

Decoding consumer voice : Sentiment analysis of web-scraped product reviews

S. Abhijith *

Anitha Poly †

Benjamin A. Jacob §

Gifty Roy ‡

Reshma Rachel Cherish ®

Department of Computer Science and Engineering

Christ (Deemed to be University)

Kengeri

Bangalore

Karnataka

India

Abstract

Nowadays online media shopping is more trending and extracting the customer's thoughts from their reviews is very useful for business analytics. So the researcher's are here trying to take a product from the online shopping platform Flipkart and apply live web scraping using a tool called BeautifulSoup. Then applied Natural language processing techniques to remove the emojis, regular expressions and stop words from the text. Also the researcher applies a stemming technique named snowballstemmer method for NLP purposes. Finally from the cleaned and extracted text, by applying vaderSentiment, the sentiment analysis is also done. Also by using the n-gram method in the end it is able to fetch most occurring words in the review.

Subject Classification: 68T01, 68Q55.

Keywords: Sentiment analysis, Product review, Natural language processing, Live web scraping, Stemming.

* E-mail: abhijith.s@mtech.christuniversity.in (Corresponding Author)

† E-mail: anitha.poly@mtech.christuniversity.in

§ E-mail: benjamin.jacob@mtech.christuniversity.in

‡ E-mail: gifty.roy@mtech.christuniversity.in

® E-mail: reshma.rachel@mtech.christuniversity.in

1. Introduction

Online shopping is where the customer can buy the product through the internet using a browser or mobile. Customers can find the product of interest by browsing the online store. It will provide all details of the product so that the customer can buy the product according to their needs. Online shopping can be business-to-consumer (B2C) or business-to-business (B2B) online shopping. In B2C, the products are directly sold to the customer without any influence of the other, whereas in B2B the products are sold from business firm to business firm.

From the review by the customer while purchasing products, the store can predict the quality and rating of their product on the market. Since most people are buying products online they give their views regarding the product. The analysis of customer reviews is an important feature to make the market a success. To read all the reviews, it will be a difficult task. This can be easily done by the method called **Sentiment Analysis**. Sentiment analysis is a perspective of finding the feeling of the customer toward a particular entity. It can predict the emotions, feelings, and attitudes of the customer toward a particular product. It can classify product reviews into good and bad. Thus, the market can know the needs of the customer and make them according to the needs of the customer.

Web scraping is a process of extracting information from websites. The information is stored in a database or spreadsheet for analysis. Scraping the information from the website involves fetching and extracting it. Fetching is downloading a page that the user views, it is done by the browser. For further processing of the data fetched we use Web Crawling, which is one of the components of web scraping.

Text preprocessing is a process of cleaning and preparing text data. It is an important step in NLP where it is used to remove unwanted texts such as punctuation, stop words. In this project a product from the online shopping platform Flipkart and apply live web scraping using a tool BeautifulSoup. Then applied Natural language processing techniques to remove the emojis, regular expressions and stop words from the text. Then applies a stemming technique named snowballstemmer method for NLP purpose. Finally from the cleaned and extracted text, by applying vaderSentiment, the sentiment analysis is also done. Also, by using the n-gram method in the end it is able to fetch most occurring words in the review.

2. Related work

The daily generation of data generated is immense as a result of the extensive utilization of the internet. One of the significant tasks in the world of unstructured data is to extract meaningful inferences. These inferences can be further utilized for making important decisions that affect the marketing of products and customer satisfaction.

Xing Fang and Justin Zhan [1] proposed a three-phase method for sentiment analysis on Amazon product reviews. They first extracted subjective sentences and performed POS tagging, then identified negative phrases and computed sentiment scores, finally categorizing sentiment polarity. Kaushal Kumar Bhagat et al. [2] analyzed sentiment on virtual learning during COVID-19, collecting and preprocessing 198 articles from search engines. Using dictionary-based and lexicon-based approaches, they determined article polarity, finding 92.2% positivity. Amjad Iqbal et al. [3] utilized deep learning to classify sentiments from various online sources, preprocessing data and employing RNN-based LSTM and Deep LSTM classifiers, measuring performance with accuracy, precision, recall, and F1 score metrics.

Anirban Adak et al. [4] reviewed ML, DL, and explainable AI techniques for sentiment analysis of food delivery reviews. They noted ML's widespread use, highlighting DL's interpretability limitations, which can be addressed with explainable AI. Saleh Naif Almuayqil et al. [5] proposed sentiment analysis on an imbalanced dataset from the Twitter US Airline Sentiment Dataset. They performed manual feature selection, text preprocessing, lemmatization, and used Random Undersampling to balance the dataset, employing ML models for sentiment prediction.

Weidong Huang et al. [6] introduced the EBLA model, combining word embedding and Bidirectional LSTM for sentiment analysis on JD.com product reviews. They optimized LSTM hidden layer weights using attention mechanisms and evaluated model performance in terms of precision, recall, and F1 score. Mohammed Jabreel et al. [7] proposed SentiRank, integrating online reviews and expert opinions for decision-making, using a restaurant dataset and the ELECTRE method for sentiment ranking.

Shivaprasad TK et al. [8] reviewed sentiment analysis on product reviews, covering data collection, preprocessing, text analytics, computational linguistics, and sentiment classification methods. Chhaya Chauhan et al. [9] explored algorithms for meaningful product review summaries, finding improved accuracy with unigrams and bigrams, along

with techniques like POS tagging, negation handling, and opinion classification.

These studies contribute to sentiment analysis advancements across various domains [11][12]. They address challenges such as handling different types of data, imbalanced datasets, interpretability of deep learning models, and enhancing accuracy through advanced techniques.

These studies collectively contribute to the advancement of sentiment analysis methodologies and their applications across diverse domains, including e-commerce, education, social media, and decision-making processes. Each approach offers insights into handling different types of data, addressing challenges such as imbalanced datasets, interpretability of deep learning models, and enhancing accuracy through advanced techniques.

3. Problem Statement Based on The Related Work

Manual analysis of review data from online shopping sites can be a challenging and time-consuming task. It can be difficult to manually analyze large datasets of review data, especially if the business is growing rapidly. Manual analysis can also be prone to errors, especially if the analyst is not familiar with the product or service being reviewed.

It can be difficult to interpret the results of manual analysis, especially if the dataset is large or complex. Fake reviews can be a major problem on online shopping sites, and it can be difficult to manually identify and remove them. Review data from online shopping sites can come in a variety of formats, including text, images, and videos. This can make it difficult to manually analyze the data in a consistent and efficient manner. Review data from online shopping sites often lacks context, such as the customer's demographics, purchase history, and reason for returning the product. This can make it difficult to understand the meaning of the reviews and draw meaningful insights from them.

4. Proposed Methodology

An online shopping platform, Flipkart, is used for this project, along with the web scraping tool BeautifulSoup. After that, natural language processing techniques are applied to remove emojis, regular expressions, and stop words from the text. Next, the snowball stemmer method is applied for NLP purposes. As a final step, the sentiment analysis is performed on the cleaned and extracted text by applying vaderSentiment.

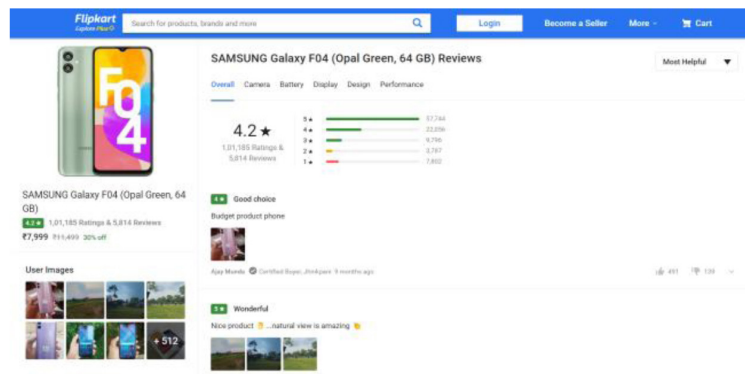


Figure 1
Flipkart data

In addition, by using the n-gram method, the most frequent words in the review were also found.

4.1 Data Set Used

Since the researcher is doing a live web scraping, review of a product named Samsung Galaxy F04 (which is a mobile phone) is taken from the online shopping site (Figure 1) and then the review is converted to dataset having 6 entities named rating, summary, review, location, date, upvotes and downvotes.[1]

4.2 Webscraping

Web scraping is extracting the live data automatically from a website which is Flipkart in this scenario. BeautifulSoup is the python library used here for web scraping. For doing this the developer has to use a header for faking the server of the online site. By using header it confuses the server that researcher is accessing through a program, and not by a program. If the server understands that the researcher is accessing the data through a programme, there is a possibility of blocking it by the server. This can be avoided using a header. Using a function find & find_all() we extract 6 entities named rating, summary, review, location, date, upvotes and downvotes from the classes from the webpage.

4.3 Natural Language Processing (NLP)

After extracting the needed information from the website, researchers convert and store it into a Dataframe to apply the natural language

	Review	Positive	Negative
0	tipe c batteri best call super nice good	0.775	0.0
1	basic android phone easili suit anyon normal ...	0.140	0.0
2	verri nice phone like phone verri best perform...	0.553	0.0
3	batteri amp charg time hrs watt charger ...	0.000	0.0
4	good	1.000	0.0
Neutral			
0		0.225	
1		0.860	
2		0.447	
3		1.000	
4		0.000	

Figure 3

Sentiment score for each reviews

4.3.4 Stemming

It is the process of reducing a word to its root form. It is a very important step to extract meaningful information while analyzing a huge volume of data.

4.4 Sentiment Analysis

It is the process of identifying the emotional tone behind a text. We use it to categorize reviews and opinions about a product or service into sentiments such as positive, negative or neutral.

Here the Sentiment Intensity Analyzer module from vaderSentiment package was used to perform the analysis. For the data used the result came to be neutral for most of the review words. The sentiment scores will be generated for each review as shown in Figure 3.

4.5 N-grams

N-gram methodology is a common approach used in sentiment analysis, which involves analyzing the occurrence of n-grams in text to understand sentiment and extract meaning. N-grams are contiguous sequences of n words or characters within a given text. By considering different values of n, such as unigrams (single words), bigrams (pairs of words), trigrams (three-word sequences), and so on, sentiment analysis models can capture different levels of context and meaning in the text.

In sentiment analysis, N-gram methodology is used to identify patterns and associations between words that may indicate sentiment or opinion. By examining the frequency and co-occurrence of specific n-grams in positive, negative, or neutral sentiment categories, machine learning algorithms or rule-based approaches can learn to classify new text samples based on their sentiment.

5. Results

In this work, the product reviews for a mobile phone from a prominent online e-commerce website, Flipkart has been subjected for analysis. Web scraping was done from the website and text processing was done using the Natural Language Toolkit (NLTK) python package. The stopwords were identified and removed from the text as part of the preprocessing step.

The following are the various graphical representations of the sentiment analysis tasks done for the web-scraped data.

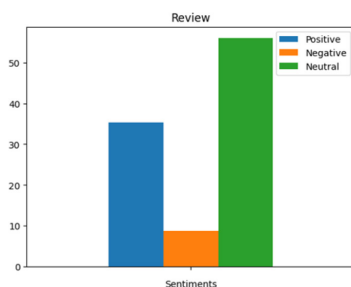


Figure 4

Categorization of Reviews

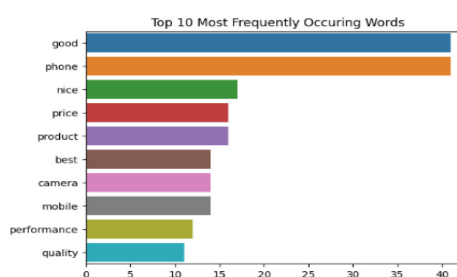


Figure 5

Most Frequently Occurring Words

Figure 4, categorizes the sentiments gathered or reviews generated from the data into three categories: positive, negative, and neutral. There were around 36 mentions of positive reviews; the negative reviews were around 8 mentions and, the neutral reviews were the highest with 55 mentions.

Figure 5, provides a list of the top 10 most frequently occurring words. From the graph it can be observed that the words namely good, phone, nice, price, product, best, camera, mobile, performance, and quality are very commonly used by product reviewers. The numbers at the bottom provide insights into the popularity and importance of these words in the context they are being analyzed. Some of the noticeable observations are that the word “good” is most commonly used by reviewers. The other nine words are of comparatively lesser frequency.

Figure 6, provides a list of the most frequently occurring bigrams in the product reviews. A context of the definition of bigrams is provided under the Sentiment Analysis section of this paper. Bigrams include phrases like ‘value money’, ‘nice product’, ‘nice phone’, ‘don’t buy’, ‘worst mobile’ and so on. The article also mentions specific terms like ‘Samsung

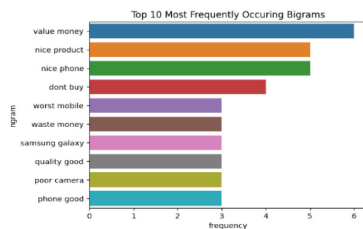


Figure 6
Most Frequently Occurring Bigrams

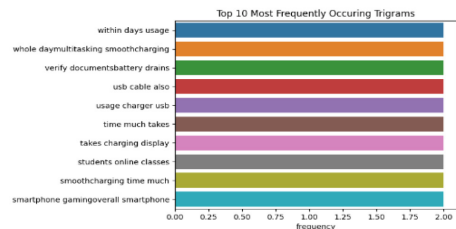


Figure 7
Most Frequently Occurring Trigrams

galaxy’, ‘quality good’, and ‘poor camera’. The frequency of these bigrams is indicated by numbers ranging from 0 to 6.

Figure 7, provides a list of the most frequently occurring trigrams in the product reviews. Here, the trigrams include phrases like ‘within days usage’, ‘usb cable also’, ‘time much takes’, ‘takes charging display’, ‘whole day multitasking smooth charging’, and so on.

6. Conclusion and Future Scope

In conclusion, leveraging web scraping for sentiment analysis offers promising avenues for understanding public opinion from online sources. Future research could explore enhancing model robustness through machine learning techniques and applying tools like selenium for automating the process of web scraping [13]. Also one can extend the scope to analyze the sentiment across various languages and domains, thereby advancing applications in market research, brand monitoring and social media analysis.

References

[1] Fang, Xing, and Justin Zhan. “Sentiment Analysis Using Product Review Data.” *Journal of Big Data*, vol.2, no. 1, 16 June (2015), journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2,<https://doi.org/10.1186/s40537-015-0015-2>.

[2] Bhagat, Kaushal Kumar, et al. “Public Opinions about Online Learning during COVID-19: A Sentiment Analysis Approach.” *Sustainabil-*

- ity, vol. 13, no. 6, 18 Mar. (2021), p. 3346, <https://doi.org/10.3390/su13063346>.
- [3] Iqbal, Amjad, et al. "Sentiment Analysis of Consumer Reviews Using Deep Learning." *Sustainability*, vol. 14, no. 17, 31 Aug. 2022, p. 10844, <https://doi.org/10.3390/su141710844>. Accessed 1 Sept. (2022).
- [4] Adak, Anirban, et al. "Sentiment Analysis of Customer Reviews of Food Delivery Services Using Deep Learning and Explainable Artificial Intelligence: Systematic Review." *Foods*, vol. 11, no. 10, 21 May 2022, p. 1500, <https://doi.org/10.3390/foods11101500>. Accessed 5 Aug. (2022).
- [5] Almuayqil, Saleh Naif, et al. "Enhancing Sentiment Analysis via Random Majority Under-Sampling with Reduced Time Complexity for Classifying Tweet Reviews." *Electronics*, vol. 11, no. 21, 6 Nov. (2022), p. 3624, <https://doi.org/10.3390/electronics11213624>.
- [6] Huang, Weidong, et al. "Sentiment Analysis of Chinese E-Commerce Product Reviews Using ERNIE Word Embedding and Attention Mechanism." *Applied Sciences*, vol. 12, no. 14, p. 7182, 16 July (2022). <https://doi.org/10.3390/app12147182>.
- [7] Jabreel, Mohammed, et al. "Introducing Sentiment Analysis of Textual Reviews in a Multi-Criteria Decision Aid System." *Applied Sciences*, vol. 11, no. 1, p. 216 28 Dec. (2020), <https://doi.org/10.3390/app11010216>. Accessed 19 Nov. 2022.
- [8] Shivaprasad, T. K., and Jyothi Shetty. "Sentiment Analysis of Product Reviews: A Review." 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), Mar. (2017), <https://doi.org/10.1109/icicct.2017.7975207>. Accessed 26 May 2021.
- [9] Chauhan, Chhaya, and Smriti Sehgal. "Sentiment Analysis on Product Reviews." *IEEE Xplore*, 1 May (2017), ieeexplore.ieee.org/document/8229825.
- [10] Poonia, Ramesh C., et al., eds. *Smart Farming Technologies for Sustainable Agricultural Development*. IGI Global (2018).

- [11] Ahuja, Shefali, and R K Tailor. "Application of Robotic Process Automation in Customer Satisfaction at Toll Plazas." *Journal of Information and Optimization Sciences*, vol. 44, no. 8, pp. 1685–1694, 1 Jan. (2024), <https://doi.org/10.47974/jios-1484>. Accessed 4 Mar. 2024.
- [12] Chhikara, Monika, and Sanjay Kumar Malik. "A Recommendation System for Online Social Semantic Network Using Knowledge Based, Content Based and Collaborative Filtering." *Journal of Information and Optimization Sciences*, vol. 44, no. 4, pp. 795–806, 1 Jan. (2023). <https://doi.org/10.47974/jios-1356>. Accessed 4 Mar. 2024.
- [13] Jain, S., Kumar, S., Sharma, V. K., & Poonia, R. C. Peregrine preying pattern based differential evolution for robot path planning. *Journal of Interdisciplinary Mathematics*, 23(2), 555–562 (2020). <https://doi.org/10.1080/09720502.2020.1731974>.