoglu.

2 Literature Review

2.1 Sentiment Analysis

Due to the development and high usage of the Internet, there has been an increase in textual data generated by online users. In response to this, the field of Natural Language Processing, specifically, sentiment analysis, has received massive growth[2].[3] Sentiment analysis, the field that intersects with linguistic,[4] analyzes the sentiment of online textual data that provide companies or enterprises with knowledge about the opinion of users about the products they are selling.

The level of sentiment analysis, depending on the range of text, can be divided into document level, sentence level, and aspect level. Document-level sentiment analysis treats the whole document as one entity. This means that each document has one sentiment. Sentence-level analysis, as the name suggests, extracts sentiment from the sentence. Classifies neutral sentiment as an objective sentence, while positive or negative sentences as subjective sentences. Lastly, aspect-level analysis, which generally involves extracting and determining the polarity of the term.[2]

There are several methods to build a sentiment analysis classifier; one of them is to use traditional machine learning models. The most promising and famous one being SVM due to its ability to distinguish different hyperplanes that maximize the boundaries of the class. It is also done using naive bayes and logistic regression due to the simpler implementation and smaller computational resource needed.[2]

2.2 Feature Extraction Methods

Feature extraction in sentiment analysis plays an important role in the performance of the model.[5] A representation of the data, like in the form of feature vectors, is needed for the model to learn from the data. Conventional machine learning approaches use feature engineering to build the representation. However, every task requires a different feature engineering approach, making it inflexible.[6]

Representation learning uses deep learning which has been used in many fields such as audio-based,[7] visual-based,[8] and text-based data.[5] In the context of textual data, word embedding is one of the implementation of representation learning, where the representation is transferred to the model of the target task.[6]

In this experiment, feature extraction is performed using BERT. Unlike conventional feature extraction methods such as BoW, BERT captures deep contextual meaning, improving sentiment prediction performance. In solving this, word embeddings were invented.[9] BERT is pre-trained on large corpus with two main self-supervised tasks that are Masked Langauge Model (MLM) and Next Sentence Prediction (NSP).[5] In which case, BERT definitely excels due to its deep contextual understanding.

To capture the actual effectiveness of BERT embeddings, the experiment will use traditional classifiers. In this case, the result can be compared with other approaches that have used BoW as the feature extractor.

2.3 Previous Study

Class	Precision	Recall	F1-Score	Support
(0) Negative	47%	81%	83%	289
(1) Neutral	31%	82%	71%	22
(2) Positive	96%	75%	79%	4215
Average	93%	93%	93%	
Total				4526

Table 1: Previous study classification report.[10]

One publication used the same data set to obtain an average of 93% F1-Score. The model classified into 3 polarities that are positive, neutral, and negative. The approach for their experiment used bidirectional LSTM with GloVe word embeddings. Although performance is satisfactory, the experiment did not use any hyperparameter tuning and did not explore more partitions for the data

set.[10] It could also be observed from the results that the model did not perform well on negative and neutral polarities.

3 Methodology

Code can be accessed: this link.

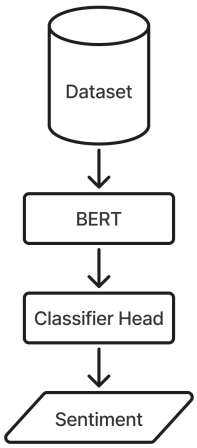

Figure 1: Methodology

3.1 Exploratory Data Analysis

The Women's E-Commerce Clothing Reviews dataset includes 23486 rows and 10 features. The important features that will be used in the experiment are 'Review Text' which is a string field containing the body of the review and 'Rating' which is an integer field containing the product score given by the reviewer.

Figure 2: Bar chart of the rating class

3.2 Dataset

In preparing this dataset, some simple processes is done such as

1. Removing null value in the 'review_text'.
2. Removing neutral polarity reviews.
3. Label mapping where when 'rating' $>= 4$ is considered as positive and when 'rating' $<= 2$ is considered negative.

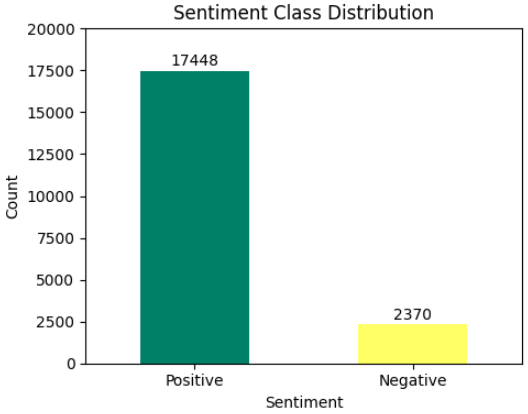

Figure 3: Bar chart of the polarity counts

4. Splitting the dataset into 80% training dataset and 20% testing dataset.
5. Tokenize the input text for the model using BertTokenizerFast tokenizer from the Hugging Face Transformers library, based on the BERT base uncased model.

3.3 Feature Extraction

As shown in Fig. 1, the experiment used the BERT-base model used as feature extractor which has approximately 110M parameters and 12 transformer layers. However, since we only use BERT for embedding extraction (without fine-tuning), the computational burden lies mostly in inference. The downstream classifiers are lightweight, with logistic regression having negligible parameter count in comparison.

3.4 Classifier Head

To implement the classifier head, three widely-used and well-established models are used: Logistic Regression, Naive Bayes, and Support Vector Machine (SVM). These models were chosen for their simplicity, interpretability, and strong performance on text classification tasks.

3.5 Evaluation Metric

From the Fig. 2 and Fig. 3, it can be observed that the dataset is imbalanced with the majority on positive polarity. Therefore, accuracy will not be able to show the effectiveness of the model.[11] To evaluate the effectiveness of the model, F1-score, precision, and recall will be used.

4 Results and Discussion

4.1 Baseline

Classifier Head	Precision	Recall	F1-Score
Logistic Regression	86%	81%	83%
Naive Bayes	68%	82%	71%
SVM	87%	75%	79%

Table 2: Macro average baseline result.

Using dataset as it is (ignoring the class imbalance problem), results in table 2. In using the macro, the minority class has the same weight as the majority class.

Logistic Regression			
Class	Precision	Recall	F1-Score
0 (Negative)	76%	65%	70%
1 (Positive)	95%	97%	96%

Table 3: Logistic regression classification report.

Comparing the macro F1-Score, logistic regression got the best result of 83%. This indicates that there is a good balance between precision and recall across both classes. However, from table 3, it is observed that the model was weaker on the minority class. This is to be expected with imbalanced dataset.

4.2 Handling Imbalance Dataset: Under Sampling

Classifier Head	Precision	Recall	F1-Score
Logistic Regression	88%	88%	88%
Naive Bayes	84%	84%	83%
SVM	88%	88%	88%

Table 4: Macro average after under sampling result.

After addressing the class imbalance problem by using under sampling, it is observed from table 4 the effectiveness from both logistic regression and SVM models are equal.

5 Conclusion

This experiment shows that using word embeddings as the feature extractor in sentiment analysis helps even traditional model such as logistic regression and SVM to achieve great performance. Both models performs better than the BoW approach, with 88% macro average on F1-Score.

6 Future Work

This experiment focused only on binary sentiment classification. Future directions include:

- Extending the model to handle neutral polarity for multi-class classification.
- Applying stratified K-fold cross-validation for more robust evaluation.
- Experimenting with BERT fine-tuning to explore performance improvements.
- Testing on other datasets to verify generalizability.
- After listening to the other groups presentation, oversampling and using other features than 'Review_Text' might help with the performance of the model.

References

[1] Brooks N. Women's E-Commerce Clothing Reviews; 2018. https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews.

[2] Mao Y, Liu Q, Zhang Y. Sentiment analysis methods, applications, and challenges: A systematic literature review. Journal of King Saud University - Computer and Information Sciences. 2024;36(4):102048. Available from: https://www.sciencedirect.com/science/article/pii/S131915782400137X.

[3] Wankhade M, Rao ACS, Kulkarni C. A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review. 2022 Feb;55(7):5731–5780.

[4] Taboada M. Sentiment analysis: An overview from linguistics. Annual Review of Linguistics. 2016 Jan;2(1):325–347.

[5] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding; 2019. Available from: https://arxiv.org/abs/1810.04805.

[6] Liu Z, Lin Y, Sun M. In: Representation Learning and NLP. Singapore: Springer Nature Singapore; 2020. p. 1-11. Available from: https://doi.org/10.1007/978-981-15-5573-2_1.

[7] Liu X, Mei X, Huang Q, Sun J, Zhao J, Liu H, et al. Leveraging Pre-trained BERT for Audio Captioning. arXiv preprint arXiv:220302838. 2022. Available from: https://arxiv.org/abs/2203.02838.

[8] Bao H, Dong L, Piao S, Wei F. BEiT: BERT Pre-Training of Image Transformers. arXiv preprint arXiv:210608254. 2022. Available from: https://arxiv.org/abs/2106.08254.

[9] Smairi N, Abadlia H, Brahim H, Lejouad Chaari W. Fine-tune BERT Based on Machine Learning Models for Sentiment Analysis. Procedia Computer Science. 2024;246:2390-9. 28th International Conference on Knowledge Based and Intelligent Information and Engineering Systems (KES 2024). Available from: https://www.sciencedirect.com/science/article/pii/S1877050924025766.

[10] Agarap AF. Statistical Analysis on E-Commerce Reviews, with Sentiment Classification Using Bidirectional Recurrent Neural Network (RNN); 2020. Available from: https://arxiv.org/abs/1805.03687.

[11] Ruman. Precision, Recall and F1 Explained [with 10 ML Use Case]; 2024. Medium. https://rumn.medium.com/precision-recall-and-f1-explained-with-10-ml-use-case-6e