Aspect Based Sentiment Analysis using VADER and RoBERTa

1st Sazid Hasan Tonmoy
dept. of Computer Science & Engineering(CSE)
School of Data & Sciences(SDS)
Brac University
Dhaka,Bangladesh
sazid.hasan.tonmoy@g.bracu.ac.bd

2nd Abdulla Al-Amin
dept. of Computer Science & Engineering(CSE)
School of Data & Sciences(SDS)
Brac University
Dhaka,Bangladesh
abdullah.al.amin.saikot@g.bracu.ac.bd

3rd Tasnim Sultana
dept. of Computer Science & Engineering(CSE)
School of Data & Sciences(SDS)
Brac University
Dhaka,Bangladesh
tasnim.sultana@g.bracu.ac.bd

4th Md Sabbir Hossain
dept. of Computer Science & Engineering(CSE)
School of Data & Sciences(SDS)
Brac University
Dhaka,Bangladesh
md.sabbir.hossain1@g.bracu.ac.bd

5th MD. Mustakin Alam
dept. of Computer Science & Engineering(CSE)
School of Data & Sciences(SDS)
Brac University
Dhaka,Bangladesh
md.mustakin.alam@g.bracu.ac.bd

6th Md Humaion Kabir Mehedi
dept. of Computer Science & Engineering(CSE)
School of Data & Sciences(SDS)
Brac University
Dhaka,Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

7th Annajiat Alim Rasel
dept. of Computer Science & Engineering(CSE)
School of Data & Sciences(SDS)
Brac University
Dhaka,Bangladesh
annajiat@gmail.com

***Abstract*—In this Digital age, there's a growing need to understand the emotions and intent behind each and every word. It is even more true for business companies that rise and fall based on their consumer reviews. Customers also decide on their preferred services based on these reviews. With that in mind, in this study, we take a look at a data-set which consists of hotel reviews and try to find out the negative and positive sentiment by analyzing the reviews based on aspects. We used VADER and RoBERTa models for our desired results. This paper aims to help companies and customers by presenting a technique to distinguish positive and negative reviews from users.**

Index Terms*—Aspect-based sentiment analysis (ABSA), RoBERTa, VADER, Natural Language Processing(NLP), Hotel Reviews

I. Introduction

People's reliance on online reviews is increasing with each passing year. They rely on these reviews to choose their restaurant, lodging, and mode of transportation. These reviews are becoming invaluable as times go on and companies are forced to focus on them. There have been many pieces of research over the years on distinguishing positive reviews from negative reviews [9]. It gives the company a chance to take a look at what they need to improve on and what they are doing right. Sentiment analysis is a branch of Natural Language Processing(NLP) that is also referred to as opinion mining [10]. It works to identify the sentiment of a text and classifies it into negative, positive, and neutral. It is challenging to extract sentiment scores from a text and potentially label it as positive or negative [11]. Traditional sentiment analysis focuses on classifying the overall attitude that is expressed in a text without distinguishing the subject of the attitude. If the text is referring to numerous subjects or things at once (also known as aspects), maybe communicating different thoughts about distinct aspects, this might not be sufficient. A more challenging method called aspect-based sentiment analysis entails identifying sentiments related to specific textual features (ABSA). In this study, we intend to assess sentiment scores and extract aspect phrases from hotel reviews using the VADER and RoBERTa model. We will ultimately determine if the reviews are good or negative after examining the sentiment scores of all the aspect phrases.

II. Related Work

In the framework of aspects-based sentiment analysis, this section tries to critically analyze prior pertinent research in the field of opinion mining. We looked at the many methods and strategies applied to get the desired result in NLP related tasks. Opinion mining encompasses computational linguistics, natural language processing (NLP), and data mining. In contrast to its counterpart, ordinary syntactical NLP, opinion mining doesn't require a thorough grasp of the text. While Syntactical NLP works on summarization and auto classification, it primarily focuses on semantic inference and affective information related to natural language.

Beyond the straightforward distinction between positive and negative, this paper established an additional classification method for evaluative language using a selection of linguistic variables, and a categorization framework that identifies several forms of evaluation in a particular text. The principal contributions of this study are the expansion of the simple positive and negative classification scheme for evaluative language, the linguistic analysis of online reviews, and experimental evidence of the classification system's effectiveness. [1]. Some researches focuses on integral theme and implicit and explicit opinions. The researchers who conducted this research aim to give readers an intuitive understanding of the various ways authors might convey emotion in their writing as well as the information that can be useful in doing so. Key themes covered include opinion models, opinion objectives, and opinion expressions and to remove uncertainty and facilitate a deeper discussion on important themes, They also provide their own definitions, examples, and explanations [2]. There has been lot of progress made in NLP field one of them being emergence of transformer based models. In this paper the researchers propose a fresh model for language representation called Bidirectional Encoder Representations from Transformers (BERT). BERT utilizes cutting-edge techniques to deliver competitive results. The ideal method concatenates the token representations from the top four hidden layers of the pre-trained Transformer, which is only 0.3 F1 slower than completing the full model. This illustrates that BERT works well for both feature-based and fine tuning techniques [3]. While BERT is one of the best task specific model, it also has its own shortcoming. In this paper, the researcher improved ABSA's performance with the help of an additional training on Review text, called Post-Training (BERT-PT) which demonstrates the advantages of having two forms of knowledge by having the best performance across all tasks in all areas. They discovered that task-awareness (MRC) post-training accounts for the majority of the performance gain of BERT-PT (as indicated by BERT-MRC). Domain knowledge post-training also provides the majority of the performance improvement, indicating the importance of contextualized representations of domain information [4]. Outside of BERT LSTM is another major model used in sentiment analysis. In order to get over recurrent neural networks' limitations when processing naturally organized data, the researcher built a machine reading simulator with a memory network that directly stores input token context without recursively compressing it. It is based on a long short-term memory architecture. [5]. AAnother paper converts (T)ABSA from a single sentence classification project to a sentence pair classification task by using a pre-trained BERT model and an auxiliary sentence. On the craft of classifying sentence pairs, they refined the pre-trained BERT model and got the most recent cutting-edge results. The researchers evaluated the advantages of sentence pair classification, contrasted the experimental outcomes of single sentence classification with sentence pair classification based on BERT fine-tuning, and validated the efficacy of their conversion strategy. [6]. Due to the substantial amount of text that it has been trained on, BERT has demonstrated the ability to perform well on NLP tasks. [7].

III. Working with Dataset

We collected Hotel Reviews from the Trip Advisor website from the internet, which consisted of 20491 reviews. Among them we had 9054 - 5 star reviews, 6039 - 4 star reviews, 2184 - 3 star reviews, 1793 - 2 star reviews and 1421 - 1 star reviews, shown below in Fig. 1.

Fig. 1. Number of reviews according to rating

We then checked if the data-set had any null value and the percentage of it, which was zero. For our purpose we used 1000 reviews from the data-set, which consisted of 437 - 5 star reviews, 315 - 4 star reviews, 112 - 3 star reviews, 82 - 2 star reviews, 52 - 1 star reviews, shown below in Fig. 2.

Fig. 2. Number of reviews according to rating from 1000 reviews

We added an extra 'Id' column in our dataset for further objectives. For cleaning purposes, we changed all the reviews into lowercase and replaced anything that is not an alphabet or a number with an empty string using Regular Expressions.

IV. Methodology

Our objective is to extract the aspects from each review and analyze the sentiments of those aspects and ultimately categorize those reviews as either positive or negative. For our purposes, we've used two models - 1. VADER, 2. The RoBERTa. In this section we will describe each of the models and their architecture.

2.1. VADER - VADER (Valence Aware Dictionary for Sentiment Reasoning) is a model which is utilized to recognize the sentiment of a text. The model utilizes a dictionary to label sentiment scores. It examines a word and obtains a sentiment score; through that, it identifies whether a text is positive, negative, or neutral. For our purposes, we only considered positive and negative scores for each aspect of a review.

2.2. The RoBERTa - RoBERTa(Robustly Optimized BERT-Pre Training Approach) is similar to BERT with some transformations to give better performances and results. The model uses dynamic masking, unlike BERT which uses Static masking. The RoBERTa model and the BERT model have the same architecture. The first fully unsupervised, bidirectional language representation model is called BERT. There were several further bidirectional unsupervised learning-based language models available prior to BERT.

In contrast to recent language representation models, BERT aims to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. Thus, utilizing the pre-trained BERT model with just one additional output layer and no significant task-specific structure modifications, state-of-the-art models for a number of tasks, including question answering and language inference, may be produced. Except for the outer layers, pre-training and fine-tuning in BERT employ identical architectures to RoBERTa. Models are initialized for various downstream tasks using the same pre-trained model parameters. All parameters are adjusted during fine-tuning. The RoBERTa is trained with bigger mini-batches and does not use the next-sentence pre-training target. Instead, RoBERTa uses a different pre-training approach and replaces the byte-level BPE tokenizer with a character-level BPE vocabulary. In this model, we also do not need to establish which token belongs to which segment.

V. Experimental Result

From our data-set we extracted Compound Nouns and Aspect keywords. We considered each line of each review and replaced words like - why, here, now, so, be etc. with a space. We used spaCy to find the parts of speech of each word and appended them to Compound Nouns and Aspect keywords according to their merits. Aspects were also extracted and added to our data-set using the same technique. The VADER model was imported for sentiment analysis of our data-set. We ran the model on our review column first. In the figure, the Positive score for rating 5 is high and is low for rating 1 as expected, and for the negative score rating 1 is highest and rating 5 is the lowest. But in Neutral all bars are almost at equal length.

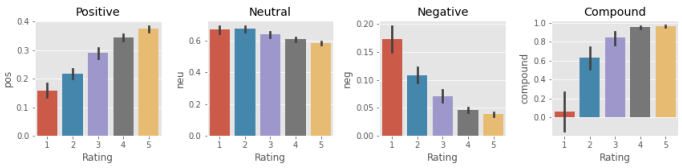

Fig. 3. VADER sentiment result

After that ROBERTA model was imported. We calculated the sentiment of the data-set by analysing the reviews. In this figure Positive score for rating 5 is much higher than rating 1, and Negative score for rating 1 is much higher than rating 5.

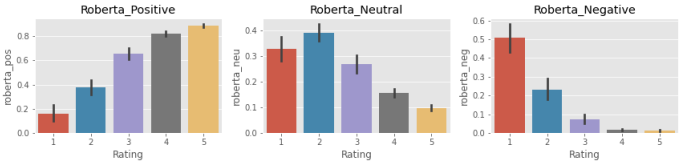

Fig. 4. RoBERTa sentiment result

We compared the results we got from our VADER model and ROBERTA model.

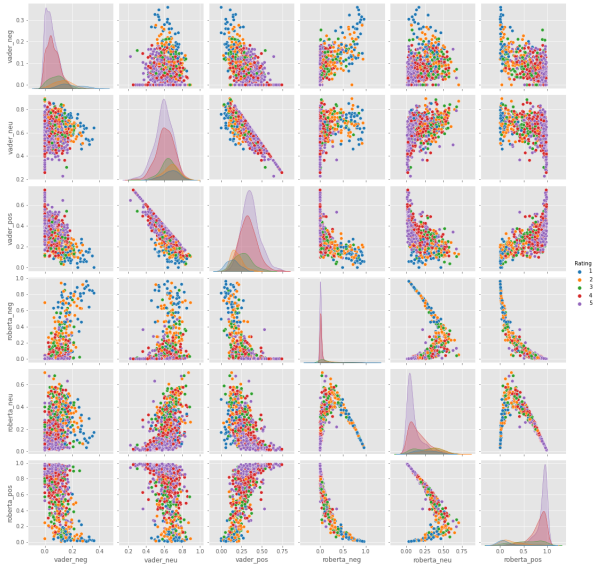

Fig. 5. Comparison between VADER and RoBERTa

The VADER model was then used for sentiment analysis for the aspect keywords, in which we only considered the positive and negative values and according to whichever value was the maximum, we classified the review as 'pos' for positive sentiment and 'neg' for negative sentiment. From the output we got 931 reviews as positive and 69 reviews as negative.

The ROBERTA model was then used in a similar way for the aspect keywords and from that got 884 reviews as positive and 116 reviews as negatives.

We then converted the sentiments of the reviews into integers for both VADER and ROBERTA results. If the sentiment was positive we converted into 1 and if it was negative we converted it into 0. We assumed the sentiment as positive if the rating is higher than 3 and if it is not the sentiment is negative. There were 752 positive reviews and 248 negative reviews.

We imported sklearn.metrics to calculate the accuracy, f1 score, recall and precision for our models. For VADER model we got - Accuracy - 0.7937937937937938, F1 score - 0.8781065088757396, Recall - 0.9880159786950732, and Precision - 0.7902023429179978.

And for RoBERTa model we got - Accuracy - 0.8228228228228228, F1 score - 0.8917431192660551, Recall - 0.9707057256990679, and Precision - 0.8246606334841629.

VI. Limitations

The majority of NLU models developed above the word level are task-specific and have trouble processing input from other domains and automated extraction of information is quite difficult due to the ambiguity and metaphorical terms that are frequently present in natural language.

Any word that is not in the VADER's lexicon is regarded as neutral which can skew the performance of the model. VADER operate best in microblog-like settings meaning it takes into account only the individual words being used, entirely ignoring the larger context. Misspellings and grammatical errors may also cause the analysis to overlook significant terms or use. Irony, sarcasm, discriminating jargon, terminology, memes, or phrasal verbs can be misunderstood.

The majority of RoBERTa's flaws are related to its size. While training the data on a big corpus improves its performance, there is another aspect to consider. Training takes longer and there are many weights to update. Just like BERT, Roberta is still limited to 512 tokens in total.

VII. Conclusion

The focus of our research is to evaluate the sentiments based on aspect terms. It is impossible to assign humans to determine the sentiment of a text or a review. We tried to implement a system that will do the work for us and give a satisfactory result. In our research, RoBERTa model worked best for us with an F1 score of 0.8917431192660551. A logical continuation of this paper would be to investigate the viability of using our system in additional contexts and languages.

References

[1] Kang, Hyun Jung, and Iris Eshkol. "An empirical examination of online restaurant reviews." *Proceedings of the 12th Language Resources and Evaluation Conference.* 2020.

[2] Diaz, Gerardo Ocampo, Xuanming Zhang, and Vincent Ng. "Aspect-based sentiment analysis as fine-grained opinion mining." *Proceedings of the 12th language resources and evaluation conference.* 2020.

[3] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[4] Xu, Hu, et al. "BERT post-training for review reading comprehension and aspect-based sentiment analysis." *arXiv preprint arXiv:1904.02232* (2019).

[5] Cheng, Jianpeng, Li Dong, and Mirella Lapata. "Long short-term memory-networks for machine reading." *arXiv preprint arXiv:1601.06733* (2016).

[6] Sun, Chi, Luyao Huang, and Xipeng Qiu. "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence." *arXiv preprint arXiv:1903.09588* (2019).

[7] Wang, Alex, et al. "GLUE: A multi-task benchmark and analysis platform for natural language understanding." *arXiv preprint arXiv:1804.07461* (2018).

[8] Cambria, Erik, et al. "New avenues in opinion mining and sentiment analysis." *IEEE Intelligent systems 28.2* (2013): 15-21.

[9] Chaturvedi, Iti, et al. "Distinguishing between facts and opinions for sentiment analysis: Survey and challenges." *Information Fusion* 44 (2018): 65-77.

[10] Bakshi, Rushlene Kaur, et al. "Opinion mining and sentiment analysis." *2016 3rd international conference on computing for sustainable global development (INDIACom).* IEEE, 2016.

[11] Thelwall, Mike, et al. "Sentiment strength detection in short informal text." *Journal of the American society for information science and technology* 61.12 (2010): 2544-2558.