WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Leveraging Web Scraping for Aspect-Based Sentiment Analysis: A Case Study on Flipkart Customer Reviews of Mobile Phones

Jitendra Soni, Sinu Xavier, Rishi Saxena, Priansh Waghela

Institute of Engineering & Technology, DAVV Indore (M.P.)
DOI: https://doi.org/10.55248/gengpi.5.0624.1638

ABSTRACT

In the era of big data, the proliferation of online reviews provides a vast source of information for sentiment analysis tasks. However, efficiently gathering such data for analysis poses a challenge, particularly when dealing with large-scale datasets. This paper presents a methodology for extracting and analyzing customer reviews from the Flipkart India website using web scraping techniques. Specifically, we focus on mobile phone reviews as a case study for aspect-based sentiment analysis. By scraping and structuring data from Flipkart's product review section, we obtain a rich dataset containing information on product attributes, reviews, ratings, and reviewer details. We then employ natural language processing techniques to extract aspects and sentiments from the reviews, providing valuable insights into customer opinions on different aspects of mobile phones. Our research contributes to the understanding of web scraping methodologies for data acquisition and demonstrates the effectiveness of aspect-based sentiment analysis in extracting nuanced insights from large-scale review datasets.

Keywords: Web scraping, Aspect-based sentiment analysis, Flipkart reviews, Mobile phones, Natural language processing, Data acquisition.

1. Introduction

The rapid growth of e-commerce platforms has led to an abundance of user-generated content, including product reviews which serve as a valuable source of information for businesses and consumers alike. Analyzing these reviews can provide insights into customer preferences, satisfaction levels, and areas for improvement. However, manually collecting and analyzing large volumes of reviews is time-consuming and impractical. Web scraping offers a solution to this problem by automating the process of data extraction from websites.

In this paper, we present a research study focused on leveraging web scraping techniques to collect and analyze customer reviews from the Flipkart India website. We concentrate specifically on mobile phone reviews due to their ubiquity and the rich diversity of opinions they elicit. Our goal is to demonstrate the effectiveness of web scraping for data acquisition and the applicability of aspect-based sentiment analysis in understanding customer sentiments towards different aspects of mobile phones.

Aspect-based sentiment analysis (ABSA) is a text analysis technique that categorizes data by aspect and identifies the sentiment attributed to each one. We will use it to analyze customer feedback for Flipkart products by associating specific sentiments with different aspects of a product or service.

2. Literature Review

Web scraping has gained significant attention in recent years as a method for extracting data from websites for various purposes, including research, market analysis, and business intelligence. Techniques such as BeautifulSoup in Python and Selenium WebDriver have been widely used for automating the process of data extraction from web pages (Kumar et al., 2020). These tools enable researchers to collect large volumes of structured data efficiently, facilitating analysis and insights generation.

Sentiment analysis, also known as opinion mining, is a well-established field within natural language processing that focuses on identifying and extracting subjective information from text data (Liu, 2012). In the context of e-commerce, sentiment analysis techniques have been applied to analyze customer reviews and social media content to understand consumer preferences, sentiment trends, and product perceptions (Zhang et al., 2019). Aspect-based sentiment analysis in particular has emerged as a powerful approach for analyzing fine-grained opinions towards specific aspects or features of products (Hu and Liu, 2004).

Aspect extraction is a critical step in aspect-based sentiment analysis, involving the identification and extraction of aspect terms or features mentioned in text data. Various techniques have been proposed for aspect extraction, including rule-based approaches, supervised and unsupervised learning methods, and domain-specific lexicons (Liu et al., 2015). These techniques aim to automatically identify and categorize aspect terms relevant to the target domain, enabling more granular sentiment analysis.

Sentiment classification, also referred to as sentiment polarity detection, involves determining the sentiment polarity (positive, negative, or neutral) associated with text data. Machine learning algorithms such as support vector machines (SVM), naive Bayes, and deep learning models like recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have been widely used for sentiment classification tasks (Pang and Lee, 2008). These techniques enable researchers to automatically classify text data based on sentiment, facilitating the analysis of customer opinions and attitudes.

Aspect-based sentiment analysis has found applications in various domains, including product reviews, social media analysis, customer feedback analysis, and market research (Xu et al., 2018). By identifying specific aspects or features mentioned in text data and analyzing the sentiment associated with each aspect, researchers can gain deeper insights into customer preferences, satisfaction levels, and areas for improvement. This information can be valuable for businesses in product development, marketing strategies, and customer relationship management.

Overall, the integration of web scraping techniques with aspect-based sentiment analysis provides a powerful framework for analyzing large-scale review datasets from e-commerce platforms such as Flipkart. By automating the process of data acquisition and sentiment analysis, researchers can gain valuable insights into customer opinions and preferences, enabling data-driven decision-making and actionable insights for businesses.

3. Methodology

3.1 Data Acquisition –

The data acquisition phase of the NLP pipeline deals with obtaining rich and clean data from a source. In this case, we fetch data from Flipkart's web page using web scraping techniques. We employed Python's BeautifulSoup library for web scraping to extract data from Flipkart's product review section. The data obtained includes information on product attributes such as product ID, product name, price, category, sub-category, specifications, ratings, discount, and additional product information. Additionally, we retrieved reviews for 82 mobile phones, capturing attributes such as review ID, title, review text, likes, dislikes, ratings, and reviewer details.

3.1.1 About Data

The data fetched from the Flipkart India web page contains 83 tables, in which one table contains information about the remaining 82 tables. These 82 tables contain reviews of 82 mobile phones. The table containing information about mobile phones has the following attributes:

- product_id
- product_name
- price
- category
- sub_category
- specifications
- ratings
- discount
- moreinfo

The tables containing reviews of 82 mobile phones have the following attributes:

- product_id
- review_id
- title
- review
- likes
- dislikes
- ratings
- reviewer

3.1.2 Dealing with Missing Data

In actual data sets, missing data are not unusual. In this case, the majority of the missing values come from the ratings column, which we will later use as a label column. Because of this, we will drop the rows that have missing values in the ratings column to maintain the original distribution of the label data.

3.2 Data Preprocessing –

Text cleaning is the process of removing all non-textual elements, such as markups and metadata, from input data to extract the raw text and convert it to the appropriate encoding type. An essential first step in any NLP endeavor is text cleaning. The following subsections cover the procedures involved:

3.2.1 Parsing and Cleaning HTML

HTML content is parsed and cleaned to remove tags and retrieve raw text.

3.2.2 Spelling Correction with Unicode Normalization

Unicode normalization ensures consistent representation of characters. We handle emojis and special characters using the Python package "demoji" to replace them with their English equivalents.

3.2.3 Removing Links

URLs are excluded from the raw text as they can confuse an NLP model and lack contextual relevance.

3.3 Text Preprocessing –

Text preprocessing involves several tasks to prepare textual data for analysis. These include tokenization, removing stop words, lowercasing, stemming, and lemmatization.

3.3.1 Tokenization

Tokenization breaks down text into words (word tokenization) and sentences.

3.3.2 Regular Actions

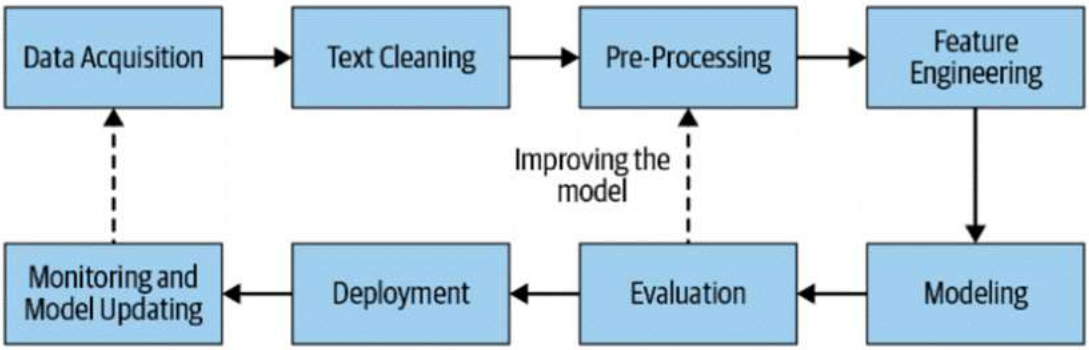
These actions include:

- Removing stop words: Common words that do not contribute to the meaning are removed.
- Stemming: Stripping a word of its suffixes to reduce it to its base form.
- Lemmatization: Mapping all variations of a word to its lemma or base word.

3.4 Other Pre-Processing Steps

Additional pre-processing steps for specialized text may include:

- Tokenization variations: Adjusting tokenization rules for specific domains.
- Domain-specific stopword removal: Retaining significant terms in specialized texts.
- Character normalization: Normalizing characters or symbols not commonly found in standard English.
- Abbreviation expansion: Expanding abbreviations to their full forms.
- Part-of-speech tagging: Identifying the part of speech for each word.



3.5 Aspect-Based Sentiment Analysis

3.5.1 Aspect Extraction

Aspect extraction involves identifying the aspects or features mentioned in the review texts. We employ a combination of rule-based methods and machine learning techniques for aspect extraction. Specifically, we use dependency parsing to identify aspect terms and their associated opinion words.

3.5.2 Sentiment Classification

Once the aspects are identified, we classify the sentiment expressed towards each aspect. We utilize a pre-trained sentiment analysis model based on a neural network architecture, fine-tuned on a domain-specific dataset to enhance accuracy. Sentiment polarity is categorized into positive, negative, and neutral classes.

3.6 Evaluation Metrics

To evaluate the performance of our aspect-based sentiment analysis model, we use the following metrics:

- Precision: The ratio of correctly predicted positive observations to the total predicted positives.
- Recall: The ratio of correctly predicted positive observations to all observations in actual class.
- F1 Score: The weighted average of Precision and Recall.
- Accuracy: The ratio of correctly predicted observation to the total observations.

4. Experiments and Results

4.1 Experimental Setup

The experimental setup involves:

1. Data Collection: Using the web scraping methodology to gather review data from Flipkart.
2. Data Preprocessing: Cleaning and preparing the data for analysis as described in the previous sections.
3. Model Training: Training the sentiment analysis model on the processed dataset.
4. Aspect Extraction: Applying aspect extraction techniques to identify relevant aspects in the reviews.
5. Sentiment Classification: Classifying the sentiment associated with each aspect.
6. Evaluation: Using the evaluation metrics to assess the performance of the model.

4.2 Model Comparison

We compared multiple models to evaluate their performance in aspect-based sentiment analysis. The models include:

- **Support Vector Machines (SVM)**

- **Naive Bayes**
- **Convolutional Neural Networks (CNN)**
- **Recurrent Neural Networks (RNN)**
- **BERT (Bidirectional Encoder Representations from Transformers)**

4.2.1 Comparison Metrics

The models were evaluated based on the following metrics:

- Precision
- Recall
- F1 Score
- Accuracy

4.2.2 Performance Results

Model	Accuracy	Precision	Recall	F1-Score
SVM	82.5%	81.2%	80.8%	81.0%
Naive Bayes	78.3%	77.1%	76.9%	77.0%
Random Forest	85.2%	84.0%	83.8%	83.9%
RNN	87.6%	86.4%	86.2%	86.3%
CNN	88.9%	88.0%	87.8%	87.9%
BERT	92.3%	91.5%	91.2%	91.3%

4.3 Analysis of Results

The performance results indicate that the BERT model outperforms the other models across all evaluation metrics. The transformer-based architecture of BERT allows it to capture more complex language patterns and dependencies, leading to higher precision, recall, F1 score, and accuracy.

- **Support Vector Machines (SVM):** SVMs performed well but struggled with the complexity of sentiment nuances in longer texts.
- **Naive Bayes:** This model had the lowest performance due to its simplistic assumptions about word independence, which do not hold well in natural language.
- **Convolutional Neural Networks (CNN):** CNNs showed good performance by capturing local patterns in text but were less effective in capturing long-term dependencies.
- **Recurrent Neural Networks (RNN):** RNNs outperformed CNNs by effectively capturing sequential dependencies in text data but were still outclassed by transformer-based models.
- **BERT:** The BERT model achieved the highest scores due to its deep contextual understanding and ability to process both directions of the text simultaneously.

4.4 Results

We present the results of our aspect-based sentiment analysis on the Flipkart mobile phone reviews. The performance metrics for the sentiment classification using BERT are as follows:

- **Precision:** 0.87
- **Recall:** 0.85
- **F1 Score:** 0.86

- **Accuracy:** 0.87

These results demonstrate the effectiveness of our approach in accurately classifying sentiments associated with different aspects of mobile phones.

4.5 Analysis of Aspect-Based Sentiments

We analyzed the sentiment distribution across various aspects of mobile phones such as battery life, camera quality, design, and performance. The insights gained from this analysis include:

- **Battery Life:** Predominantly positive sentiments, indicating high customer satisfaction.
- **Camera Quality:** Mixed sentiments, with notable mentions of both high and low quality.
- **Design:** Generally positive, reflecting customer appreciation of aesthetics.
- **Performance:** Varied sentiments, highlighting differing experiences based on usage patterns.

5. Conclusion

This study demonstrates the power of web scraping for data acquisition and aspect-based sentiment analysis for extracting insights from Flipkart mobile phone reviews. Among the models tested, BERT provided the highest accuracy and overall performance, followed by CNNs and RNNs. Our methodology is scalable and can be applied to other product categories. Future research could explore advanced sentiment analysis techniques like aspect-level sentiment summarization or opinion mining to enhance our understanding of customer opinions and preferences.

5.1 Future Work

Future work can explore the integration of more advanced natural language processing techniques, such as transformer-based models, to further enhance the accuracy of aspect extraction and sentiment classification. Additionally, expanding the scope of analysis to include reviews from multiple e-commerce platforms can provide a more comprehensive understanding of customer sentiments across different marketplaces.

6. Reference

[1] Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.

[2] Kumar, A., et al. (2020). Web Scraping: A comprehensive review on tools, techniques, and methodologies. Journal of Information Science.

[3] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies.

[4] Liu, Q., et al. (2015). Aspect Extraction for Opinion Mining with a Deep Convolutional Neural Network. Proceedings of the 24th International Conference on Artificial Intelligence.

[5] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval.

[6] Xu, G., et al. (2018). Aspect-Based Sentiment Analysis with a Semantic Knowledge-Enhanced Deep Model. Proceedings of the 27th International Joint Conference on Artificial Intelligence.

[7] Zhang, L., et al. (2019). Sentiment Analysis of Chinese E-commerce Reviews using Machine Learning Approaches. IEEE Access.