**Exploring Sentiment Analysis using VADER and RoBERTa for Organizations**

Tyler Loest

Department of Data Science, University of Wisconsin- Oshkosh

DS 785

Dr. Thomas Buck

December 11, 2022

**Abstract**

This paper explores how sentiment analysis is helpful in various types of organizations. retail (Amazon), healthcare, and education are the three areas this paper will explore, but the techniques and results can be used in other areas. With the growth of social media and online presence, people are expressing their opinions daily. Sentiment analysis can help determine whether something is negative, neutral, or positive right away and not have to read every single review or tweet. Using an amazon dataset and gathering tweets using the Twitter API, I conducted sentiment analysis using the VADER (Valence Aware Dictionary for sEntiment Reasoning) and RoBERTa (Robustly Optimized BERT Pre-training Approach) methods and then compared them. The comparison results were largely the same, with the RoBERTa method being more accurate due to being a more powerful tool. Overall, the results will help show that organizations should utilize sentiment analysis to have a better relationship with their customers or clients.

**Contents**

Abstract……………………………………………………………………………………………2

List of Figures……………………………………………………………………………………. 5

Chapter 1: Background……………………………………………………………………………6

What is the problem?...........................................................................................................6

What is the purpose of your project?...................................................................................6

What are the project objectives?................................................…………………………..6

What is the significance of the project?...............................................................................6

The History of Sentiment Analysis………………………………………………………..7

The importance of Sentiment Analysis……………………………………………………7

What are the limitations of the project?...............................................................................8

Chapter 2: Literature Review……………………………………………………………………10

Chapter 3: Data Collection/Methodology………………………………………………………..18

Research Questions………………………………………………………………………18

Study Design……………………………………………………………………………..18

Participants……………………………………………………………………………….18

Data Sources……………………………………………………………………………..18

Data Collection…………………………………………………………………………..19

Data Analysis…………………………………………………………………………….20

Ethical Considerations…………………………………………………………………...23

Assumptions, Delimitations, and Limitations……………………………………………23

Chapter 4: Findings/Results……………………………………………………………………...24

Chapter 5: Summary/Recommendation/Conclusion……………………………………………..30

Summary of findings……………………………………………………………………..30

Interpretation of findings………………………………………………………………...30

Amazon…………………………………………………………………………..30

Education and Healthcare Tweets………………………………………………..35

Context of findings………………………………………………………………………35

Implications of findings………………………………………………………………….36

Discussion on limitations of the study.…………………………………………………..36

Discussion on future directions of research/field………………………………………...37

Conclusion……………………………………………………………………………….37

References………………………………………………………………………………………..38

Appendix A: Link to the Python Code…………………………………………………………..41

**List of Figures**

**Figure 1:** Amazon Pantry Positive VADER Score……………………………………………...25

**Figure 2:** Amazon Pantry Positive VADER Score- Cleaned……………………………………25

**Figure 3:** Amazon Pantry Neutral VADER Score………………………………………………26

**Figure 4:** Amazon Pantry Neutral VADER Score- Cleaned…………………………………….26

**Figure 5:** Amazon Pantry Negative RoBERTa Score…………………………………………...27

**Figure 6:** Amazon Pantry Negative RoBERTa Score- Cleaned………………………………...27

**Figure 7:** Amazon Pantry Neutral RoBERTa Score…………………………………………….28

**Figure 8:** Amazon Pantry Positive RoBERTa Score- Cleaned………………………………….28

**Figure 9:** Education Tweets Pair Plot Cleaned………………………………………………….29

**Figure 10:** Amazon Pantry Negative VADER Score……………………………………………31

**Figure 11:** Amazon Pantry Negative VADER Score- Cleaned…………………………………31

**Figure 12:** Amazon Pantry Compound VADER Score…………………………………………32

**Figure 13:** Amazon Pantry Compound VADER Score- Cleaned……………………………….32

**Figure 14:** Amazon Pantry Positive RoBERTa Score…………………………………………..33

**Figure 15:** Amazon Pantry Neutral RoBERTa Score- Cleaned…………………………………33

**Figure 16:** Amazon Pantry VADER/RoBERTa Positive Scores Pair Plot- Cleaned…………...34

**Figure 17:** Amazon Pantry VADER/RoBERTa Negative Scores Pair Plot- Cleaned…………..34

**Figure 18:** Amazon Pantry VADER/RoBERTa Neutral Scores Pair Plot- Cleaned……………35

**Chapter 1:** Background

**What is the problem?**

More people are online today than ever, posting on social media and leaving reviews on everything. It is almost impossible to go over what people are saying about companies and organizations that they would need help going over them and determine if they are positive or negative. This is where sentiment analysis comes in.

**What is the purpose of your project?**

This case study aims to show companies and organizations the importance and usefulness of sentiment analysis. Nowadays, companies are hiring data scientists who are most likely skilled in Python, which has sentiment analysis capabilities.

**What are the project objectives?**

This case study has four objectives. The first is to detail the history of sentiment analysis, such as what tools data scientists used in the past. Next would be discussing why sentiment analysis is essential. Furthermore, it demonstrates what data scientists are using today for sentiment analysis. Lastly, explore what sentiment analysis can be like in the future and what technologies are being developed for it.

**What is the significance of the project?**

The significance of this case study is to convince organizations and companies, particularly in education, healthcare, and retail, to use sentiment analysis to improve their operations and be on good terms with their customers. The importance of sentiment analysis section expands on it.

**The History of Sentiment Analysis**

Sentiment Analysis can trace its modern beginnings in the 1950s, long before computers were widespread. Nowadays, sentiment analysis is being implemented on websites like Twitter, blogs, Facebook, news articles, reviews, and comments to get personal information. Many tools perform sentiment analysis, like statistics, machine learning, and natural language processing. Recently, the Obama Administration used sentiment analysis to see the public’s opinions on his policy announcements.

The two types of sentiment analysis are subjectivity/objectivity identification and feature/aspect-based. Subjectivity/objectivity identification puts a sentence or partial text into two categories, subjective or objective. One disadvantage of this type of sentiment analysis is that words or phrases often depend on the context and may not be entirely accurate when classifying. Feature/aspect-based sentiment analysis determines different opinions or features of an object’s different qualities. This allows for a subtle overview of views and feelings.

**The importance of Sentiment Analysis**

Today, sentiment analysis is vitally crucial as “Social networking, blogging, and online forums have turned the Web into a vast repository of comments on many topics, generating a potential source of information for social science research.” (quoted in Thelwall et al., 2011 as cited in Thelwall, Wouters, & Fry, 2008). Websites have a lot of valuable data and information that businesses are interested in. With customers writing their opinions more openly on the internet, companies can analyze them to give insights into their reputation and discover new opportunities to improve their business and advertise their products. Using sentiment analysis, companies can save time and resources in determining the most critical aspects of their customer’s opinions. They also don’t have to rely on their employees who may have biases or ill intent toward the organization. “Today’s robots are able to accurately, unbiasedly, without fatigue, interpret more language-based data than humans” (Ajitha et al., 2021). There are many applications for sentiment analysis, including social media monitoring, marketing, data mining, and political commentary.

With its ever-increasing popularity, social media has become a place where people are unafraid to share their opinions on anything. This treasure trove of valuable information that businesses and organizations can use. One advantage of looking at social media for sentiment analysis is that many of them already have tools that have some sentiment analysis abilities or to collect data for analysis, such as Twitter API. Social media is helpful to businesses as it can get people’s opinions on them promptly (Singh, 2022). They can use the information to try and fix customers’ complaints in an easy and fast way.

Another application for sentiment analysis is marketing. People like leaving reviews on the product’s site, so businesses can use sentiment analysis to save time and not rely on surveys to determine people’s opinions on their products. They can use the research to determine the general tone of the reviews to get a more accurate view of their products.

Lastly, one of the essential applications of sentiment analysis is data mining. Similar to the previous paragraph, companies can track other competitors’ social media mentions or the web to see how consumers feel about them. They can use the information to make changes or improvements that the competitors are not doing to gain an advantage in the marketplace.

**What are the limitations of the project?**

While this is a general case study and an overview of sentiment analysis, it would likely cause businesses and organizations to seek further data specific to their industry. Also, since there are many different analysis programs, the ones shown in this case study might not fit certain businesses and organizations.

**Chapter 2:** Literature Review

After discussing the case study’s main objectives and general overview, this literature review will explore sentiment analysis more by showing how it can be applied to many different research areas. This would be accomplished by discussing peer-reviewed journal articles that would help convince businesses and organizations to use sentiment analysis in their operations. This chapter is going to use six journal articles.

The first journal article is by Geler et al. (2021) mentions two sentiment analysis techniques but focuses on machine learning on food services and restaurant reviews. The report would also be helpful for businesses that combine studies and a number scale based on satisfaction in specific areas as it uses different regression models to predict customer satisfaction. They converted the top keywords from the comments into a bag-of-words to use the textual data. Then they took each of the bag-of-words and made new attributes based on the number of unique keywords. Finally, they figured that the values of the new attributes could be based on different frequencies, so they ended up with two different datasets. The two frequencies they did were the inverse document frequency (IDF) and a 0/1 score depending on if the keyword was in it the top-10 keyword or not. This article’s authors define IDF as the “value of a given keyword for all instances that had the particular keyword in their top-10 keywords list.” In other words, how common or rare a keyword is in a document. This is an intelligent way of transforming the text as you can easily count the repeated keywords in the 0/1 dataset and see how vital a keyword would be in the IDF dataset. This article is an excellent example of how transforming data can yield different results by having two different datasets and using six different regression techniques. More importantly, it shows how sentiment analysis can be applied to structured analysis techniques such as prediction. One disadvantage I found when reading this article is that some sentences were structured weirdly, and I needed to research some terms. One positive aspect of this article was how they converted textual words into something they could use for regression.

The following article, authored by Ivaturi and Bhagwatwar (2020), tackles this question: “How can firms develop intelligence by strategically analyzing customer reactions during crisis events such as security hacks?” This article would suit businesses who want to respond to a crisis intelligently. They answered this question by looking at two retailers, Target and Home Depot, when they had a security hack by looking at Facebook. Along with sentiment analysis, they also included justice theory to provide context, so the research is just not “positive,” “neutral,” or “negative.” According to the authors, “The justice theory is an appropriate lens for this work as it explicitly explores how customers react to firms’ actions in response to service complaints.” In other words, justice theory is how people react to a firm’s decision and if those decisions are perceived to be fair. This shows how sentiment can be combined with other concepts to create a more advanced way to analyze data. The authors concluded that their study suggested that customers like policies that help them recuperate any losses with justice more than the security hack’s interaction and outcome. Companies will have more customer satisfaction by having companies realize this and how they convey and interact on their social media pages. One aspect I didn’t like about this article is that they explained the Target and Home Depot security breach background in more detail than needed. The aspect that I liked about the article is that they used sentiment analysis and combined it with a different concept.

The next article by Saura et al. (2021) focuses on start-ups in India. Their goal is to identify problems in the Indian startup environment and detect people’s feelings about those problems. This will lead investors to obtain the information and help decision-making approaches when investing in start-ups. This article would benefit startup companies to help get money from investors or for investors to look and evaluate prospective startup companies they are willing to invest in. To figure this out, the researchers decided to explore these three questions:

* “Can topics of interest for investors in Indian startups be identified from the UGC (user-generated content) on Twitter?”
* “What are the sentiments (positive, negative, or neutral) of the identified topics about Indian startups, and what is their connection with possible investments?”
* “Can indicators that help investors invest in Indian startups from UGC in Twitter be identified? Is it possible to divide these indicators into positive, negative, and neutral to help investors make better decisions?”

To answer these questions, they did a three-step process based on a paper from Saura and Bennett (2019). The first process uses a latent Dirichlet (LDA) allocation to group words into topics. The second, they used an algorithm using machine learning and data mining to perform sentiment analysis. They used textual analysis to apply data classified using the previous steps in the last step. Their research provided an overview of which startups liked and disliked. Some examples that they like were “startups that develop business models based on innovation” and startups that create answers centered around artificial intelligence and augmented reality. Ones disliked were startups that produce Indian hardware and startups that participate in competitive projects in seminars to get investment. This article was powerful as I don’t see a specific weakness of it, but an aspect I liked is that the conclusions they came up with were for particular types of startups.

The article by Yaqub et al. (2021) explores sentiment analysis and topic modeling on Twitter and Facebook of nine different public sector organizations in the Northeast US, consisting of law enforcement, transportation, utility services, and the department of motor vehicles. They aim to compare and contrast sentiment analysis on the content and topics of discussion on social media. Their goals are:

* “Observe frequency, sentiment and content of social media messages by organizations operating in disparate areas of public service.”
* “Evaluate how and why their messages are similar or different from each other.”
* “Perform topic modeling on their social media messages to identify main themes of their messages and how these themes are similar or different for these organizations.”

The authors also developed four researched questions for the article:

1. “How the nature of organization’s operations affects the sentiment of it’s social media messages?”
   1. They found that it was primarily negative, but it’s because it’s from law enforcement, and most of the messages are crime-related, not really for public communication.
2. “How frequently these organizations are using social media for communication? How does it relate with their operations?”
   1. They found that they quickly relayed delays and emergencies with the public but not so much promoting the benefits of public transport.
3. “How many organizations use # to make their tweets more searchable to users and how much do these organizations engage in direct messaging with other users, creating a conversation and answering their queries directly?”
   1. They did not find a clear pattern.
4. “How the nature of organization’s operations affects the topics being discussed in their messages on social media? Do organizations operating in similar areas have similar topics in their online messages?”
   1. They found the same results as the first question, only using social media to post crime-related events.

To answer these questions and satisfy their goals, they gathered posts from Facebook and Twitter using Facepager and Twitter API, collected them into CSV files, and cleaned and added them to MySQL. Next, they did sentiment analysis using a tool called SentiStrength. It is used to analyze short informal texts. They then use the same model that Saura et al. (2021) used, the LDA model. This article was very well-written and easy to understand, and I couldn’t find anything wrong with it.

The article by Alharbi et al. (2021) is about “This is research will evaluate different sentiment analysis approaches for the dataset of the mobile phone reviews in order to predict consumers’ satisfaction for a mobile phone reviews by using deep learning algorithms, including five different RNN models to be evaluated based on their performance. Nevertheless, this will also help buyers to make better decisions when considering the purchases of a specific mobile phone.” After preparing the dataset for analysis, they divided it based on the review ratings where 1-2 were considered negative, three were neutral, and 4-5 were positive. They did this so they could implement what they were going to use for analysis; as quoted by them, “Many machine learning algorithms and virtually all architectures of the deep learning are unable to process strings or plain text in their raw form or to perform any kind of work, such as the classification and the regression.” They used three different programs:

1. word2vec Embedding
   1. It uses math to find similarities in words, according to the authors.
2. Glove Vector Embedding
   1. It is from Global Vectors and uses a factorization matrix on a word-context matrix. It is also a model that uses an unsupervised learning algorithm to obtain vector representation for words.
3. FastText Vector Embedding
   1. It Is similar to Glove, but it also can split the words into subwords for further analysis.

They then used for different RNN models to perform Long Short-Term Memory Networks (LRNN), Group Long ShortTerm Memory Networks (GLRNN), gated recurrent unit (GRNN), and update recurrent unit (UGRNN). They concluded that the GLSTM with FastText was the best one that resulted in accuracy, precision, and recall. One problem with the paper is that they didn’t define what RNN stood for, which isn’t good if people don’t know what it stood for. Also, it was very technical, which was hard to understand in some places.

The last article by Peng et al. (2022) is about analyzing text using sentiment analysis but taking it further by researching how students respond to the text. To do this, they did two models using their own words:

1. “In order to improve the text feature extraction ability of the proposed model, it uses the CNN model to set up convolution windows of different sizes to extract the binary and ternary features of the text. In addition, BLSTM is used for sequence feature extraction, thereby providing highly reliable text features for subsequent classification.”
2. “In view of the lack of consideration of the relationship between emotions in traditional analysis models, the proposed model introduces an attention mechanism. It can adaptively combine context information and student emotion information to extract key text features that affect student emotions, effectively improving the accuracy of emotion classification.”

CNN stands for “Convolutional neural network,” and BLSTM stands for “Bidirectional Long Short Term Memory.” They concluded that there were more misclassifications for sadness and no emotion. By adding an attention mechanism, it can improve the accuracy of the analysis. This article was easy to read and explained the concepts well, but they sometimes didn’t define the acronyms. For example, they didn’t say what CNN and BLSTM stood for, so I had to look them up.

The previous articles provide ways to use sentiment analysis individually, but I believe all six pieces could be shared. One example is using the first and second articles. One could study past security breaches and their social media leading up to it and convert them into structured data so they possibly predict if they would be vulnerable to a hack. The company could have been not doing so good in customer satisfaction, so a hacker could decide to breach their website and cause havoc to them. If they keep up with reasonable satisfaction and standing, they would be less likely to get hacked and move to a different company that does not have a good reputation.

The first and third articles could be combined to predict whether an investor would invest in a startup company. You can take a dataset similar to what article one used and take the reviews or comments to analyze what article 3 did. Then you can compare them to see if the predictions are similar or not. If they are identical, you might have an excellent prediction model.

The second and third articles could be similarly combined, like one and three as three is also prediction. They can take information from social media and perform their three-step process to see if they can predict a potential hack threat.

The second and fourth articles both use the LDA model. It seems like the LDA model is a valuable tool in the topic of sentiment analysis. It is beneficial for organizations interested in how the public feels about them. It would be a helpful tool to look into for them.

The fifth article also talks about prediction, like some other articles use. This demonstrates many research opportunities to use prediction models with sentiment analysis. This can be useful in the future so people can see how their messages come across before sending them to people.

The sixth article could be combined with the other prediction articles so teachers can predict how their students will respond to them. This can be used to better communicate with students and teachers, with the teachers predicting how the students will react.

In conclusion, there are many ways that sentiment analysis can be used, such as using it with regression and with different concepts such as justice theory. By showing these examples, I hope to show businesses why they should be using sentiment analysis and how sentiment analysis can be applied to many different topics. The next chapter will describe the collection process to demonstrate current techniques and explain why those tools are excellent for sentiment analysis.

**Chapter 3:** Data Collection/Methodology

**Research Questions**

Many businesses and organizations want to be better at helping and understanding their customers’ feelings about them. They can do surveys and polls, but they are often inaccurate and might not solve the customers’ problems. So how can a company do better in that department? By using sentiment analysis.

**Study Design**

I wanted to show different ways of sentiment analysis and show the evolution of it through the VADER (Valence Aware Dictionary for sEntiment Reasoning) approach, which is a basic approach that only looks at each word separately to the RoBERTa (Robustly Optimized BERT Pre-training Approach) approach, which evolved from the VADER method to take into account how the words are related to each other in the sentences. By doing the analysis, on the uncleaned and cleaned datasets, it can show how much of a difference the results will be, and eliminating the “stop words” allow more rows (text) to be analyzed. This will be shown in graphs and other supplementary data explained in the next chapter.

**Participants**

There are no direct participants in this case study, but user reviews and tweets from people all over that are anonymous.

**Data Sources**

The data sources are from an existing dataset in a JSON file and later converted into a CSV file. Also, tweets from Twitter using the Twitter API.

**Data Collection**

Data Collection was straightforward, as tweets were collected using the free Twitter API elevated package. The number of tweets was 10,000 each for education and healthcare. The existing Amazon dataset contains Amazon Pantry reviews between May 1996 and Oct 2018. It is part of a massive dataset of various Amazon reviews from many categories. They have two sections, one containing the raw reviews and one reduced to their k-cores for smaller experimentation such as this. After cleaning and analysis, the initial dataset contained 137788 rows but was decreased slightly.

First, I signed up for the Twitter API, which requires a Twitter account. After I signed up, I used the Jupyter IDE to write the Python code to utilize the Twitter API and clean and analyze the datasets I used. Then I used the tweepy package to collect the tweets using the mentioned hashtags. I cleaned the dataset to prepare it for sentiment analysis. I used the exact cleaning method for each of the three datasets. I removed the punctuation from the text by using a function and applying it to each of the text columns. Then I tokenized each of the words in the text. The next step was to remove “stop words” or words that don’t add any sentiment, such as “the” and “a,” using the python package NLTK (Natural Language Toolkit), which has its list of “stop words.” After this, I converted the data frame into a CSV file to download for manual cleaning, as the tokenized words had some characters that prevented me from analyzing the data. Such characters included the brackets, coma, and single quotes (‘). Once cleaned, it was uploaded back to jupyter, and finally did, the analysis.

**Data Analysis**

There are many tools for sentiment analysis, but for this study, we did the VADER and the RoBERTa methods. Both can be used in python, costing the analyst nothing. The python libraries I used were:

* os
* tweepy
* pandas
* numpy
* matplotlib.pyplot
* seaborn
* plt.style.use('ggplot')
* string
* regex
* %matplotlib inline
* import nltk
  + nltk.download('punkt')
  + nltk.download('averaged\_perceptron\_tagger')
  + nltk.download('maxent\_ne\_chunker')
  + nltk.download('words')
  + nltk.download('vader\_lexicon')
  + nltk.download('stopwords')
* from tqdm.notebook import tqdm
* transformers import AutoTokenizer
* from transformers import AutoModelForSequenceClassification
* from scipy.special import softmax
* from nltk.sentiment import SentimentIntensityAnalyzer

Both methods use both qualitative and quantitative data by taking qualitative data and transforming it into quantitative data. The data that was analyzed came from two sources, primary data, and secondary data. The primary data was from analyzing tweets gathered by the Twitter API by using the hashtag #education and then again by the hashtag #healthcare. The secondary data is from reviews from the prime pantry section of Amazon, collected by Jianmo Ni from the University of California San Diego. By analyzing this data using the free VADER and RoBERTa methods, the companies will benefit tremendously by knowing what their customers feel about them and making better products and services for them.

The first method of analysis was the RoBERTa method. “RoBERTa is an improved version of the BERT (Bidirectional Encoder Representations from Transformers) framework, owing to several modifications such as dynamic masking, input changes without the next-sentence-prediction loss, large mini-batches, etc. The BERT framework itself is based on transformers, a deep learning model built upon several encoder layers and multi-heads self-attention mechanism.” referenced by Liu et al., 2019, Devlin et al., 2018, and Vaswani et al., 2017 as quoted by Yan, T., & Liu, F. (2022).

The method takes a model to classify the data to recognize the relationship between words. This robust model is based on the opened source BERT (Bidirectional Encoder Representations from Transformers) model. In other words, it’s a model that understands language well and can accomplish many language tasks. An excellent example of using BERT is if you ever Google searched something. “To achieve better vector representation, the Bert model uses a transformer structure to learn the contextual information of the input words, which integrates a multi-headed self-attentive mechanism to comprehensively mine the information in different locations under different subspaces and encode the information representation to each location. The main innovation of BERT is using a masked language model and next sentence prediction to capture word and sentence representations, respectively.” (Cao et al., 2022). Facebook Research AI agency then took BERT and improved upon it, which resulted in the RoBERTa model. That is the reason why I did RoBERTa instead as it’s the evolution of BERT. The one used in all the datasets is the twitter RoBERTa base sentiment from Hugging Face, a company known for the transformers library. This part of the analysis took the longest as the Amazon dataset; it took about two hours. Once done, I converted the list of scores and combined them with the original cleaned dataset.

Then I did the VADER analysis on the combined dataset with the RoBERTa results. The VADER analysis was released in 2014. Its goal was to use sentiment analysis that senses the polarity (positive and negative) and the intensity of the emotion. As mentioned in Chiny (2021, quoted in Majidpour, J., & Al-Barznji, K., 2022). “In order to obtain a reliable point estimate of the sentiment valence (intensity) of each lexicalattribute, VADER is based on a wisdom of crowds (WotC) technique.” “Humans evaluate and approve the lexicons. They use qualitative techniques to enhance the emotion analyzer's effectiveness” as mentioned by Reshi et al. (2022) as quoted in Majidpour, J., & Al-Barznji, K. (2022). I separated the dictionary into separate columns, which took much less time as it only takes each word and scores them as negative, neutral, and positive, with no relationship between them. The Amazon dataset was in a JSON file, so I had to convert the text into a string. I did the same analysis but did not clean the original dataset to compare the two.

After the analysis, I took the education and healthcare analysis datasets and then grouped them into 11 different groups; 0-999, 1000-1999,…,9000-9999, and 10,000+. I then took the average of the sentiment scores to get it ready to be able to see graphs and plots a bit better by taking out unnecessary columns that were unnamed.

**Ethical Considerations**

For the study itself, a possible ethical consideration is that a tweet collection reveals the person who tweeted it, but it also sets it as an ID, so you can use that if you want to point out specific tweets that stand out. The primary ethical consideration is the possible contact with the person who made the tweet or review. One can easily fall into harassment of the person, but it can be avoided if the review can be commented on and replied to and see if they respond and should only be contacted once.

**Assumptions, Delimitations, and Limitations**

A limitation of this study is the Twitter API, as after about 450 tweets were gathered, it had to take about a ~13-minute break before collecting more tweets. Depending on how many tweets you wanted, the gathering could take a long time. Another limitation is that using the RoBERTa method can only take so many characters that some texts can be considered too long and will result in an error, so you would have to try and except code to skip them. A related limitation is that the RoBERTa analysis took much time to complete, with about 140,000 rows.

**Chapter 4:** Findings/Results

The first analysis method I did was the VADER method. This model is good for giving a general feeling of the text Cristescu et al. (2019). It results in four different scores, negative, neutral, positive, and compound scores. The first three scores use a scale from zero to one. It takes each word and scores them as neutral, negative, or positive. It then calculates a compound score that takes each of the three scores and results in a number between -1 (more negative sounding) to +1 (more positive sounding).

The second analysis I did was the RoBERTa method which is a more complex model than the VADER in terms of scoring the words. In RoBERTA, you use a pre-trained model so the analysis can use the relationship between the words in the sentence. It is different from VADER as VADER only scores each individual word with no relation to the other words. It also does not include a compound score, only negative, neutral, and positive scores from a scale of zero to one. It is one of the better models in terms of performance, according to Junaid et al. (2022).

I analyzed the uncleaned and cleaned data sets using these methods to showcase how words and other factors can influence the results. The first dataset I looked at was the Amazon Pantry reviews using a bar graph; I first compared the VADER scores to the review score on a scale of one to five. Since there was a rating, I didn’t use a frequency graph like Fattahoh et al. (2022) did. The results were expected to be more negative scores had more “one star” reviews while the more positive scores had more “five star” reviews by using bar graphs to compare the two. With the datasets being big, there are some outliers to be expected. After analyzing the cleaned dataset, the scores were more significant compared to the uncleaned dataset (Figures 1-2). An interesting result is that the neutral scores were the same for one to three stars but noticeably dropped for the four and five stars (Figures 3-4).

Chart, bar chart

Description automatically generated

**Figure 1:** Amazon Pantry Positive VADER Score

Chart, bar chart

Description automatically generated

**Figure 2:** Amazon Pantry Positive VADER Score- Cleaned

Chart, bar chart

Description automatically generated

**Figure 3:** Amazon Pantry Neutral VADER Score

Chart, bar chart

Description automatically generated

**Figure 4:** Amazon Pantry Neutral VADER Score- Cleaned

Another interesting observation comparing the uncleaned and cleaned datasets was that the uncleaned one had 154 errors, having too many text and index errors. The cleaned dataset reduced it to only fourteen too much text and no index errors. This is an example of how crucial cleaning the data is.

The RoBERTa results were quite interesting as the analysis differs from what one would expect. Comparing the bar graphs, the negative RoBERTa is what is expected, a more negative score with the one-star reviews than five stars (Figures 5-6). There seem to be no noticeable changes between the bar graph when showing neutral scores. That is expected for the neutral RoBERTa scores (Figure 7). For the positive score, there is a slight increase in the scores but not much of a change compared to the negative scores (Figures 5 and 8).

Chart, bar chart

Description automatically generated

**Figure 5:** Amazon Pantry Negative RoBERTa Score

Chart, bar chart

Description automatically generated

**Figure 6:** Amazon Pantry Negative RoBERTa Score- Cleaned

Chart, bar chart

Description automatically generated

**Figure 7:** Amazon Pantry Neutral RoBERTa Score

Chart, bar chart

Description automatically generated

**Figure 8:** Amazon Pantry Positive RoBERTa Score- Cleaned

Another way to compare the Amazon dataset is to use a pair plot between the VADER and RoBERTa scores using the “overall” column as the data points. The pair plots confirmed what the bar plots show us: the RoBERTa scores are similar to the review scores. The pair plot will be shown in the next chapter.

The healthcare and education datasets were a bit harder to do some preliminary exploration as the best category to compare the scores was the “Number of favorites” column. This is not a good way to indicate tone as people may “favorite” something that they agree with but is still negative sounding. It can still be helpful to look at individual tweets to see how the public feels about it and indicate that something is wrong with the topic. The preliminary exploration I did was to group the “Number of Favorites ” into 11 groups; 0-999, 1000-1999,….,9000-9999, 10,000+. This way, I can explore some graphs and plots to see how the tweets behave. The bar graphs didn’t have anything interesting as it was all over the place and wasn’t a clear pattern. The pair plots had some positive correlation (Figure 9).

Chart, scatter chart

Description automatically generated

**Figure 9:** Education Tweets Pair Plot Cleaned

**Chapter 5:** Summary/Recommendation/Conclusion

**Summary of findings**

The results of the Amazon dataset for both VADER and RoBERTa methods had some surprises but mostly what to expect. For the VADER method, the positive, negative, and compound scores resulted in what was expected, but the neutral scores seemed to behave differently.

For the RoBERTa method, there were a few surprises. Each score has about the same number of positive scores for the positive RoBERTa scores compared to the overall review scores. The negative score have the same expected result but is not as dramatic as the negative VADERS score. The neutral scores are what is to be expected in that all over the review scores have about the same number.

The healthcare and education tweets were analyzed using the VADER and RoBERTa methods. In order to interpret the findings, I grouped the number of favorites into groups and took the average pf the different scores. It was then compared using a pair plot between the two different scores,

**Interpretation of findings**

**Amazon**

For the Amazon uncleaned and cleaned datasets, the VADERS has a distinct difference between the overall scores and the sentiment scores. It makes sense that the 1-star scores have a higher negative sentiment score than the 5-star scores (Figures 10 and 11). The same goes for the 5-star scores and a higher positive sentiment score (Figures 1 and 2).

Chart, bar chart

Description automatically generated

**Figure 10:** Amazon Pantry Negative VADER Score

Chart, bar chart

Description automatically generated

**Figure 11:** Amazon Pantry Negative VADER Score- Cleaned

An interesting finding to point out is the compound sentiment score. As you can see in figures 12 and 13, the 1-star score is negative for the uncleaned dataset. This makes sense as no words were deleted, so there was more words to bring down the score. As mentioned before, the compound score is the score taking into account the other three scores.

Chart

Description automatically generated

**Figure 12:** Amazon Pantry Compound VADER Score

Chart, bar chart

Description automatically generated

**Figure 13:** Amazon Pantry Compound VADER Score- Cleaned

The RoBERTa scores it was quite different from the VADER scores, which was unexpected. As mentioned before, the positive scores were even (Figures 8 and 14). The cleaned version had more of a difference, but not a lot. An explanation of this is that the RoBERTa scores take into account the relationships between the words. As most things are grayer than “black and white.” So this would make sense as a sentence can have a lot of negative words in it, but it actually means something positive.

Chart, bar chart

Description automatically generated

**Figure 14:** Amazon Pantry Positive RoBERTa Score

Figures 3-4 are interesting in that the scores drop when the overall scores are 3 and up. One would assume it would be the same since neutral words. One explanation is that there are fewer neutral words in the sentence. Another one is that the neutral words could be marked as more negative individually. Contrast with the RoBERTa scores (Figures 7 and 15) where they are almost even. This is because of the different process the VADER and RoBERTa uses.

Chart, bar chart

Description automatically generated

**Figure 15:** Amazon Pantry Neutral RoBERTa Score- Cleaned

Using a pair plot to compare the VADER and RoBERTa scores shows that the bar graphs show that the more 5-stars reviews, the higher the positive scores are (Figure 16) and vice-versa (Figure 17).

Chart, scatter chart

Description automatically generated

**Figure 16:** Amazon Pantry VADER/RoBERTa Positive Scores Pair Plot- Cleaned

**Chart, scatter chart

Description automatically generated**

**Figure 17:** Amazon Pantry VADER/RoBERTa Negative Scores Pair Plot- Cleaned

The neutral scores are mainly evenly distributed, with a slight leaning towards the left. This is most likely because the VADERS scores had a more leaning toward zero, which would make the comparison not even (Figure 18).

**Chart

Description automatically generated**

**Figure 18:** Amazon Pantry VADER/RoBERTa Neutral Scores Pair Plot- Cleaned

**Education and Healthcare Tweets**

As mentioned in the last chapter, the bar graphs didn’t show much as the results were sporadic, but the pair plots shown interesting results. When compared to their similar scores, they form a positive line going up. This means that the scores have a correlation which makes sense as you are comparing two similar scores and grouping the number of favorites (figure 9).

**Context of findings**

This study is in agreement with the literature that I found. For instance, the Cristescu et al. (2019) article used VADER to get a general overview of the sentiment scores. This paper’s goal is to show organizations why to use sentiment analysis. Using VADER is a simple way to show that. Using RoBERTa was used in the Junaid et al (2022) article and they confirmed that the RoBERTa method is one of the best tools for sentiment analysis.

**Implications of findings**

This case study was nothing new or groundbreaking in the field of sentiment analysis, but it did satisfy the goals of the paper. It is consistent with the field as it uses current techniques of sentiment analysis such as VADER and RoBERTa which have been used numerous times. I believe this paper is for organizations that are considering incorporating sentiment analysis into their organization. It will give them a glimpse of what sentiment analysis can do and how powerful it can be when expanded on.

**Discussion on limitations of the study**

One of the biggest limitations of this study was that depending on how long the dataset was, the RoBERTa method took a very long time to calculate the resulting scores. It could be the pre-trained model that was used as it contained millions of different tweets and their relationship to the words. One could use a different smaller model to be used instead. Another big limitation is the Twitter API itself as it also takes a long time to gather the tweets. Also, Twitter could not always be around or be drastically changed in the future so it won’t be a reliable source of data anymore.

Furthermore, the Tweets themselves could be a limitation as a lot of tweets could be bots or generic tweets that only contain links to advertisements. This can affect the larger exploratory research such as bar graphs.

A surprising limitation is that the RoBERTa method has a character limitation and can’t analyze more than a certain number. It results in an error message, so you have to ignore that using an except/try code.

**Discussion on future directions of research/field**

I can see sentiment analysis becoming very popular as technology continues to grow and everyone becomes more connected online. The next step for the future will be making sentiment analysis more accurate and somehow figuring out how to detect sarcasm better. In the future, I can see having a program on every computer that analyzes what you type and determine if its positive, neutral, or negative. It’s starting to develop as Grammarly can now determine if the text is positive sounding.

**Conclusion**

Overall, this paper’s results were both expected and unexpected. It was expected because some results showed what people would naturally assume, but others did not. The VADER results were expected mainly for positive and negative scores but not neutral ones. The RoBERTa scores were largely unexpected as the positive scores had little difference. At the same time, the neutral and negative RoBERTa scores were consistent with what people would believe. This paper will help organizations to consider utilizing sentiment analysis in their business. Implementing sentiment analysis tools would allow you to address your customers’ and even employees’ needs and concerns.

**References**

Ajitha, P., Sivasangari, A., Immanuel Rajkumar, R., & Poonguzhali, S. (2021). Design of text sentiment analysis tool using feature extraction based on fusing machine learning algorithms. Journal of Intelligent & Fuzzy Systems, 40(4), 6375–6383. <https://doi-org.www.remote.uwosh.edu/10.3233/JIFS-189478>

Alharbi, N. M., Alghamdi, N. S., Alkhammash, E. H., & Al Amri, J. F. (2021). Evaluation of Sentiment Analysis via Word Embedding and RNN Variants for Amazon Online Reviews. Mathematical Problems in Engineering, 1–10. <https://doi-org.www.remote.uwosh.edu/10.1155/2021/5536560>

Calderon, P. (2018, March 31). *Vader sentiment analysis explained*. Medium. Retrieved November 24, 2022, from <https://medium.com/@piocalderon/vader-sentiment-analysis-explained-f1c4f9101cd9>

Cao, Y., Sun, Z., Li, L., & Mo, W. (2022). A Study of Sentiment Analysis Algorithms for Agricultural Product Reviews Based on Improved BERT Model. Symmetry (20738994), 14(8), 1604. <https://doi-org.www.remote.uwosh.edu/10.3390/sym14081604>

Cristescu, M. P., Nerisanu, R. A., Mara, D. A., & Oprea, S.-V. (2022). Using Market News Sentiment Analysis for Stock Market Prediction. Mathematics (2227-7390), 10(22), 4255. <https://doi-org.www.remote.uwosh.edu/10.3390/math10224255>

Fattoh, I. E., Kamal Alsheref, F., Ead, W. M., & Youssef, A. M. (2022). Semantic Sentiment Classification for COVID-19 Tweets Using Universal Sentence Encoder. Computational Intelligence & Neuroscience, 2022, 1–8. <https://doi-org.www.remote.uwosh.edu/10.1155/2022/6354543>

Geler, Z., Savić, M., Bratić, B., Kurbalija, V., Ivanović, M., & Dai, W. (2021). Sentiment prediction based on analysis of customers assessments in food serving businesses. Connection Science, 33(3), 674–692. <https://doi-org.www.remote.uwosh.edu/10.1080/09540091.2020.1870436>

Ivaturi, K., & Bhagwatwar, A. (2020). Mapping sentiments to themes of customer reactions on social media during a security hack: A justice theory perspective. Information & Management, 57(4). <https://doi-org.www.remote.uwosh.edu/10.1016/j.im.2019.103218>

Jianmo Ni, Jiacheng Li, Julian McAuley  
Empirical Methods in Natural Language Processing (EMNLP), 2019

Junaid, T., Sumathi, D., Sasikumar, A. N., Suthir, S., Manikandan, J., Khilar, R., Kuppusamy, P. G., & Janardhana Raju, M. (2022). A comparative analysis of transformer based models for figurative language classification. Computers & Electrical Engineering, 101, N.PAG. <https://doi-org.www.remote.uwosh.edu/10.1016/j.compeleceng.2022.108051>

Majidpour, J., & Al-Barznji, K. (2022). Opinions for Receiving COVID-19 Vaccines Based on Sentiment Analysis. Journal of Pharmaceutical Negative Results, 13, 1648–1659. <https://doi-org.www.remote.uwosh.edu/10.47750/pnr.2022.13.S01.197>

Peng, H., Zhang, Z., & Liu, H. (2022). A Sentiment Analysis Method for Teaching Evaluation Texts Using Attention Mechanism Combined with CNN-BLSTM Model. Scientific Programming, 1–9. <https://doi-org.www.remote.uwosh.edu/10.1155/2022/8496151>

Saura, J. R., & Bennett, D. R. (2019). A Three-Stage method for Data Text Mining: Using UGC in Business Intelligence Analysis. Symmetry (20738994), 11(4), 519. <https://doi-org.www.remote.uwosh.edu/10.3390/sym11040519>

Saura, J. R., Reyes-Menéndez, A., deMatos, N., & Correia, M. B. (2021). Identifying Startups Business Opportunities from UGC on Twitter Chatting: An Exploratory Analysis. Journal of Theoretical & Applied Electronic Commerce Research, 16(6), 1929–1944. <https://doi-org.www.remote.uwosh.edu/10.3390/jtaer16060108>

Sims, S. (n.d.). Sentiment analysis 101. KDnuggets. Retrieved March 5, 2022, from <https://www.kdnuggets.com/2015/12/sentiment-analysis-101.html>

Singh, A. (2021, July 9). *Evolving with bert: Introduction to Roberta*. Medium. Retrieved November 24, 2022, from <https://medium.com/analytics-vidhya/evolving-with-bert-introduction-to-roberta-5174ec0e7c82>

Singh. (2022). Multitask learning for complaint identification and sentiment analysis. Cognitive Computation., 14(1).

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. Journal of the American Society for Information Science & Technology, 62(2), 406–418. <https://doi-org.www.remote.uwosh.edu/10.1002/asi.21462>

Yan, T., & Liu, F. (2022). COVID-19 sentiment analysis using college subreddit data. PLoS ONE, 17(11), 1–18. <https://doi-org.www.remote.uwosh.edu/10.1371/journal.pone.0275862>

Yaqub, U., Chun, S. A., Atluri, V., Vaidya, J., Sandoval-Almazan, R., Kavanaugh, A., & Ignacio Criado, J. (2021). Analyzing social media messages of public sector organizations utilizing sentiment analysis and topic modeling. Information Polity: The International Journal of Government & Democracy in the Information Age, 26(4), 375–390. <https://doi-org.www.remote.uwosh.edu/10.3233/IP-210321>

**Appendix A:** Link to Python Code

<https://drive.google.com/drive/folders/1mexeIy1qRWGA8RlSMWjGomE1rVg7G-M5?usp=share_link>