North East University Bangladesh

Department of Computer Science and Engineering

Sentiment Analysis on Product Reviews Collected by Web Scraping from E-Commerce Websites

By

Md. Jahidul Islam
Reg. No: 200103020029
B.Sc. (Engg.) in CSE
4th year 2nd semester

Md. Ashrafuzzaman Sunny
Reg. No: 200203020002
B.Sc. (Engg.) in CSE
4th year 2nd semester

Supervised By

Razorshi Prozzwal Talukder
Lecturer
Dept. of CSE
North East University Bangladesh

Submission: 29th November 2023

Sentiment Analysis on Product Reviews Collected by Web Scraping from E-Commerce Websites

A Project submitted to the Department of Computer Science and Engineering, North East University Bangladesh, in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering

By

Md. Jahidul Islam
Reg. No: 200103020029
B.Sc. (Engg.) in CSE
4th year 2nd semester

Md. Ashrafuzzaman Sunny
Reg. No: 200203020002
B.Sc. (Engg.) in CSE
4th year 2nd semester

Supervised By

Razorshi Prozzwal Talukder
Lecturer
Dept. of CSE
North East University Bangladesh

Submission: 29th November 2023

Recommendation Letter from Project Supervisor

These Students, *Md. Jahidul Islam and Md. Ashrafuzzaman Sunny*, whose project entitled *"*Sentiment Analysis on Product Reviews Collected by Web Scraping from E-Commerce Websites*",* is under my supervision and agrees to submit for examination.

Signature of the Supervisor:

Razorshi Prozzwal Talukder
Lecturer
Dept. of CSE
North East University Bangladesh

Qualification Form of B.Sc. (Engg.) Degree

Student Name： **Md. Jahidul Islam, and Md. Ashrafuzzaman Sunny**.

Project Title: Sentiment Analysis on Product Reviews Collected by Web Scraping from E-Commerce Websites.

This is to certify that the thesis was submitted by the student named above in November 2023. It is qualified and approved by the following persons and committees.

Abstract

In the era of digital commerce, understanding customer sentiments toward products is crucial for both consumers and businesses. In the dynamic landscape of e-commerce, customer opinions and sentiments play a pivotal role in shaping purchasing decisions. Understanding the sentiment behind product reviews is valuable for consumers seeking informed choices and essential for businesses aiming to enhance their offerings and customer satisfaction. This project involves employing advanced data analysis techniques to extract meaningful insights from the plethora of product reviews on various e-commerce platforms. The purpose of this project is to develop a robust sentiment analysis model that can automatically classify product reviews into positive, negative, or neutral sentiments. For collecting the dataset; firstly, we gathered product reviews from e-commerce websites using NLP techniques and stored them in a database. Secondly, we perform sentiment analysis based on the dataset.

Keywords: Positive, Neutral, Negative, NLTK, BeautifulSoup, Scrapy, Confusion Matrix, Pre-process.

Table of Contents

List of Figures

Chapter 1

Introduction

For this project, we picked out two products from Amazon.com to perform sentiment analysis on – a PlayStation 5 and a Lawn Washing product. The PlayStation 5 has very good reviews (mostly 4 and 5 stars out of 5) while the Lawn Washer has mostly negative and neutral reviews (2 or 3 stars).

Firstly, we have used Web Scraping techniques, using the Python library BeautifulSoup, to parse the Amazon web pages of the required products, go to the reviews section, and automatically extract the reviews onto an Excel file. Similarly, the 'Rating Scores' out of 5, were also extracted onto another Excel file.

Next, we cleaned and pre-processed the data, and then performed sentiment analysis on it using two models – VADER (Valence Aware Dictionary for Sentiment Reasoning) and RoBERTa. These models are discussed in detail in the Methodology section.

Finally, we presented some results comparing the two models and showed that the RoBERTa model performed very well in identifying that the Playstation reviews were mostly positive and that the Lawn Washer reviews were mostly neutral and negative.

1.1 Challenges

Some of the Challenges we have faced doing this project are:

i) We could not perform Web Scraping on Amazon.com directly, as is usually done using the 'requests' module. This is because Amazon has recently disallowed web scraping, and it is now against the Terms of Services to do so. Therefore, we had to first save the pages locally onto our computers, and then scrape the saved webpages.

ii) It proved difficult to extract the 'reviews' and 'ratings' together, and we are working on fixing this issue to make the extraction more efficient. For the moment, we have extracted them separately.

Chapter 2

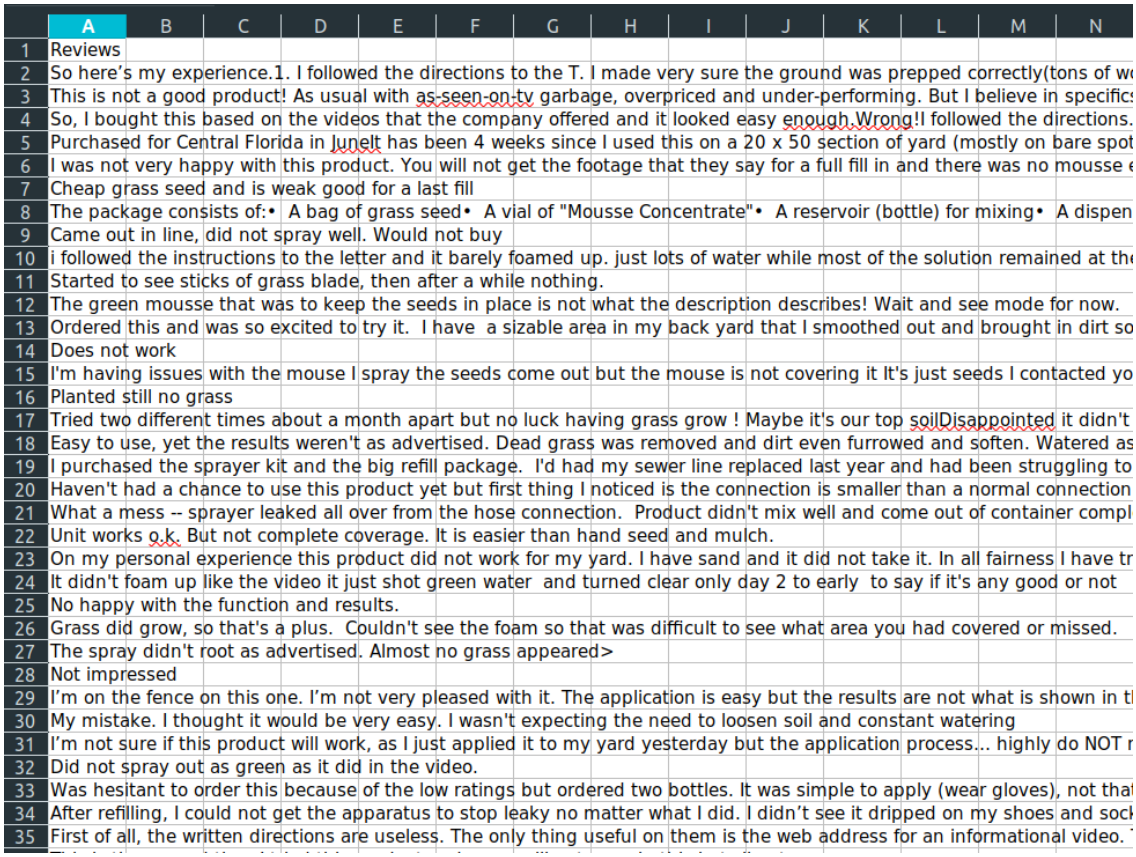
Dataset of Product Reviews

Our dataset consists of two columns, the **'Reviews'** section, and the **'Stars'** section. The reviews column contains the description of the reviews of the selected product. The 'Stars' column contains the ratings score out of 5 stars.

We have chosen two products for our dataset, a **PlayStation 5** and a **Lawn Washing Product**. We have extracted 50 reviews from each product and their rating scores, and stored them in two Excel spreadsheets, shown below:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Reviews												
2	I ordered this the day it released on Friday November 10th. It arrived at my house at 11:30AM on the following Tues												
3	Took a pic on the side of my 55" tv for size reference. Been waiting since og launch for the slim release since typical												
4	I've waited 3 years for Spider-Man 2 game bundle for the PS5 and its absolutely worth the wait! Set up takes a little												
5	Open the box to find my ps5 that i paid 499 for to be tampered with like someone was trying to steal it. Other then												
6	It was amazing. I was worried about my package coming in terrible condition but that was not the case at ALL. It was												
7	It works great and the extra storage space is awesome!												
8	I bought the console from Colombia, and the experience was great. It arrived 8 days earlier than I expected. I love th												
9	Forgot about all the YouTube videos you've seen about it not being smaller it definitely is and looks so much better												
10	I got the call of duty one and everything was fine setting it up! The PS5 works great and fast!												
11	Been waiting on the slim for months, and it was worth the wait. Love the new system Spiderman 2 bundle is sick												
12	Realice el pedido desde Colombia y tenia miedo por los comentarios sin embargo llego en perfecto estado, buena co												
13	It's perfect.... a must have if you're buying a new PS5.... remember while playing to set it horizontal so it doesn't ove												
14	I bought the console from Colombia, and the experience was great. It arrived 8 days earlier than I expected. I love th												
15	I am just a kid but overall i love the game and the consle is great to if your looking to buy one its a little pricy												
16	Even if my tv doesn't support 120 frames this looks better than my one x by a lot. Thanks Sony												
17	This is the CFI-2015 model PS5 with 1 TB hard drive. Came a day late but arrived just fine in Amazon box. No signatu												
18	It is powerful, has a 1 Terabyte hard drive, and is perfect.												
19	I already owned a ps4 before and setting up the console was amazing. It transferred all my games and data to the p												
20	It a gift for my son for Christmas he is going to love it thank you Amazon all your staff you people are great happy												
21	The price finally came to where it was affordable and the game play is wonderful.												
22	I already owned a ps4 before and setting up the console was amazing. It transferred all my games and data to the p												
23	I am just a kid but overall i love the game and the consle is great to if your looking to buy one its a little pricy												
24	It looks great, so much smaller and lighter than the original. I just wish I knew ahead of time that I couldn't get a free												
25	Even if my tv doesn't support 120 frames this looks better than my one x by a lot. Thanks Sony												
26	Todo bien, todo correcto y yo que me alegro.												
27	It is powerful, has a 1 Terabyte hard drive, and is perfect.												
28	It a gift for my son for Christmas he is going to love it thank you Amazon all your staff you people are great happy												

Fig 1: **ps5 reviews**



	A
1	Reviews
2	So here’s my experience.1. I followed the directions to the T. I made very sure the ground was prepped correctly(tons of w
3	This is not a good product! As usual with as-seen-on-tv garbage, overpriced and under-performing. But I believe in specifics
4	So, I bought this based on the videos that the company offered and it looked easy enough.Wrong!I followed the directions.
5	Purchased for Central Florida in JuneIt has been 4 weeks since I used this on a 20 x 50 section of yard (mostly on bare spot
6	I was not very happy with this product. You will not get the footage that they say for a full fill in and there was no mousse e
7	Cheap grass seed and is weak good for a last fill
8	The package consists of:• A bag of grass seed• A vial of "Mousse Concentrate"• A reservoir (bottle) for mixing• A dispen
9	Came out in line, did not spray well. Would not buy
10	i followed the instructions to the letter and it barely foamed up. just lots of water while most of the solution remained at th
11	Started to see sticks of grass blade, then after a while nothing.
12	The green mousse that was to keep the seeds in place is not what the description describes! Wait and see mode for now.
13	Ordered this and was so excited to try it. I have a sizable area in my back yard that I smoothed out and brought in dirt so
14	Does not work
15	I'm having issues with the mouse I spray the seeds come out but the mouse is not covering it It's just seeds I contacted yo
16	Planted still no grass
17	Tried two different times about a month apart but no luck having grass grow ! Maybe it's our top soilDisappointed it didn't
18	Easy to use, yet the results weren't as advertised. Dead grass was removed and dirt even furrowed and soften. Watered as
19	I purchased the sprayer kit and the big refill package. I'd had my sewer line replaced last year and had been struggling to
20	Haven't had a chance to use this product yet but first thing I noticed is the connection is smaller than a normal connection
21	What a mess -- sprayer leaked all over from the hose connection. Product didn't mix well and come out of container compl
22	Unit works o.k. But not complete coverage. It is easier than hand seed and mulch.
23	On my personal experience this product did not work for my yard. I have sand and it did not take it. In all fairness I have tr
24	It didn't foam up like the video it just shot green water and turned clear only day 2 to early to say if it's any good or not
25	No happy with the function and results.
26	Grass did grow, so that's a plus. Couldn't see the foam so that was difficult to see what area you had covered or missed.
27	The spray didn't root as advertised. Almost no grass appeared>
28	Not impressed
29	I’m on the fence on this one. I’m not very pleased with it. The application is easy but the results are not what is shown in t
30	My mistake. I thought it would be very easy. I wasn't expecting the need to loosen soil and constant watering
31	I’m not sure if this product will work, as I just applied it to my yard yesterday but the application process... highly do NOT r
32	Did not spray out as green as it did in the video.
33	Was hesitant to order this because of the low ratings but ordered two bottles. It was simple to apply (wear gloves), not that
34	After refilling, I could not get the apparatus to stop leaky no matter what I did. I didn’t see it dripped on my shoes and sock
35	First of all, the written directions are useless. The only thing useful on them is the web address for an informational video.

Fig 2: **Lawn washer reviews**

	A	B
1	Stars	
2	5.0 out of 5 stars	
3	5.0 out of 5 stars	
4	5.0 out of 5 stars	
5	4.0 out of 5 stars	
6	5.0 out of 5 stars	
7	4.0 out of 5 stars	
8	5.0 out of 5 stars	
9	5.0 out of 5 stars	
10	5.0 out of 5 stars	
11	5.0 out of 5 stars	
12	5.0 out of 5 stars	
13	5.0 out of 5 stars	
14	5.0 out of 5 stars	
15	5.0 out of 5 stars	
16	5.0 out of 5 stars	
17	5.0 out of 5 stars	
18	5.0 out of 5 stars	
19	5.0 out of 5 stars	
20	5.0 out of 5 stars	
21	5.0 out of 5 stars	
22	5.0 out of 5 stars	
23	5.0 out of 5 stars	
24	5.0 out of 5 stars	
25	5.0 out of 5 stars	
26	5.0 out of 5 stars	
27	5.0 out of 5 stars	
28	5.0 out of 5 stars	
29	5.0 out of 5 stars	
30	5.0 out of 5 stars	

Fig 3: **ps5 rating**

	A	B
1	Stars	
2	3.0 out of 5 stars	
3	2.0 out of 5 stars	
4	2.0 out of 5 stars	
5	3.0 out of 5 stars	
6	3.0 out of 5 stars	
7	3.0 out of 5 stars	
8	3.0 out of 5 stars	
9	3.0 out of 5 stars	
10	3.0 out of 5 stars	
11	3.0 out of 5 stars	
12	3.0 out of 5 stars	
13	2.0 out of 5 stars	
14	3.0 out of 5 stars	
15	3.0 out of 5 stars	
16	3.0 out of 5 stars	
17	3.0 out of 5 stars	
18	3.0 out of 5 stars	
19	2.0 out of 5 stars	
20	3.0 out of 5 stars	
21	2.0 out of 5 stars	
22	3.0 out of 5 stars	
23	2.0 out of 5 stars	
24	3.0 out of 5 stars	
25	3.0 out of 5 stars	
26	3.0 out of 5 stars	
27	3.0 out of 5 stars	
28	3.0 out of 5 stars	
29	3.0 out of 5 stars	

Fig 4: **lawn ratings**

Chapter 3

Dataset Pre - Processing

At first, the ‘Reviews’ and ‘Stars’ columns for each product were merged into one Pandas data frame, to make the dataset cleaner and easier to analyze. Also, an “ID” column is added to the start of the data frame, to identify each review easily. The result is shown below:

	ID	Reviews	Stars
0	1	I ordered this the day it released on Friday N...	5.0 out of 5 stars
1	2	Took a pic on the side of my 55" tv for size r...	5.0 out of 5 stars
2	3	I've waited 3 years for Spider-Man 2 game bund...	5.0 out of 5 stars
3	4	Open the box to find my ps5 that i paid 499 fo...	4.0 out of 5 stars
4	5	It was amazing. I was worried about my package...	5.0 out of 5 stars

Fig 5: **Dataset during pre-processing**

Finally, the 'Stars' section was converted to number format, to make sentiment analysis easier.

	ID	Reviews	Stars
0	1	I ordered this the day it released on Friday N...	5.0
1	2	Took a pic on the side of my 55" tv for size r...	5.0
2	3	I've waited 3 years for Spider-Man 2 game bund...	5.0
3	4	Open the box to find my ps5 that i paid 499 fo...	4.0
4	5	It was amazing. I was worried about my package...	5.0

Fig 6: **Final dataset**

Finally, **tokenization** was done for the RoBERTa model, using the AutoTokenizer class in the Transformers library, which splits each sentence into its constituting words and maps each word to an integer value.

Below is shown an example of a review text after tokenization.

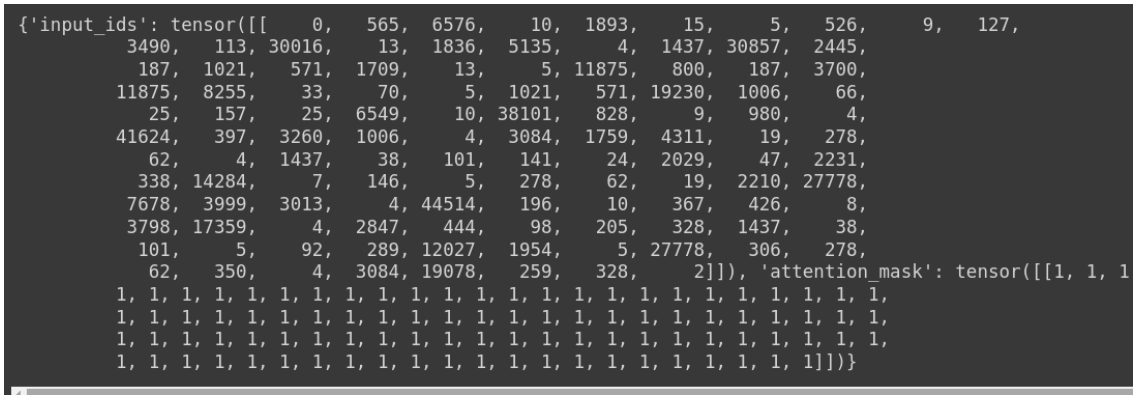


Fig 7: **Text after tokenization**

Chapter 4

Methodology

We have studied Scrapy, BeautifulSoup, NLTK, VADER model, RoBERTa, and transformer model in various online resources. From these resources, we will create a dataset that best fits our model and allows us to perform sentiment analysis over 1-1.5 months.

4.1 Our contributions

Our project contributions are as follows:
i) Code of Product reviews and rating score.
ii) VADER Model.
iii) RoBERTa model.

4.1.1. Code of Product reviews and rating scores

This has been discussed in the Dataset section. The following is the code applied for Web Scraping

```python
from bs4 import BeautifulSoup
import openpyxl

# Read HTML file
with open('amazon5_lawn.html', 'r', encoding='utf-8') as file:
    html_content = file.read()

# Parse HTML with BeautifulSoup
soup = BeautifulSoup(html_content, 'html.parser')

# Find all divs with class="a-row a-spacing-small review-data"
review_divs = soup.find_all('div', class_='a-row a-spacing-small review-data')

# Extract reviews
reviews = []
for div in review_divs:
    review_span = div.find('span', {'data-hook': 'review-body', 'class': 'a-size-base review-text review-
    if review_span:
        reviews.append(review_span.get_text(strip=True))

# Create and write to Excel file
excel_file = 'amazonLawn_reviews5.xlsx'
workbook = openpyxl.Workbook()
sheet = workbook.active

# Write header
sheet['A1'] = 'Reviews'

# Write reviews to Excel
for index, review in enumerate(reviews, start=2):
    sheet.cell(row=index, column=1, value=review)

# Save Excel file
workbook.save(excel_file)

print(f'Reviews extracted and saved to {excel_file}')
```

```python
# Read HTML file
file_path = 'amazon5_lawn.html'
with open(file_path, 'r', encoding='utf-8') as file:
    html_content = file.read()

# Parse HTML with BeautifulSoup
soup = BeautifulSoup(html_content, 'html.parser')

# Find all divs with class="a-row"
divs_a_row = soup.find_all('div', class_='a-row')

# Create Excel workbook and sheet
wb = openpyxl.Workbook()
sheet = wb.active

# Add headers to the Excel sheet
sheet.append(['Stars'])

# Loop through each div with class="a-row" and extract the required information
for div in divs_a_row:
    review_title = div.find('a', class_='a-size-base a-link-normal review-title a-color-base review-title

    # Check if review title is found
    if review_title:
        review_star_rating = review_title.find_next('i', attrs={'data-hook': 'review-star-rating'})

        # Check if review star rating is found
        if review_star_rating:
            stars_span = review_star_rating.find('span', class_='a-icon-alt')

            # Check if span with class='a-icon-alt' is found
            if stars_span:
                stars = stars_span.text

                # Append the stars value to the Excel sheet
                sheet.append([stars])

# Save the Excel file
excel_file_name = 'stars_lawn5.xlsx'
wb.save(excel_file_name)
```

Fig 8: **Web scraping reviews and rating scores.**

4.1.2 VADER Model

VADER (Valence Aware Dictionary for Sentiment Reasoning) is a model used for text sentiment analysis that is sensitive to both polarity (positive, negative, and neutral) and intensity (strength) of emotion. It is available in the NLTK package and can be applied directly to unlabeled text data.

VADER sentimental analysis relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores. The sentiment score of a text can be obtained by summing up the intensity of each word in the text.

After applying VADER to our product review data, the result looks like the following

```
{1: {'neg': 0.0, 'neu': 0.801, 'pos': 0.199, 'compound': 0.9853},
 2: {'neg': 0.027, 'neu': 0.849, 'pos': 0.124, 'compound': 0.837},
 3: {'neg': 0.0, 'neu': 0.894, 'pos': 0.106, 'compound': 0.8266},
 4: {'neg': 0.09, 'neu': 0.758, 'pos': 0.152, 'compound': 0.296},
 5: {'neg': 0.084, 'neu': 0.573, 'pos': 0.343, 'compound': 0.9236},
```

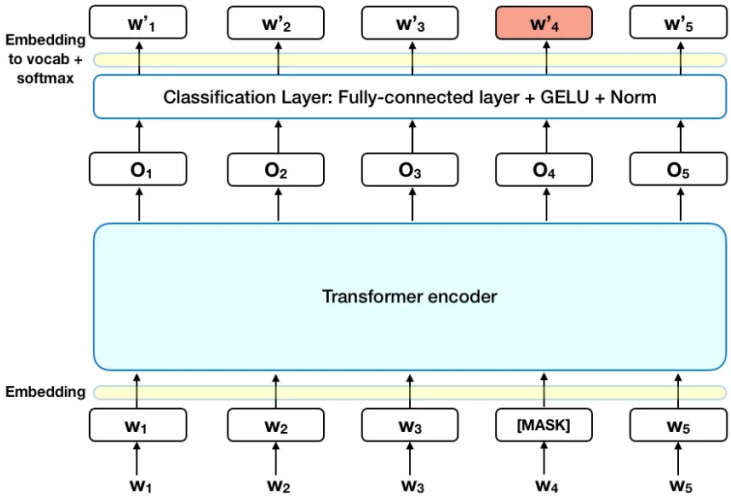
Fig 9: **VADER results.**

For each review, there is a negative, neutral, positive score, between 0 and 1. Finally, there is also a compound score which is between -1(for negative) to 1(for positive).

4.1.3 RoBERTa Model

RobERTa (Robustly optimized BERT approach) is a transformer-based model similar to BERT (Bidirectional Encoder Representations from Transformers) but with some modifications in its architecture and training strategy. It has been shown to achieve state-of-the-art results on various natural language processing (NLP) tasks, including sentiment analysis. Using RoBERTa for sentiment analysis often leads to state-of-the-art results due to its ability to capture intricate contextual information in the text. The model's deep understanding of language allows it to discern subtle nuances and context-dependent sentiments, making it well-suited for sentiment analysis tasks across various domains and languages.

Here is the main architecture of RobERTa model:



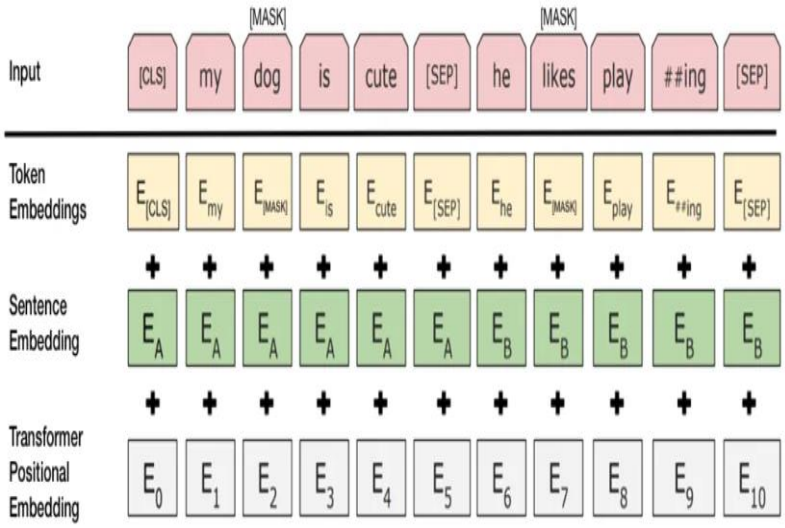


Fig 10: **Architecture of RoBERTa Model**

After applying RoBERTa to our Product Reviews dataset, it looks like the following:

```
{1: {'vader_neg': 0.0,
  'vader_neu': 0.801,
  'vader_pos': 0.199,
  'vader_compound': 0.9853,
  'r_neg': 0.0073414254,
  'r_neu': 0.053010404,
  'r_pos': 0.9396482},
 2: {'vader_neg': 0.027,
  'vader_neu': 0.849,
  'vader_pos': 0.124,
  'vader_compound': 0.837,
  'r_neg': 0.0046368847,
  'r_neu': 0.023539243,
  'r_pos': 0.9718239},
 3: {'vader_neg': 0.0,
  'vader_neu': 0.894,
  'vader_pos': 0.106,
  'vader_compound': 0.8266,
  'r_neg': 0.006855508,
  'r_neu': 0.04677482,
  'r_pos': 0.94636965},
```

Fig 11: **After applying the RoBERTa model**

The ‘r_neg’, ‘r_neu’, and ‘r_pos’ values show the results after RoBERTa. Similar to VADER, it calculates positive, negative, and neutral scores, but at a much **higher accuracy.**

Chapter 5

RESULTS AND DISCUSSION

Results so far:

We gathered two product reviews from amazon.com websites one is ps5 and the other is a lawn washer. Then we measure sentiment analysis on the entire dataset.

Below shown results about ps5 reviews and lawn washer reviews

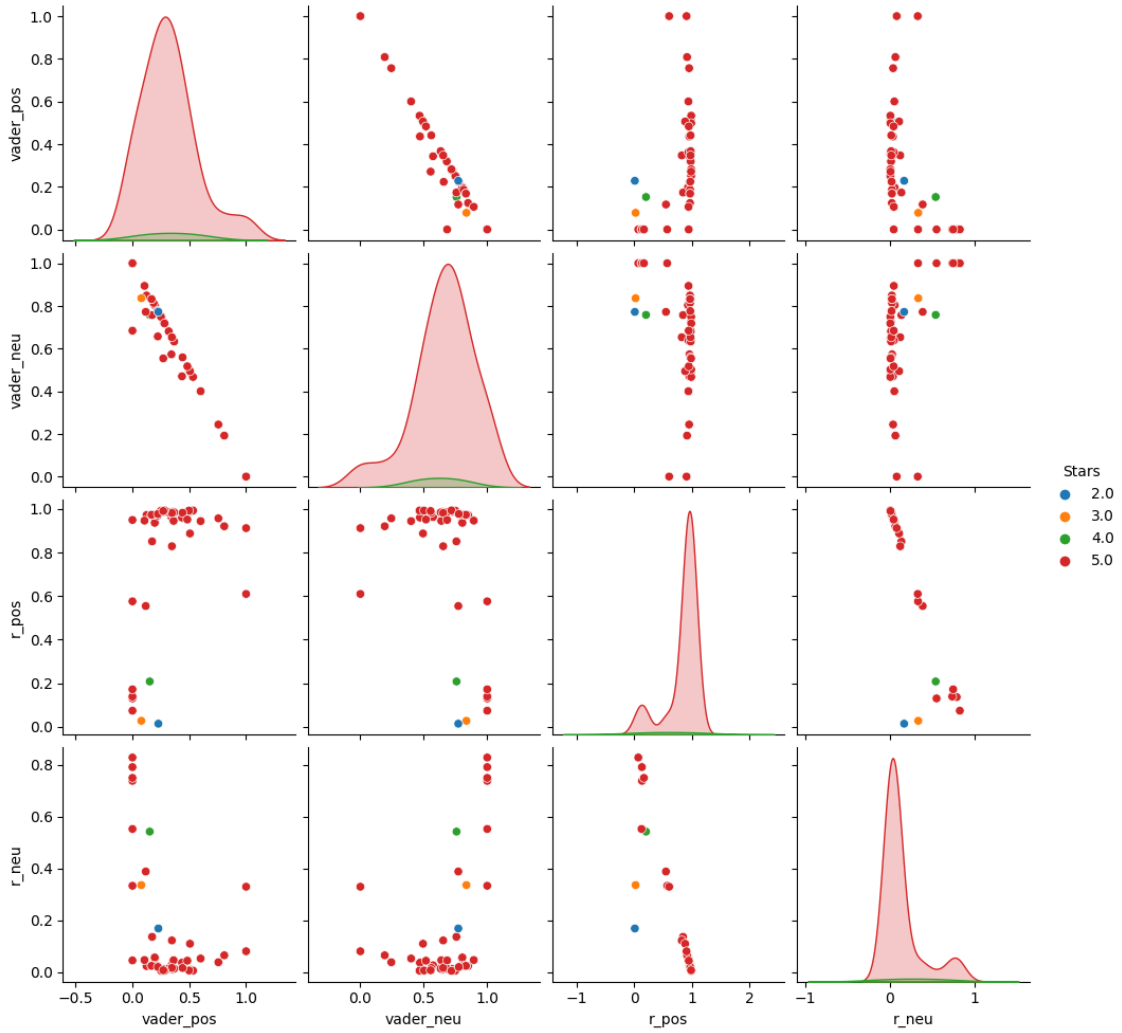


Fig 12: **Results of ps5 Reviews.**

The above graph is for the PS5 reviews, which were mostly positive. As we can see, VADER puts most of the scores in the 'Neutral' section (around 0.5) while RoBERTA can put the scores mostly in the positive section, concentrated around 1.

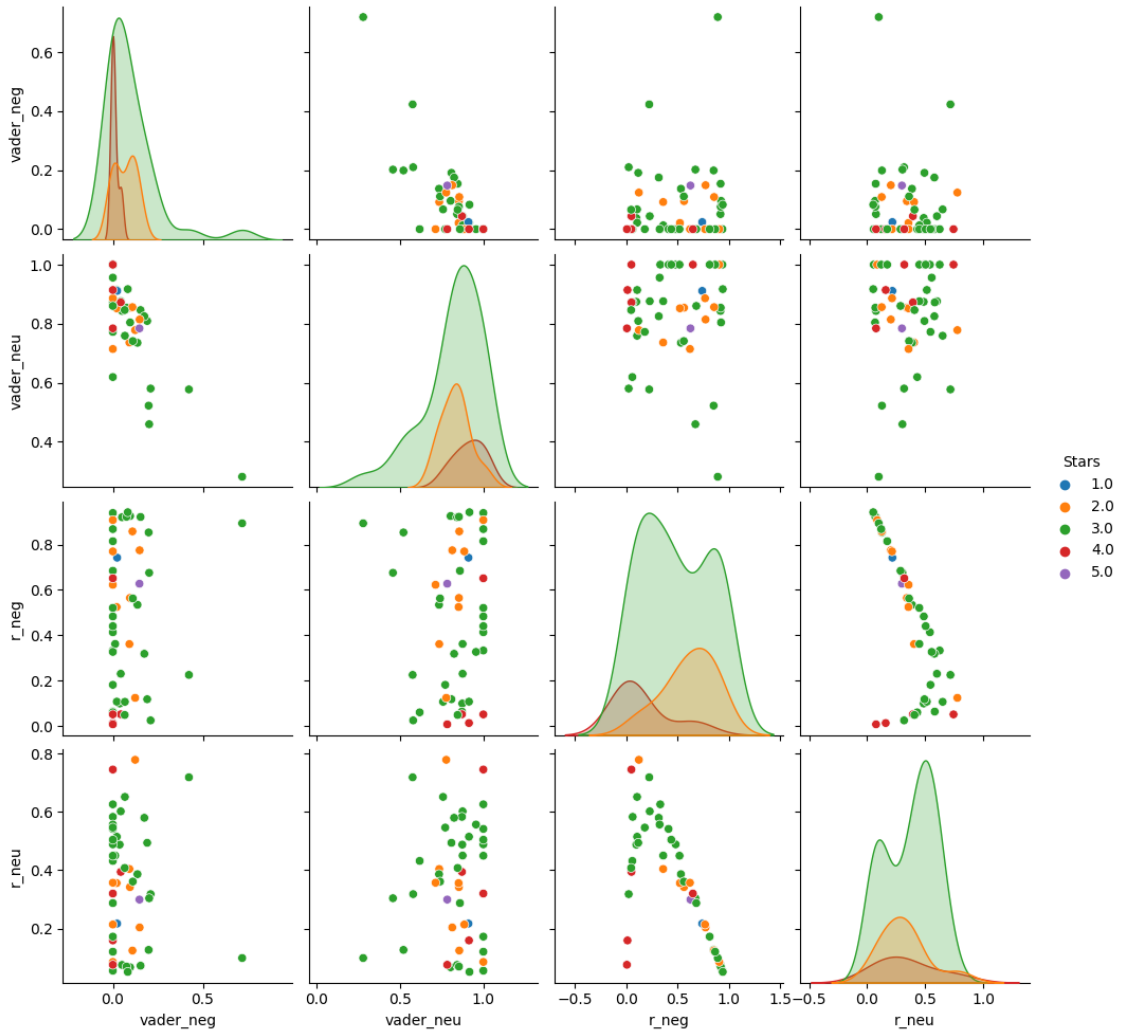


Fig 13: **Results of Lawn Washer reviews.**

The above graph is for the Lawn Washer reviews, which were mostly negative. Both these models perform similarly here, putting the scores concentrated around the neutral area (around 0.5).

Overall, both models were able to distinguish that PS5 had positive/neutral reviews, while Lawn Washer had negative/neutral reviews. RoBERTa was better at identifying the positive bias of the PS5 reviews than VADER.

Chapter 6

Lacking and Limitation

Lacking:

i) Several products reviewed. Analyzing reviews of more products will give a better indicator
of how the models are performing.

ii) The VADER model lacks in considering the context of the whole sentence, only taking the value of
each word without considering the rest of the sentence.

Iii) Token limitations. BERT has a token limit of 512 tokens, so longer passages will need to be split or
truncated, losing important context.

Limitations:

i) Unable to scrape Amazon webpages. Amazon has disabled the scraping of their web pages; therefore, the pages have to be saved locally to the computer first.

ii) Due to this limitation, scaling the project to scrape thousands of product reviews will be cumbersome.

Chapter 7

CONCLUSION

Our project consisted of two parts – web scraping reviews of two products from Amazon.com, and performing sentiment analysis on the reviews, using VADER and RoBERTa. The results showed both models performing well at identifying the overall sentiment of both products.

References

[1] Google Colab

[2] https://huggingface.co/docs/transformers/model_doc/roberta

[3] Visual Studio Code

[4] https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/

[5] https://www.geeksforgeeks.org/implementing-web-scraping-python-beautiful-soup/