Sentiment Analysis of Customer Reviews: A Comparative Analysis

Karim Hamdar
University of Padua
karimeugenio.hamdar@studenti.unipd.it

Luca Tusini
luca.tusini@studenti.unipd.it

Davide Christian Mancosu Bustos
davidechristian.mancosubustos@studenti.unipd.it

Abstract

This report presents a comparative analysis of sentiment analysis techniques applied to customer reviews, utilizing VADER, BERT, RoBERTa, and a multi-label logistic regression model. Our study focuses on classifying sentiments into positive, negative, and neutral categories using the Amazon US Customer Reviews Dataset. We highlight the superior performance of transformer-based models, BERT and RoBERTa, which demonstrate enhanced contextual understanding and sentiment classification accuracy compared to traditional methods like VADER and logistic regression. The findings underscore the transformative potential of advanced NLP models in effectively deciphering consumer sentiments.

1. Introduction

Sentiment analysis is a critical task in natural language processing that involves extracting sentiments from textual data to understand the attitudes and emotions expressed. In the context of consumer reviews, sentiment analysis plays a pivotal role in deciphering customer opinions and feedback. With the exponential growth of e-commerce platforms like Amazon, where customer reviews and ratings influence purchasing decisions, the ability to accurately classify sentiments as positive, negative, or neutral has profound implications for businesses.

Positive reviews can enhance product visibility and consumer trust, directly impacting sales and brand reputation. Conversely, negative reviews may deter potential customers, highlighting areas for product improvement and customer service enhancement.

The goal of our project is to conduct sentiment analysis on consumer reviews using the Amazon US Customer Reviews Dataset. Specifically, we aim to classify these sentiments into positive and negative categories. We adopt a comparative approach, fine tuning and evaluating the performance of various sentiment analysis models: VADER, a lexicon-based method; BERT and RoBERTa, advanced transformer-based models; and a multi-label logistic regression model. Additionally, we explore the impact of different preprocessing techniques on model performance.

Our primary objectives include assessing the effectiveness of these models and understanding the influence of preprocessing steps on sentiment classification accuracy. By addressing these objectives, our research contributes to advancing sentiment analysis methodologies and provides valuable insights for businesses aiming to leverage customer feedback to optimize product development and marketing strategies.

Throughout our investigation, we observed significant variations in the performance of sentiment analysis models. Transformer-based models, particularly BERT and RoBERTa, consistently demonstrated superior accuracy and robustness compared to VADER and traditional logistic regression. Specifically, BERT and RoBERTa exhibited higher precision in identifying both positive and negative sentiments, showcasing their proficiency in contextual understanding and semantic analysis.

By leveraging advanced machine learning models and comprehensive preprocessing strategies, businesses can gain deeper insights into consumer preferences, improve product offerings, and strengthen market competitiveness.

2. Dataset

The dataset used in this project is sourced from the Amazon US Customer Reviews Dataset on Kaggle, which encompasses a diverse collection of consumer reviews across various product categories. Amazon, as a leading e-commerce platform, facilitates customers to provide detailed feedback and ratings on purchased products, thereby generating a rich source of textual data for analysis.

2.1. Data Collection

The dataset consists of 37 files, each corresponding to a different product category, such as Digital Software, Gift Cards, Personal Care Appliances, Mobile Electronics, Major Appliances, Digital Video Games, Software, Musical Instruments, Video, and Digital Music Purchase. Each file contains customer reviews accompanied by star ratings, serving as labels indicating the overall sentiment expressed in the review. The dataset contains 15 variables for each review, our analysis focuses on classifying the sentiment of a review into positive, neutral, or negative based on the review body.

2.2. Preprocessing

Before analysis, the dataset underwent several preprocessing steps to ensure data quality and consistency:

- **Handling Missing Values**: Any reviews with missing essential information were removed. In our case the indipendent variable review body had 217 NA values removed from the dataframe.

- **Noise Removal**: Special characters, irrelevant information, and excessive whitespace were removed to clean the text data.

- **Balancing the Dataset**: Initially, the dataset exhibited significant class imbalance, with a majority of reviews categorized as positive sentiment. This imbalance can potentially bias the performance and interpretation of sentiment analysis models. To mitigate this issue, we balanced the dataset by ensuring an equal number of positive, neutral, and negative reviews, we performed resampling, where we randomly selected an equal number of observations from each sentiment class. Specifically, we selected 20,000 samples from each of the positive, neutral, and negative sentiment classes. This approach helps in creating a more representative dataset for training and evaluating sentiment analysis models.

- **Text Normalization**: Techniques such as tokenization, stopword removal, and lemmatization were applied to standardize the textual data and enhance model performance:

  - **Tokenization**: We used `word_tokenize` from NLTK to split text into tokens (words).
  - **Stopword Removal**: To to filter out common English stopwords we used NLTK's stopwords corpus .
  - **Lemmatization**: The code we employed `WordNetLemmatizer` from NLTK to reduce words to their base or root form.

2.3. Data Features

Each data instance consists of the following key features:

- **review_body** (string): The main textual content of the customer review.

- **review_headline** (string): The headline or summary of the review, which provides a concise overview of the sentiment expressed.

- **star_rating** (integer): The numerical rating given by the customer (ranging from 1 to 5 stars), used as the label for sentiment classification.

- **verified_purchase** (string): A flag indicating whether the review was from a verified purchase (Y/N). In our case we considered for our analysis only verified purchase.

2.4. Dataset Size and Distribution

The entire dataset collects more than 130 million reviews. Given computational constraints and the complexity of models like BERT and RoBERTa, we decided to work with a subset of 60,000 reviews in total. To ensure a fair evaluation, we extracted 20,000 positive reviews (4 or 5 stars), 20,000 neutral reviews (3 stars), and 20,000 negative reviews (1 or 2 stars). This balanced dataset helps in training models effectively and avoids bias towards any particular sentiment class.

.

3. Methods

In this section, we present our approach to conducting sentiment analysis on consumer reviews using state-of-the-art natural language processing (NLP) models. Our approach leverages advanced models such as BERT (Bidirectional Encoder Representations from Transformers), RoBERTa, and VADER, alongside traditional machine learning techniques. Each model is evaluated rigorously to determine its effectiveness in capturing nuanced sentiment nuances from e-commerce reviews. We also explore various preprocessing techniques to enhance model performance. The four different models we used for our sentiment analysis task are:

- **VADER**: A rule-based model for sentiment analysis.

- **BERT**: A transformer-based model fine-tuned on the dataset for sentiment classification.

- **RoBERTa**: An optimized version of BERT, also fine-tuned for sentiment classification.

- **Multi-label Logistic Regression**: A traditional machine learning model trained on a TF-IDF representation of the dataset.

3.1. VADER: Valence Aware Dictionary and sEntiment Reasoner

3.1.1 Overview

VADER, short for Valence Aware Dictionary and sEntiment Reasoner, is a lexicon and rule-based sentiment analysis tool specifically attuned to sentiments expressed in social media. VADER is designed to be sensitive to both polarity (positive/negative) and intensity (strength) of sentiment. It uses a combination of a sentiment lexicon (a list of lexical features labeled according to their sentiment polarity) and a set of heuristics that incorporate grammatical and syntactical conventions typically used in social media, such as capitalization, punctuation, degree modifiers, and more.

3.1.2 Architecture

The architecture relies on the following components:

- **Sentiment Lexicon**: A precompiled list of words and phrases, each annotated with a sentiment score. These scores range from -1 (most negative) to +1 (most positive).
- **Heuristics**: Rule-based adjustments applied to the sentiment scores based on context.

3.1.3 Implementation Details

Input Representation VADER requires the input text to be in a plain string format. The text is then parsed and analyzed using the sentiment lexicon and heuristics.

Sentiment Scoring VADER computes a sentiment score for each piece of text. This involves:

- **Lexicon Matching**: Each token is matched against the VADER lexicon to retrieve sentiment scores.
- **Aggregation**: The adjusted scores are combined to produce a compound score, which is a normalized score between -1 (most negative) and +1 (most positive). Additionally, scores for positive, negative, and neutral sentiments are provided.

Evaluation The performance of VADER was evaluated based on metrics such as accuracy, precision, recall, and F1 score. VADER's rule-based approach provides quick and interpretable sentiment analysis but may lack the contextual understanding of more complex models like BERT and RoBERTa.

3.2. BERT: Bidirectional Encoder Representations from Transformers

3.2.1 Overview

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a transformer-based model developed by researchers at Google in 2018. It is designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers. This bidirectional training approach enables BERT to understand the context of a word based on its surroundings, making it highly effective for various NLP tasks, including sentiment analysis.

3.2.2 Architecture

BERT's architecture is based on the Transformer, a model architecture that relies entirely on self-attention mechanisms to compute representations of input sequences. The key components of BERT's architecture are:

- **Transformer Encoder**: BERT uses only the encoder part of the original transformer model. The encoder consists of multiple layers of self-attention mechanisms and feed-forward neural networks. There are two versions of BERT: BERT_base with 12 layers and 110 million parameters, and BERT_large with 24 layers and 340 million parameters.
- **Bidirectional Context Understanding**: Unlike previous models that read text input sequentially (either left-to-right or right-to-left), BERT reads the entire sequence of words at once, allowing it to derive context from both directions simultaneously.
- **Pre-training Tasks**: BERT employs two unsupervised tasks during pre-training:
  - **Masked Language Model (MLM)**: Randomly masks some tokens in the input sequence and attempts to predict them. This helps BERT understand the context of a word based on its surrounding words.
  - **Next Sentence Prediction (NSP)**: Predicts whether a given pair of sentences appears consecutively in the original text, aiding the model in understanding the relationship between sentences.

3.2.3 Implementation Details

In our study, we utilized the BERT_base uncased model due to its balance between performance and computational efficiency. The BERT model was fine-tuned on our sentiment classification task, which involves classifying customer reviews into positive, neutral, and negative categories based on the review body.

Input Representation BERT requires input text to be represented in a specific format, which includes:

- **[CLS]**: A special token added at the beginning of each input sequence to aggregate the representation of the entire sequence for classification tasks.
- **[SEP]**: A special token used to separate sentences or segments within the input sequence.
- **Tokenization**: The input text is tokenized using WordPiece tokenization, which breaks words into subword units to handle out-of-vocabulary words effectively.

Training and Fine-Tuning The pre-trained BERT model was fine-tuned on our dataset with the following setup:

- **Dropout Regularization**: Applied with a dropout rate of 0.1 to prevent overfitting.
- **Softmax Classifier**: A softmax layer was added on top of the BERT model to output probabilities for the sentiment classes.
- **Optimization**: The Adam optimizer was used with a learning rate of 2e-5. The model was trained for 5 epochs, and various learning rates were tested to determine the optimal configuration.
- **Evaluation**: The performance of the model was evaluated based on standard metrics such as accuracy, precision, recall, and F1 score.

Fine-Tuning Strategies During fine-tuning, we applied gradual unfreezing of BERT's layers, starting from the top layers and progressively unfreezing lower layers. This strategy helped in adapting BERT's general language understanding to our specific sentiment classification task.

3.3. RoBERTa: Robustly Optimized BERT Pretraining Approach

3.3.1 Overview

RoBERTa, which stands for Robustly Optimized BERT Pretraining Approach, is an advanced version of BERT developed by Facebook AI. RoBERTa enhances BERT by adjusting key hyperparameters, training with larger mini-batches, and removing the Next Sentence Prediction (NSP) task. These optimizations make RoBERTa more robust and better suited for a wide range of NLP tasks, including sentiment analysis.

3.3.2 Architectural Enhancements and Pre-training

RoBERTa incorporates several key enhancements over BERT:

- **Increased Model Parameters**: RoBERTa uses a larger number of model parameters, totaling 125 million compared to BERT's 110 million.
- **Larger Batch Size**: The training batch size is increased to improve training efficiency and model performance.
- **Extended Training Data**: RoBERTa is trained on significantly more data, totaling 160GB, which includes diverse datasets like BookCorpus, English Wikipedia, CC-News, OpenWebText, and Stories.
- **Dynamic Masking**: Unlike BERT's static masking, RoBERTa employs dynamic masking during training, which changes the masked tokens in each epoch, leading to better generalization.
- **Removal of Next Sentence Prediction (NSP)**: The NSP task is removed, simplifying the training process and allowing the model to focus on the Masked Language Model (MLM) task.
- **Byte-Level Byte-Pair Encoding (BPE)**: RoBERTa uses a larger BPE vocabulary of 50,000 subwords, compared to BERT's 30,000, which helps in better handling of rare words and subword units.

3.3.3 Implementation Details

For our sentiment classification task, we employed the RoBERTa_base model. This model was fine-tuned on our dataset to classify customer reviews into positive, neutral, and negative categories based on the review body.

Training and Fine-Tuning The pre-trained RoBERTa model was fine-tuned on our dataset with the following setup:

- **Dropout Regularization**: Applied with a dropout rate of 0.1 to prevent overfitting.
- **Softmax Classifier**: A softmax layer was added on top of the RoBERTa model to output probabilities for the sentiment classes.
- **Optimization**: The AdamW optimizer was used with a learning rate of 2e-5. The model was trained for 5 epochs, and various learning rates were tested to determine the optimal configuration.
- **Evaluation**: The performance of the model was evaluated based on standard metrics such as accuracy, precision, recall, and F1 score.

Handling of Rare Words RoBERTa's expanded Byte-Pair Encoding (BPE) vocabulary of 50,000 subwords facilitated superior handling of rare and out-of-vocabulary words compared to BERT. This capability ensured more accurate representation of complex linguistic nuances present in customer reviews.

3.4. Multi-label Logistic Regression

3.4.1 Overview

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is typically binary (0 or 1, true or false, positive or negative). In the context of sentiment analysis, logistic regression can be extended to multi-label classification to handle multiple sentiment categories. Despite being a relatively simple model compared to modern transformer-based models, logistic regression remains popular due to its interpretability and efficiency.

3.4.2 Architecture

Multi-label logistic regression applies the principles of logistic regression to multiple labels. It models the probability that a given input belongs to a certain class by fitting a logistic function to the input features. For sentiment analysis, this means predicting whether a review is positive, neutral, or negative based on its textual features.

TF-IDF Representation Logistic regression requires numerical input, so text data must be converted into a numerical format. The TF-IDF (Term Frequency-Inverse Document Frequency) approach is commonly used:

- **Tokenization**: Splitting the text into individual words (tokens).
- **TF-IDF Vectorization**: Creating a vector representation of text where each element corresponds to the TF-IDF score of a word in the text.

Implementation Details For our sentiment classification task, we used a multi-label logistic regression model trained on a TF-IDF representation of the dataset.

3.5. Evaluation Metrics

Evaluation of the model performance was conducted using several key metrics:

- **Accuracy**:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision**:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall**:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1 Score**:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

These metrics were used to evaluate the performance of the model in classifying sentiment categories (positive, neutral, negative) based on customer reviews.

article graphicx float

4. Experiments and Results

The sentiment analysis models were trained and evaluated on a comprehensive dataset to assess their performance using metrics such as accuracy, precision, recall, and F1-score. This section presents the outcomes and comparative analysis of each model.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.6856	0.68	0.69	0.68
BERT	0.7215	0.73	0.72	0.73
VADER	0.4649	0.46	0.46	0.40
RoBERTa	0.7450	0.75	0.74	0.75

Table 1. Comparative results of Logistic Regression, BERT, VADER, and RoBERTa models on sentiment classification.

4.1. Logistic Regression

The logistic regression model achieved an accuracy of 68.56%. It demonstrated moderate performance across sentiment categories, with F1-scores of 0.58 for neutral, 0.71 for negative, and 0.76 for positive sentiments. While providing a solid baseline, its overall performance was surpassed by more advanced models like BERT and RoBERTa.

4.2. BERT

BERT exhibited notable improvements over logistic regression, achieving an accuracy of 72.15%. It maintained balanced precision, recall, and F1-scores around 0.73, indicating robust performance across sentiment categories. BERT's capability to contextualize and interpret reviews resulted in enhanced sentiment classification compared to traditional logistic regression.

4.3. RoBERTa

RoBERTa demonstrated the highest accuracy among the evaluated models, achieving 74.50%. It consistently performed well across sentiment categories with F1-scores of 0.66 for neutral, 0.75 for negative, and 0.82 for positive sentiments. RoBERTa's advanced pre-training and optimization led to a refined understanding of nuanced sentiments, outperforming both BERT and logistic regression models.

4.4. VADER

VADER, a rule-based sentiment analysis method, yielded an accuracy of 46.49%. Despite its simplicity and fast execution, the precision and recall metrics were approximately 0.46, with an F1-score of 0.40. These results indicate that VADER struggled significantly compared to more sophisticated models like BERT and RoBERTa. The lower performance underscores its limitations in capturing the nuanced sentiment expressed in reviews, particularly in contexts where language subtleties and complexities are prevalent. While VADER remains a quick and straightforward option for sentiment analysis tasks, its efficacy may be compromised in scenarios requiring higher accuracy and nuanced understanding of sentiments.

In conclusion, BERT and RoBERTa represent significant advancements in sentiment analysis, offering substantial improvements over traditional logistic regression and rule-based methods like VADER. These findings underscore the importance of model sophistication and data richness in achieving accurate sentiment classification, with ongoing research likely to further refine these capabilities.

5. Conclusion

This study provides a comprehensive evaluation of various sentiment analysis models applied to customer reviews, revealing significant performance differences among them. Transformer-based models, specifically BERT and RoBERTa, exhibit remarkable accuracy and robustness in sentiment classification, significantly outperforming traditional methods like VADER and multi-label logistic regression.

Traditional sentiment analysis methods, such as rule-based approaches and logistic regression, often struggle to capture the nuanced and context-dependent nature of human language. In contrast, transformer-based models leverage deep learning techniques and large-scale pre-training on diverse text corpora, enabling them to understand complex language patterns and subtle sentiment expressions.

Our results demonstrate that BERT and RoBERTa significantly outperform VADER and logistic regression in terms of accuracy, precision, recall, and F1-score. The superior contextual understanding and semantic analysis capabilities of BERT and RoBERTa underscore their effectiveness in handling nuanced sentiment expressions. These findings highlight the potential of advanced NLP models to transform sentiment analysis, offering deeper and more accurate insights into consumer preferences and sentiments. Future work can also explore aspect-based sentiment, n-gram models implemnetations, different pre-processing and the relationship between the length of a review and the results of a model. Future research should explore aspect-based sentiment analysis for more granular insights, integrating sentiment analysis with other data sources, such as social media potentially enhancing the ability of businesses to leverage customer feedback for product development and marketing strategies. .

References

6. Bibliography

6.1. BERT: Bidirectional Encoder Representations from Transformers

- Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Vol. 1, pp. 4171-4186). Association for Computational Linguistics.

- Geetha, M.P., Renuka, D.K. (2021). Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model. *Materials Today: Proceedings*, *45*, 1941-1946. https://doi.org/10.1016/j.matpr.2021.03.644

6.2. RoBERTa: Robustly Optimized BERT Pretraining Approach

- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

- Narayanaswamy, G.R. (2020). Exploiting BERT and RoBERTa to Improve Performance for Aspect Based Sentiment Analysis. *Procedia Computer Science*, *167*, 464-474. https://doi.org/10.1016/j.procs.2020.03.326

6.3. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text

- Hutto, C. J., Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM 2014).*

- Borg, A., Boldt, M. (2020). Using VADER sentiment and SVM for predicting customer response sentiment. *Proceedings of the 31st Irish Conference on Artificial Intelligence and Cognitive Science*. https://arrow.tudublin.ie/scschcomdis/232/