Sentiment Analysis of an E-commerce Clothing Review Dataset: Transformers roBERTa vs VADER Model Comparison

Jamian Ikel Ross M. Huang
Department of Computer, Information Sciences and Mathematics
University of San Carlos
Cebu City, Philippines
19100253@usc.edu.ph

***Abstract*—Sentiment Analysis, is the process of analyzing text to determine and classify the emotional tone of said text, most commonly being split into "positive", "negative", and "neutral" tones. In this term paper, we perform sentiment analysis on a public dataset about clothing products reviews and use two different models, the VADER Model and the roBERTa model. We then compare the results we get from applying these models onto said dataset. In doing so, we're able to get more insight as to which dataset gives a more accurate result onto the dataset or datasets similar in nature to the one used.**

***Index Terms*—Sentiment Analysis, Dataset, Model**

I. INTRODUCTION

Natural language processing (NLP) is a branch of artificial intelligence (AI) that enables computers to comprehend, generate, and manipulate human language. Natural language processing has the ability to interrogate the data with natural language text or voice [1]. The type of study being done, called "Sentiment Analysis" belongs to the field of NLP, under the category of text classification.

Text classification is a machine learning technique that assigns a set of predefined categories to open-ended text. Text classifiers can be used to organize, structure, and categorize pretty much any kind of text – from documents, medical studies and files, and all over the web [2]. Similar to how some texts can be classified as "spam" mail or regular mail, Sentiment Analysis works similarly in that in can identify whether certain texts are of a *positive, negative, or neutral* tone, and these results are often summarized and used for analysis by businesses regarding product reviews, customer feedback, ratings, and many more.

Sentiment analysis is the process of analyzing digital text to determine if the emotional tone of the message is positive, negative, or neutral [3]. A crucial tool in NLP are called "Models". In this context, Models are trained to better understand and manipulate the human language, to extract meaning from them.

A machine learning model is a program that can find patterns or make decisions from a previously unseen dataset. For example, in natural language processing, machine learning models can parse and correctly recognize the intent behind previously unheard sentences or combinations of words [4].

For our use, the Models being used are *pre-trained* and are being used to classify and review the different emotional tones commonly found in a shop's review section. The models in question are the VADER (Valence Aware Dictionary for Sentiment Reasoning) Model, a rule-based/lexicon sentiment analyzer, and the roBERTa Model from huggingface, which is inspired by the BERT Model, a deep learning model in which every output element is connected to every input element, and the weightings between them are dynamically calculated based upon their connection [5].

The objectives for this paper are as follows:

1) Perform Sentiment Analysis on the E-commerce Clothing Review Dataset using the two Models.
2) Use Direct Comparison, Accuracy, Precision, Recall, Specificity, F1 Score, and Irony/Sarcasm to test the results from both Models.
3) Conclude the results for each model.

These objectives will allow the paper to present a clearer understanding on how each model works, particularly how well they perform on a feedback-review based dataset on products such as clothing; when to use each model and what other considerations should be made when using these models, when used against datasets of similar nature to the one being analyzed.

II. REVIEW OF RELATED LITERATURE

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

In the research paper "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text" [6], VADER is described as a simple, rule-based model for general sentiment analysis.

VADER's creation and methods were as followed: "to leverage the advantages of parsimonious rule-based modeling to construct a computational sentiment analysis engine that 1) works well on social media style text, yet readily generalizes to multiple domains, 2) requires no training data, but is constructed from a generalizable, valence-based, human-curated gold standard sentiment lexicon 3) is fast enough to be used online with streaming data, and 4) does not severely suffer from a speed-performance tradeoff.", focusing efforts in the development and validation of a gold standard sentiment lexicon sensitive to sentiment polarity / intensity found in social media microblog, identifying and evaluating generalized rules on conventional uses of grammatical / syntactical aspects of text to assess sentiment intensity, and finally, comparing the performance of the VADER model compared to the other established sentiment analysis baselines.

The model's results were then evaluated by getting the mean sentiment rating from 20 pre-screened and appropriately trained human raters, and compared the F1 score, precision, and recall scores compared to other sentiment analysis lexicons such as SCN, GI, SWN, LIWC, ANEW, WSD, and Hu-Liu04, across four different data sets such as Social Media Tweets, Movie Reviews, Amazon Product Reviews, and NY Times Editorials.

The results showed that VADER was able to outperform human raters, having a (F1 = 0.96) score whereas the human raters only had a (F1 = 0.84) score when it came to correctly classifying tweet sentiments. The final conclusion was that "VADER performed as well as (and in most cases, better than) eleven other highly regarded sentiment analysis tools.".

B. RoBERTa: A Robustly Optimized BERT Pretraining Approach

In the research paper "RoBERTa: A Robustly Optimized BERT Pretraining Approach" [7], the researchers found that BERT was significantly undertrained, and proposed an improved method in training BERT models, dubbed RoBERTa, that can match / exceed the performance of post-BERT methods. The research's modifications included: " (1) training the model longer, with bigger batches, over more data; (2) removing the next sentence prediction objective; (3) training on longer sequences; and (4) dynamically changing the masking pattern applied to the training data. We also collect a large new dataset (CC-NEWS) of comparable size to other privately used datasets, to better control for training set size effects.".

The conclusion for the creation of RoBERTa was that the model could achieve significant improvements by allowing the model to train longer with bigger batches over more data; removing the next sentence sequence prediction objective; training on longer sequences; and by changing the masking pattern applied to the training data.

RoBERTa was able to achieve "state-of-the-art" results via GLUE, RACE, and SQuAD.

Ultimately, the suggestion was that "BERT's pretraining objective remains competitive with recently proposed alternatives.".

III. METHODOLOGY

Dataset

The dataset used is entitled "Women's E-Commerce Clothing Reviews". Columns used include *Rating* and *Review Text*. The dataset is quite straightforward as it is review-based from an online store, unlike if it was twitter / reddit posts or online articles where there are other factors to take into account, such as sarcasm, spite-reviews, nuance, inside jokes, obscure references, and among many others.

Sentiment Analysis and Tools

Tools used include Python, Jupyter Notebook, NTLK package, VADER Model, and cardiffnlp/twitter-roberta-base-sentiment. The sentiment analysis performed is limited to the classifications of the *positive, neutral,* and *negative* emotional tones, and does not go further than that (such as including emotional intensity, etc.)

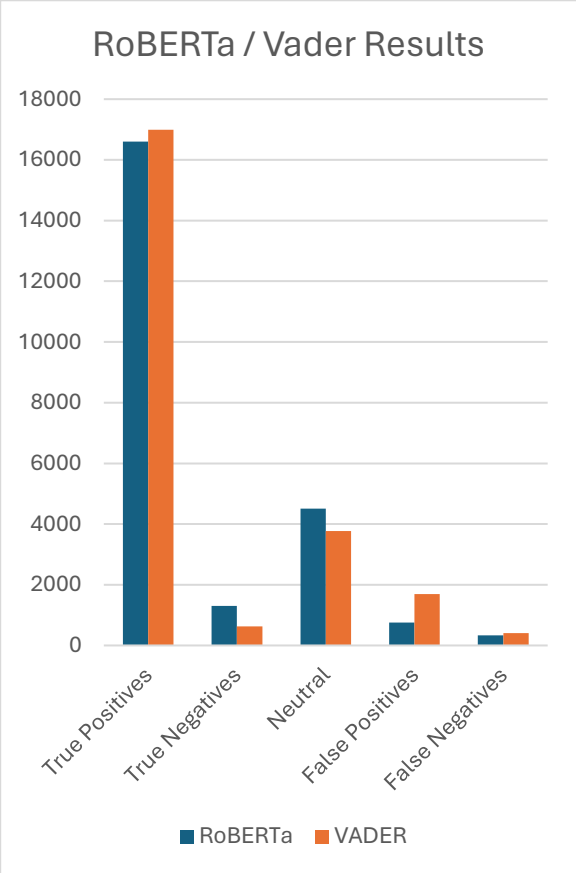
Testing

Testing will include Direct Comparison of results, Accuracy, Precision, Recall, Specificity, F1-Score, and testing how the models respond to *Sarcasm / Irony*. The testing all place emphasis on the accuracy of the Models' classifications after performing sentiment analysis.

IV. RESULTS AND CONCLUSIONS

Final Results

	RoBERTa	*VADER*
True Positives	16597	16989
True Negatives	1301	631
Neutral	4505	3768
False Positives	753	1696
False Negatives	330	402



Looking at the results directly, we see that in terms of True Positives and Neutral Identification, that both are nearly identical in numbers, however when we look at the quantity for True Negatives and False Positives, we see that RoBERTa has a higher rate of correct identification whereas VADER has nearly double the number of False Positives, indicating wrong predictions, at a much higher rate than RoBERTa.

Accuracy

Vader Accuracy: 89.35997565676031 %

roBERTa Accuracy: 94.29429429429429 %

The accuracy for both are quite high, and between the two are very close. The straightforwardness of product reviews from paying customers might contribute to the high-accuracy count from the sentiment analysis, as there are less occurrences of sarcasm and spite-reviewing, compared to game ratings and reviews.

Precision

Vader Precision: 0.9092320042815092

roBERTa Precision: 0.9565994236311239

Recall

Vader Recall: 0.976884595480421

roBERTa Recall: 0.9805045194068648

Specificity

Vader Specificity: 0.2711645896003438

roBERTa Specificity: 0.633398247322298

Here we see a considerable difference, where RoBERTa's specificity score is nearly triple that of VADER's. If we look at the graph from *Final Results,* we see that the number of true negatives from RoBERTa is double that of VADER, implying a higher rate of correct identification.

F1-Score

Vader F1-Score: 0.9418449939017629

roBERTa F1-Score: 0.9684044694693236

Interpreting the results, we see that the F1-Score of both VADER and RoBERTa are both > 0.9, indicating "excellence".

Sarcasm / Irony

Sentence used is **The quality of the product was so high that it broke after 1 use! Definitely recommend!**

	RoBERTa	VADER
Positive	0.9719866	0.281
Neutral	0.023617167	0.583
Negative	0.004396171	0.136

From here we can infer that the models are not equipped to deal with sarcastic comments, and correctly identifies them as "Positive" with the high positive and neutral scores and low negative scores.

Conclusions

Both models performed excellently, with both of them having an F1-score > 0.9. The RoBERTa Model performs significantly better in some areas, with identifying True Negatives and avoiding False Positives nearly double in effectiveness compared to VADER.

For Datasets with expected straightforward responses like Clothing Store Reviews or anything similar, especially for small scale datasets, VADER and RoBERTa proved effective.

However, with bigger datasets like Amazon's datasets, the probability of encountering spite-reviews, nuanced sentences, obscure references, and sarcasm becomes higher, especially on Datasets like forums and post-based websites like Facebook. It is recommended that on Datasets like that, you opt for RoBERTa Models that has also been trained to detect sarcasm. For most cases and applications RoBERTa outperforms VADER. For smaller scale /simpler datasets VADER can also perform well, but generally you'd want to use RoBERTa, for in this case Sentiment Analysis.

REFERENCES

1. Oracle, “What Is Natural Language Processing (NLP)?”: https://www.oracle.com/ph/artificial-intelligence/what-is-natural-language-processing/#:~:text=Natural%20language%20processing%20(NLP)%20is,natural%20language%20text%20or%20voice
2. MonkeyLearn, “Text Classification: What it is And Why it Matters”: https://monkeylearn.com/text-classification/
3. AWS, “What is Sentiment Analysis?”: https://aws.amazon.com/what-is/sentiment-analysis/#:~:text=Sentiment%20analysis%20is%20the%20process,social%20media%20comments%2C%20and%20reviews
4. DataBricks, “Machine Learning Models”: https://www.databricks.com/glossary/machine-learning-models#:~:text=A%20machine%20learning%20model%20is,sentences%20or%20combinations%20of%20words.
5. C. Hashemi-Pour, and B. Lutkevich, “BERT language model”: https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model#:~:text=BERT%2C%20which%20stands%20for%20Bidirectional,calculated%20based%20upon%20their%20connection.
6. Hutto, C.J. & Gilbert, E.E., “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text”, Jan. 2015: https://www.researchgate.net/publication/275828927_VADER_A_Parsimonious_Rule-based_Model_for_Sentiment_Analysis_of_Social_Media_Text
7. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, “RoBERTa: A Robustly Optimized BERT Pretraining Approach”, 2019: https://arxiv.org/pdf/1907.11692.pdf