Sentiment Analysis on Amazon Customer Reviews

Joel Alphonso
Computer Engineering
Pune Institute of Computer Technology
Pune, India
joelalph@yahoo.com

Chiranjeev Patil
Computer Engineering
Pune Institute of Computer Technology
Pune, India
chiranjeevpatil0720@gmail.com

Charul Nampalliwar
Computer Engineering
Pune Institute of Computer Technology
Pune, India
charulnampalliwar@gmail.com

***Abstract*—This research paper gives a short introduction to sentiment analysis, its importance, and its machine learning methods. We give a brief summary of the underling mechanics of the BERT and VADER models. Our study looks into the sentiment anaylsis classifier using NLTK's VADER and Huggingface RoBERTa Transformers on amazon customer reviews. We intend to look into the difference between model outputs from the two packages and compare results and assess performance. There are disadvantages and advantages to every model. We will discuss each one and how particularly suited they would be in wide-ranging applications.**

***Index Terms*—Text data, Customer reviews, BERT, VADER, Performance evaluation, Sentiment analysis.**

I. Introduction

Sentiment Analysis has come to be a mainstream area, under the NLP tree, at which subdiscipline, it views the emotions and opinions an individual has in regard to objects expressed in text data. Understanding the sentiments of the users in the content is quite important for a number of applications such as business analytics and social media monitoring in the current digital world, where information flows continuously without any barriers. With the increase in reviews online, forums, and other social media channels, nowadays businesses and researchers employ sentiment analysis techniques in order to find insights about consumer preferences and opinions.

Sentiment analysis techniques heavily rely on tools and models in their functioning. In this regard, our study would be highly interested in two of the most widely-used approaches-VADER (Valence Aware Dictionary and sEntiment Reasoner) and BERT (Bidirectional Encoder Representations from Transformers). Both VADER and BERT are very popular NLP algorithms, which become known for their ability to discover sentiments in a given text. VADER is a combination of lexicon and rule-based modeling, but it remains one of the simplest solutions to derive swift sentiment analysis, while BERT, which is a state-of-the-art deep learning model, claims much more nuanced sentiment patterns.

The paper will be focused on using VADER and BERT for sentiment analysis on Amazon reviews, a very precious resource in the pool of user-generated content, and numerous varieties of sentiments. Amazon is one of the world's largest e-commerce sites, hosting an enormous number of product reviews, making it an ideal environment in which to test techniques for sentiment analysis. In summary, we use these tools to parse the emotions mentioned in Amazon reviews and calculate their relative strengths and weaknesses in terms of complexity captured in consumer emotions as well as real-world performance.

This research paper is structured in the following way: subsequent sections introduce the VADER and BERT models, detail our dataset and methodology used, present the results of a sentiment analysis task on Amazon reviews, and comprise an overall comparison between the approaches undertaken. Ultimately, it aims to determine which of the two methods, VADER or BERT, is most effective for the application of sentiment analysis while giving insights into its potential practical applications in understanding consumer sentiments within the Amazon e-commerce platform.

II. Background

A. Sentiment Analysis

Sentiment analysis or opinion mining is an active area of a study in the field of natural language processing that focuses on opinions, sentiments, appraisals, attitudes of people and emotions through the computational processing of subjectivity in text [4].

Document-level sentiment analysis: A document type that shows the overall sentiment of a whole document, maybe an article or a review or a tweet. Overall, it gives a sense about whether the document is positive or negative, or perhaps neutral.

Sentence-level sentiment analysis: Sentence-level sentiment analysis categorizes individual sentences within a document. This comes in handy especially when you want to catch variations in the sentiment across a text.

Aspect-based sentiment analysis: Aspect-based sentiment analysis is a more fine-grained approach, as it is centered on extracting sentiments of particular aspects, features, or entities within the text. For instance, product attributes on a review can indicate the sentiment that exists towards various attributes of a product. [11].

B. Lexicon Approach

Sentiment analysis using a lexicon-based approach is the use of predefined lists, mostly called opinion lexicons, in classifying the opinions that a piece of text conveys. The reason behind this method is the fact that positive words are represented by positive sentiments and that negative words are represented by negative sentiments. Generally, lexicon-based methods focus on finding such opinion words and applying them in an analysis of the text. The lexicon-based approach has mainly two procedures: the corpus-based and the dictionary-based approaches.

Corpus-Based Approach: The corpus-based approach begins with a seed list of opinion words, usually a small set of known words endowed with clear sentiment orientations. The intention is to inflate the list with words from a large corpus of text that share the same sentiment orientations. Typically, such an approach finds more words carrying sentiment within the corpus relying upon the grammatical patterns or cooccurrence of words with seed opinion words. For example, if ”excellent” is in your seed list, then the strategy would be something like this: In that case, it tries to identify words which frequently co-occur with ”excellent” in some specific corpus in order to detect more positive sentiment words. Such words, in the corpus-based approach, may either exploit statistical methods (statistical approach) or semantic relationships between words (semantic approach) in determining the technique to be used.

- Statistical approach: It is used in many applications that have a relation in the field of SA. The most famous of them is the one that can detect the manipulation of the review by conducting a statistical test of randomization which is called runs. Test.
- Semantic approach: It gives values to sentiments while relying on more than a principle to calculate the affinity and similarity of different words. The basis of this principle is to support the Sentiment value in the words and words close.

Dictionary-Based Approach:: Dictionary-based method proposed an all-inclusive strategy for dictionary-based method. In this popular strategy, a small group of words is handpicked with known trends. Then we come to plant this set of words by searching for all synonyms and antonyms in the known approach corpora, thesaurus or WorldNet. The new words found are added to the seed list, and the following repetition then starts. This continues until there is no further repetition of new words [2].

C. Machine Learning Approach

In the domain of opinion mining, two major approaches come forward. The lexicon-based approach relies on hand-curated sentiment lexicons to extract sentiments from text. This process, however, is time consuming because compiling rich and accurate lexicons is tedious. On this account, automatic methods have emerged, which is based on machine learning techniques. These techniques enable the system to decipher the sentiment-rich features in text and thus bring a less importance to the handcrafting of the lexicon. For such a reason, modern opinion mining incorporates these machine learning approaches for effortless identification of patterns in sentiment. [2] [4].

Naive Bayes classifier is the smallest classifier which is based on Bayesian probability and the naive assumption that feature probabilities are independent of each other. [4].

Support Vector Machines: SVM’s are non-probability classifiers that work based on the principle of separating data points in space using one or more hyperplanes. [4].

Neural Networks: Neural network is a continuum of algorithms based on the recognition of the relationships inherent in several sets of data through a process similar to that which the human mind does.[2].

D. Tokenization:

SentencePiece Tokenization: SentencePiece is an unsupervised text tokenization and detokenization tool mainly for Neural Network-based text generation tasks. It implements subword units (like byte-pair-encoding (BPE) and unigram language model with the extension of direct training from raw sentences. With SentencePiece, we can make a purely end-to-end system that does not depend on language-specific pre/postprocessing. SentencePiece implements subword units, such as byte-pair-encoding (BPE) and unigram language model with the extension of direct training from raw sentences. SentencePiece enables us to build a purely end-to-end system that does not rely on language-specific pre/postprocessing. SentencePiece is actually generic and is thus optimized for use in Neural Network-based text generation tasks. A subword unit implementation (for example, byte-pair-encoding and unigram language model with the extension of direct training from raw sentences). It allows us to build an end-to-end system purely - a system that does not depend on language-specific pre/postprocessing. [12].

E. BERT

BERT is abbreviation for the language representation in the form of bidirectional encoder representation of the Transformer. It is meant to be jointly pre-trained for deep bidirectional representations from unlabeled text on all its layers for training both left and right contexts. [3].

We train a deep bidirectional representation simply by predicting the tokens that have been masked after randomly masking some percentage of the input tokens. Though we refer to this procedure more often in the literature as a Cloze task, we will actually call it a ”masked LM” (MLM). As in a standard LM, the last hidden vectors in this case represent the mask tokens and pass into an output softmax over the vocabulary. For each of our experiments, we randomly mask 15 percent of all WordPiece tokens in each sequence. Unlike denoising auto-encoders Vincent et al. 2008, we do not reconstruct the entire input; rather, we predict only the words that are masked.

Understanding the relationship between two sentences is the backbone of many other important downstream tasks, such as Question Answering (QA) and Natural Language Inference (NLI), that are not directly represented by language modeling. We pretrain a model on a binarized next sentence prediction task easily derivable from any monolingual corpus for training a sentence relationship understanding model. That is, for each pre-training example, half the pair of sentences A and B are sampled randomly from the corpus (labeled as NotNext), and half are realizations of the sentence that follows A (labeled as IsNext).

The Robustly Optimized BERT Pretraining Approach, briefly known as RoBERTa, is an updated version of the BERT model. We refer to our new training recipe for a BERT model as RoBERTa, and it can match or outperform every post-BERT approach since BERT was woefully undertrained. Simple revisions included training the model on longer sequences with bigger batches of data, and for much more extended periods of time; removing the requirement for the next sentence prediction objective; instead, dynamically change the masking pattern on the training set [5].

F. VADER

VADER, or Valence Aware Dictionary and Sentiment Reasoner, is a Lexicon and rule-based sentiment analysis tool designed to analyze the sentiment expressed in text data. VADER is a rule-based sentiment analysis tool that draws on lexicon to express sentiment in social media. To further add to its efficiencies in sentiment analysis, VADER features a well-crafted sentiment lexicon as well as a syntactic rule set. VADER supports emoticons and acronyms that are unique to Twitter. Emotional symbols known as emojis are widely employed on the Internet [8]. For social media, the VADER lexicon is quite excellent. The correlation coefficient reveals that VADER (r = 0.881) describes ground truth as well as an individual human rater with a correlation coefficient of r = 0.888 for each tweet's sentiment intensity aggregated across the mean from 20 human raters. Interestingly, if we were to examine the accuracy of this categorization closely, we then observe that in fact VADER actually classifies the sentiment of the tweet into positive, neutral, and negative classes significantly better than individual human raters (F1 = 0.96 vs. F1 = 0.84). [4].

G. The GloVe Model

All unsupervised methods for learning word representations rely on the statistics of word occurrences in a corpus as their main source of information. While many such methods are currently in use, it remains unclear how meaning is derived from such statistics and how the resulting word vectors might represent that meaning. In this section, we clarify somewhat this question. We draw upon our intuition to define a new word representation model that we refer to as GloVe, or Global Vectors, since the model directly captures the global statistics of the corpus. [6].

III. Methodology

A. Proposed Method

1) Data: Amazon reviews have been employed in this research. The dataset has been acquired from Kaggle. From the whole dataset only 10,000 subsets of Amazon review have been chosen for analysis. It is a review data set containing reviews along with textual content and star ratings. The textual data is very informative, directly created by users. These reviews regarding other products were different in terms of expression and emotions, making it perfect for sentiment analysis.

2) VADER Sentiment Scoring: To compute the sentiments for the reviews on Amazon, I applied the VADER sentiment scoring tool. For instance, I utilized the following from the Natural Language Toolkit : this tool incorporates a pre-trained model namedSentimentIntensityAnalyzer, which determines the polarity scores for each review in the dataset to measure sentiment in text data. These polarity scores, including positive, negative, neutral, and a compound score, provide insight to the emotions reflected within reviews. Polarity scores were calculated on the basis of the entire dataset, giving a basis for the following comparative analysis.

3) RoBERTa Sentiment Scoring: Apart from VADER, the RoBERTa model was also used for sentiment analysis. The ”cardiffnlp/twitter-roberta-base-sentiment” is used as the model for this paper. Its corresponding tokenizer was used as such for the sequence classification task in classifying the Amazon reviews. This particular model is efficient for use in various pre-trained capabilities to have an in-depth capture of a fine sentiment expression in text. It gives a prediction for the negatives, neutrals, and positives of sentiments for each review for deep analysis of the sentiment in the dataset.

4) Combined Analysis: To understand sentiment deep in the reviews available on Amazon, a combined analysis was carried out that included merging VADER and RoBERTa outputs obtained through sentiment analyses. This approach provided for one interesting comparative analysis of performance by VADER and RoBERTa in sentiment analysis. Percentages were then computed to determine the distribution of the different categories of sentiment. Data visualizations created included bar plots, violin plots, and finally, box plots that were used to visualize and interpret their findings. In these visualizations, they could explore the differences between the two approaches on the appearance of the reviews by star ratings.

5) Data Visualization: There is a set of data visualizations which represents the result of sentiment analysis: after all, the distribution of sentiment scores in combination with Amazon star reviews has been represented using bar plots. Violin plots will be drawn to present the distribution of both VADER and RoBERTa's sentiment scores for the sake of contrast in the results. In addition to this, box plots are used to get an idea about the spread of sentiment scores, and the pair plots visually describe relationships and correlations between different sentiment scores.

6) Analysis and Comparison: The results of the VADER and RoBERTa models for the sentiment analysis were compared in a more comprehensive manner. Through the analysis of the sentiment scores and percentages, an overall assessment of both methods regarding performance in sentiment analysis was done. This analysis sought to discern any forms of difference, pattern, or trend in the expression of sentiment in this dataset that would both showcase each of the methods' strengths and weaknesses with respect to the nuanced capturing of sentiment in Amazon reviews. To further unfold the empirical findings from this analysis, their relative performance can be grasped comprehension through the succeeding pages of this research paper.

B. Tools

NLTK: NLTK stands for Natural Language Toolkit. It is a powerful library written in Python, tailored to natural language processing and text analysis. For the purpose of this research work, I used NLTK with the VADER sentiment tool to perform sentiment analysis. In fact, all the polarity scores, including positive, negative, neutral, and compound scores, for each Amazon review in the data set were generated through NLTK's SentimentIntensityAnalyzer.

Transformers library: The Transformers library is amongst the best libraries for NLP and deep learning. It was particularly important to extend the application of the RoBERTa model for sentiment analysis. The model ”cardiffnlp/twitter-roberta-base-sentiment” and its tokenizer were applied for sequence classification. This pre-trained model allowed the evaluation of the sentiment, which gave predictions regarding whether the reviewed products on Amazon were negative, neutral, or positive.

Torch(PyTorch): PyTorch, or Torch in some common tongues, was just a deep learning framework that happened to be what one needed to work with RoBERTa and manage deep learning models. It would provide the structure that is necessary to run the RoBERTa model and make predictions for sentiment classification

IV. Results and Discussion

In this section, we elaborate on the findings of the same sentiment analysis on the same data for both VADER and RoBERTa, in pursuit of identifying sentiment patterns, comparing the performance of tools used, and, above all, drawing useful insights from our research.

TABLE I
VADER Results

Score	Positive (%)	Neutral (%)	Negative (%)
1	9.87	80.59	9.54
2	7.57	80.21	12.22
3	5.48	80.79	13.74
4	3.73	77.27	19.00
5	3.01	74.71	22.28

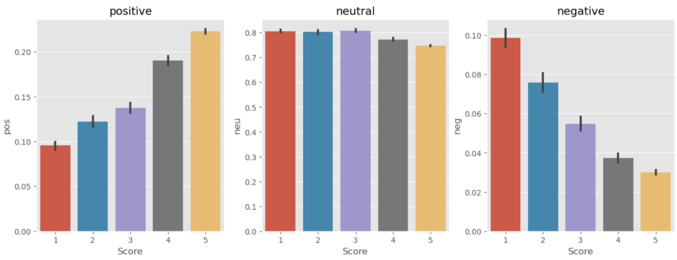

Fig. 1. Distribution of Sentiment Scores by Rating.

From Fig. 1. and Table I VADER's sentiment scoring was more conservative, especially in rating the positive reviews. VADER appears to be giving lower scores to extremely positive reviews as well, which indicates that the tool is relatively stingy in terms of rating positive sentiment. This can be directly attributed to the inherent nature of VADER's sentiment analysis, which does tend to find itself taking a relatively middle-of-the-road approach to rating categories for all intents and purposes. While a predominance of negative sentiment prevails in lowest ratings for VADER's sentiment analysis,.

TABLE II
RoBERTa Sentiment Analysis Results

Score	Positive (%)	Neutral (%)	Negative (%)
1	10.95	18.55	70.50
2	20.74	26.89	52.38
3	42.19	28.59	29.22
4	76.18	16.31	7.51
5	88.63	8.21	3.17

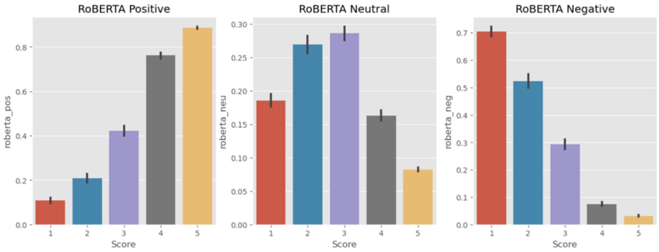

Fig. 2. Distribution of Sentiment Scores by Rating.

From Fig. 2. and Table II 0As review scores increase from the lowest to the highest, there is an observable, increasing trend in the proportion of positive sentiment expressed while the percentage of negative sentiment significantly decreases. This means that higher-rated reviews are much more likely to express positive sentiment than negative sentiment, whereas lower-rated reviews tend to have a greater proportion of negative sentiment. Specifically, the neutral class is represented equally across all categories of ratings, thus implying that there are independent or balanced opinions in the corpus. The diversities of the sentiment scores from the RoBERTa model are incredible, showing that it captures subtle shades of expression of feeling very well. That makes RoBERTa a more significant model for sentiment analysis, when the nuance, in the expression of the sentiment, has to be captured more fine-grainedly.

V. Conclusion

Thus, we conclude, from our analysis of their sentiment with VADER and RoBERTa, that there are particular patterns and performance differences between them. From results presented in Table I and Figure 1, the outcome is that VADER used a relatively conservative approach in its sentiment scoring. It gave even the most positive reviews lower scores in terms of sentiment; thus, it actually appears fairly middle-of-the-road by rating categories on average. Whereas, negative emotion tendency was higher in the lowest ratings according to the analysis by VADER.

On the other hand, results of sentiment analysis of RoBERTa, as shown in Table II and Figure 2 indicates an evident upward curve in positive sentiment as the rating of the reviews move from the lowest rating to the highest rating. Thereby, negative sentiment is relatively lower when the ratings are higher where the overall tendency is towards positive sentiment when the ratings were high. The prominent persistency of neutral sentiment across all rating categories indicates the presence of objective or impartial reviews in the dataset. Extremely high variability in RoBERTa's sentiment scores is indicative of its rich ability to capture more complex nuances related to the expression of sentiment, which makes the tool effective in sentiment analysis when a finer degree of understanding of sentiment nuances is at stake.

As can be seen from the above analysis, there are contrasting approaches of VADER and RoBERTa in sentiment analysis, while conservatism is the tendency of VADER, RoBERTa excels in nuanced sentiment capture. Hence, the choice will depend on the analysis-specific needs and the desired level of granularity in the interpretation of sentiment.

References

[1] W. Medhaat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineering Journal, vol. 5, pp. 1093-1113, 2014.

[2] A. A. Q. Aqlan, B. Manjula, and R. L. Naik, "A Study of Sentiment Analysis: Concepts, Techniques, and Challenges," in Proceedings of International Conference on Computational Intelligence and Data Engineering, Lecture Notes on Data Engineering and Communications Technologies 28, 2019, pp. 16.

[3] J. Devlin, et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1810.04805

[4] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of social media Text," Georgia Institute of Technology, Atlanta, GA, 2014. [Online]. Available: http://eegilbert.org/papers/icwsm14.vader.hutto.pdf

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, 2019.

[6] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, pp. 1532-1543.

[7] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," Neural Information Processing Systems, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1706.03762

[8] M. K. Bhagya Laxmi, B. Yamini, C. Rakshitha, and D. Keerthi, "Twitter Sentiment Analysis Using VADER on Python," International Journal of Emerging Technologies and Innovative Research, vol. 7, no. 5, pp. 1025-1031, May 2020. [Online]. Available: http://www.jetir.org/papers/JETIR2005456.pdf

[9] S. Elbagir and J. Yang, "Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment," in Proceedings of the International MultiConference of Engineers and Computer Scientists 2019 (IMECS 2019), Hong Kong, March 13-15, 2019, pp. 12-16.

[10] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, et al., "Transformers: State-of-the-Art Natural Language Processing," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38-45, 2020. http://dx.doi.org/10.18653/v1/2020.emnlp-demos.6

[11] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," Social Network Analysis and Mining, vol. 11, no. 81, pp. 1-2, 2021. https://doi.org/10.1007/s13278-021-00776-6

[12] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing," in Conference on Empirical Methods in Natural Language Processing, 2018.