**CUSTOMER PRODUCT SEARCH AID USING TEXT ANALYTICS AND WEB SCRAPING**

**By**

**MOHAN RAJENDRAN**

A Capstone Project Paper Submitted in Partial Fulfillment of the

Requirement for the Degree of

Master of Science

In

Data Science

University of Wisconsin – La Crosse

La Crosse, Wisconsin

August 2020

**ABSTRACT**

Customer Product Search Aid Using Text Analytics and Web Scraping

Mohan Rajendran

Capstone project: Case Study

Master of Science in Data Science

University of Wisconsin – La Crosse

La Crosse, Wisconsin

Summer, 2020

This project aims to develop a consolidated product review analytics dashboard, which will help consumers make an informed decision. Reviews are opinions expressed by the current owners of the products and the experts in the industry on the product features. These online reviews replace the traditional Word of Mouth recommendations and are termed as eWOMs. The consumption of these eWOMs is significantly impacted by the age, technical knowledge, and time availability of the user. The main objective of this project is to consolidate these eWOMs from various platforms like Expert Review sites (TechRadar) and social media platforms like Amazon and Twitter using Web Scraping techniques and platform provided APIs. The secondary objective is to develop a dashboard that can provide the output to the end-user and design a data collection strategy that will enable the tool to be used for different products. Two options of data collection are discussed, and 'on the fly' data collection method is considered for this project owing to the dynamic nature of the query, which will allow us to search for more products. Sentiment analysis is performed on the extracted reviews, and aggregated sentiment scores are computed using the python TextBlob package. These aggregated sentiment scores lessen the effects of effectively worded fake reviews and spams, which are more pronounced when users manually scan the reviews. Considering this is the first level of development, next steps like automatic spam detection, emoji detection are proposed for future development. Limitations such as restrictions imposed by social media platforms and Amazon are discussed, and possible solutions are outlined.

Table of Contents

[CHAPTER 1 1](#_Toc47556104)

[INTRODUCTION 1](#_Toc47556105)

[Project Background 1](#_Toc47556106)

[Objectives 3](#_Toc47556107)

[Procedure 4](#_Toc47556108)

[Significance and Limitations 5](#_Toc47556109)

[Final Document Organization 5](#_Toc47556110)

[CHAPTER 2 7](#_Toc47556111)

[LITERATURE REVIEW 7](#_Toc47556112)

[History 7](#_Toc47556113)

[Amazon 8](#_Toc47556114)

[Social Media 9](#_Toc47556115)

[Current Scenario 9](#_Toc47556116)

[CHAPTER 3 14](#_Toc47556117)

[DATA COLLECTION AND ANALYSIS 14](#_Toc47556118)

[Data Collection Strategy 14](#_Toc47556119)

[Primary Data 15](#_Toc47556120)

[Data Extractor Modules 16](#_Toc47556121)

[Expert Review Data Module 17](#_Toc47556122)

[Amazon Review Data Module 18](#_Toc47556123)

[Twitter Review Data Module 18](#_Toc47556124)

[Text Analytics Module 19](#_Toc47556125)

[CHAPTER 4 22](#_Toc47556126)

[RESULTS 22](#_Toc47556127)

[CHAPTER 5 24](#_Toc47556128)

[CONCLUSION 24](#_Toc47556129)

[Recommendations/Next Steps 25](#_Toc47556130)

[References 27](#_Toc47556131)

**Table of Figures**

Fig 1: Expert Review Word Cloud ............................................................................................ 22

Fig 2: Expert Review Sentiment Distribution ............................................................................ 23

Fig 3: Amazon Review Word Cloud .......................................................................................... 24

Fig 4: Amazon Review Sentiment Distribution ........................................................................ 24

# CHAPTER 1

# INTRODUCTION

## Project Background

Google search currently forms an integral part of all the aspects of people's life. Right from research to daily mundane tasks like grocery shopping involves google search in one way or the other. Anjana S.S. (2018) explains a customer purchase includes the following steps: Recognition of problem, Search for information, Evaluation of alternatives, Purchase decision, and Outcome. In the current social media world, typical consumer purchase involves searching the various options available in that category, going through their descriptions to understand their features and options, go through review portals, do a search amazon, twitter, and other social media platforms for the comments and feedback provided by current users and finally decide on whether to make the purchase or not.

X. Li, L.M. Hitt states that online reviews not only reflect product quality but also affect perceived product value, which is the difference between quality and price. When customers come across two similar products with similar pricing, they often depend on reviews to come up with a decision. When there is a more significant number of positive reviews, the perceived value of the product increases. When the perceived value increases, the customer is able to justify and trust the price quoted.

Customer involvement with the product also determines how likely they are to purchase the product. Ying-Ping Liang concludes through the regression analysis that the higher the consumer product involvement, the higher product knowledge, and impulse buying behavior. Arezou Ghiassaleh, Bruno Kocher, Sandor Czellarc states that products that have characteristics of experience goods or credence goods tend to be purchased via mobile and Internet channels that provide detailed and additive information like consumer reviews.

On the other end, product makers do an exactly similar search to understand customer pulse and gather their feedback and complaints. Ramanathan, U., Subramanian, N. and Parrott, G. (2017) states that "combining social media reviews, marketing, and operations, businesses will be better placed to survive in the ultra-competitive social media-influenced era." This kind of analysis helps them not only to improve the product in subsequent release and updates but also to understand customer satisfaction so they can come up with effective marketing strategies and promotions.

Another important aspect of online reviews is its ability to bring customer trust to products. Zhu, F., Zhang, X. study on video gaming industry sales indicates that online reviews are more useful and impactful for less popular games. This ability to gain customer trust proves to be a helpful aid for new products and less popular ones. Also, Yan Li, Ruijuan Wu, Dongjin Li, Hejun Fan states that "a product or service that has a high level of scarcity usually has a higher conversational value, and this makes consumers more willing to discuss it, and ultimately increases the WOM willingness of consumers in social networks."

So, we can see the reviews, comments, feedback, complaints collectively form the Voice of Customer, which is valuable information not only for the product makers but also for the prospective buyers. Mudambi, S., & Schuff, D. (2010) states that "online customer reviews can be defined as peer-generated product evaluations posted on the company or third-party websites." Customers use these evaluations to get more information about the product and also use them to evaluate alternative products available.

Though product makers gather this information from different sources and refine them as per their needs, there is no unifying solution for prospective buyers/consumers in general. Often consumers end up going to various portals and platforms to go through the information on their own.

So, in this paper, our focus is to create such an analytics solution for consumers so they can get all this info in a single place. This project is also an effort to make data science cater to ordinary people in everyday daily tasks. An attempt is made to bring an executive summary dashboard-style reporting to the general public's daily needs. The dashboard will provide the user with sentiment analysis about the product. When stating about the goal of sentiment analysis Olivier Kraaijevel, Johannes De Smedt says that "the main goal is to assign a positive, negative, or neutral sentiment polarity score to unstructured text." We will subscribe to this definition and differentiate and display the polarity of the reviews along with a word cloud based on the text of the reviews.

## Objectives

The primary purpose of this project is to develop a data science product to aid consumers who look to make a retail purchase, to make an informed decision. Though much of this information is available on the internet, they are scattered around different sources, and this project tries to bring an executive summary like a dashboard for general public consumption. For the scope of the project, we will restrict the analysis to 2-5 broad product categories (such as for example, Camera, Laptop). The output will encompass word clouds that make the product features stand out and charts depicting the sentiment scores.

The secondary objective is to build an efficient text analytics tool that can be utilized in similar scenarios with minimal changes. This objective is needed since there is a potential to use this for far more products than the ones being analyzed in this paper. We will also try to quantify the quality scores so the users can easily compare different options.

The study will try to answer the following questions

1. What are the main features of the product?
2. What is the aggregated sentiment score of the reviews?
3. How do the products compare against one another in the same category?
4. What is the voice of the customer of current users of the product?
5. What is an efficient way to make this study customized to different products and scenarios?

## Procedure

This study is based on text analytics and will be conducted by following the below steps

* Decide on the number of product categories and their details to target for analysis. For now, we will set this number to be between 2-5 depending on the data availability and data volume.
* Gather data from different sources, likes review pages, amazon comments, and other sources for these products. These will be done using APIs provided by the platform or through web scraping technologies using python.
* Perform sentiment analysis using one of the python packages
* Provide aggregated word cloud of reviews and recommend alternatives if any
* An attempt will be made to quantify the quality sentiment scores so the users can easily compare different options.
* Finally, a data science executive summary like a dashboard will be made available to the user using reporting options available within python or using visualization tools like Tableau.

## Significance and Limitations

The study is significant in terms of its usefulness to both the prospective buyers and product makers. This kind of product will significantly reduce the research time for the consumers. Due to its consolidated nature, it will also reduce the various issues related to reviews like fake reviews, spams, and under-reporting.

One of the significant limitations is in data gathering. For this study, we are utilizing the APIs and data gathering procedures without paying for the same. Tech companies have set limits when it comes to free use of their social media feeds. This significantly reduces the number of reviews being collected and analyzed.

## Final Document Organization

The final document will be a detailed report on the entirety of the project. It will include:

* An abstract about the project
* A brief introduction detailing the project scope, purpose, and description.
* Details on Data collection methods with details on why a specific data collection method is chosen.
* Data cleansing process
* Text analytics performed and their procedure
* A summary of the findings with necessary visualizations and insights
* Recommendations on next steps on how to enhance the product further and discuss any roadblocks encountered
* Conclusion

The report will also include all the references used and other supporting information related to the project.

# CHAPTER 2

# LITERATURE REVIEW

## History

In the early days around until the 1990s, shopping used to be more of an in-person experience. People often visit the shopping hubs/malls or retail stores in person, try the product sample, interact with the salesperson, negotiate on price, and make a purchase. In this model, retailers had the upper hand in their local market, and salespersons are the critical piece in making or breaking a sale. It is not uncommon to find monopolies in local markets, which made customers go to the same store despite past bad experiences.

In this historical model, customer options to understand the product are limited. This limitation not only an issue for the customer but also for the shop owners. Shop owners rely on advertisements and marketing to reach their customers. Both shop owners and customers rely significantly on Word of Mouth reviews, Personal Recommendations, Product Trails, and Product Demos to learn about products. Though each of these has significant success in terms of conveying both positive and negative reviews, their reach is often limited.

Around the same time, there is an electronic revolution happening that made affordable home computing possible. B. Bickart, R. Schindler suggests, "online communities that offer consumers the ability to exchange product information and product experiences directly and to develop relationships with others sharing similar interests may have the potential to generate product interest in large numbers of people effectively." This affordable home computing paved the way to some of the initial online review platforms like rateitall.com, deja.com, and Epinions.

The problem with these forums is that they are not trustworthy. T. Hennig-Thurau, K.P. Gwinner, G. Walsh, D.D. Gremler argues that "consumers desire for social interaction, desire for economic incentives, their concern for other consumers, and the potential to enhance their self-worth are the primary factors leading to eWOM behavior." More and more product owners started to encourage people to leave reviews for economic incentives. Often competitors also used paid services to tarnish other brands or promote friendly businesses. This action by the product owners made people lose trust in the reviews. Once the trust is lost, the reviews lost their significance among the customers. Most of these websites are defunct now.

This issue highlights how the expectation is different between customers and companies when it comes to reviews. This difference is vital for the review sites and retailers to make sure they have a satisfied customer. For example, a negative review is not favored by the product companies while it is beneficial for customers and the review sites/retailers. E. Maslowska, E.C. Malthouse, V. Viswanathan states that "it is in the retailer's interest to have an accurate and unbiased review ecosystem. Any efforts to manipulate the review system may damage its credibility and drive customers to a retailer that has unbiased reviews."

## Amazon

Then came Amazon in the late nineties. Previous review platforms just collected and displayed reviews. However, in the case of Amazon, people buy products and leave reviews about the products on the same platform. The reviews left are mostly from ordinary customers who purchased the products. So, these reviews are from people who personally had their hands-on with the product and reviewed after personally using them. These reviews become the new Word of Mouth and Personal Recommendations. R. Iyer, M. Griffin states that "in the face to face WOM, an opportunity exists for the source to provide both factual information as well as sharing their personal experiences. Comparatively, online format WOM recommendations have a lower ability to provide statistical evidence." They depend more on personal use and thereby have higher trust value.

This new age of online customers slowly adapted online reviews and gained better confidence compared to the traditional word of mouth feedback. These eWOM has far reach compared to traditional WOMs. The ease at which users can consume the reviews on the same page where they intend to purchase enabled the easy adoption of these reviews.

## Social Media

Almost at the same time, social media and blogging started taking over the internet. They opened new avenues where customers can let out their feedback and thoughts on products. The informal and personal nature of these platforms made it easy for consumers to provide reviews as well as consume reviews provided by others.

## Current Scenario

Now online platforms are filled with reviews and product details. Companies often mine these rich data sources to understand their product reach and complaints. Amazon made it easy to consume these reviews by providing them in the same place where they list the products, but it is not the case with social media reviews. Despite their difficulty to use, social media reviews provide rich Voice of Customer reviews, which is not restricted to one platform. So, it becomes one of the essential data sources for reviews. In short, customers tend to believe other customers more than any other form of product knowledge.

Khalid Saleh compiled an infographic on the importance of online customer reviews. He finds that this new age customer, on average, read about 2-10 reviews before making a purchase. More than 92% of customers read some form of online reviews before making any purchase decision. Furthermore, more than 88% of them trust these reviews in the same line as personal recommendations. It implies that customers spend a significant amount of time reading online reviews before deciding on a purchase. The process usually forms a pattern where the customer performs an internet search or social media search on the reviews related to the product he is interested in and digests the information. This will provide them with different product exposure, which will sometimes end up exploring an alternate brand or product. Though some of them are paid up reviews, platforms make it easy to distinguish them from unpaid regular customer reviews.

Though each platform makes it easy for the customer to consume the information, there is no cross-platform collaboration. However, this is not an issue for the younger generation. They are able to understand each platform and utilize or provide these reviews quickly. It is often a tough task for the older generation and those who are not familiar with social media platforms.

Another critical aspect of these reviews is how negative reviews have more impact than positive reviews. Jason Q.Zhang, Georgiana Craciuna, Dongwoo Shin states that "consumers do not give equal weights to positive and negative product reviews." N. Purnawirawan, M. Eisend, P. De Pelsmacker, N. Dens (2015) argues that "negative reviews have the most impact on attitudes and usefulness for the end customer." Also, J. Lee, D.h. Park, I. Han concludes that "high-quality negative online consumer reviews influence consumer attitudes more than low-quality negative online consumer reviews." F.L. Weisstein, L. Song, P. Andersen, Y. Zhu stated in their experimental research concluded that "participants placed greater weight on online reviews when the majority of the comments are negative than when the number of negative remarks is negligible."

Apart from eCommerce companies, this issue with negative reviews is relevant to other industries as well. Researching hotel booking reviews Beverley A. Sparks, Victoria Browning, states that "consumers seem to be more influenced by early negative information, especially when the overall set of reviews is negative." However, positively framed information, together with numerical rating details, increases both booking intentions and consumer trust. Cesare Amatulli, Matteo De Angelis, Giovanni Pino, Gianluigi Guido states that "consumers are highly sensitive to negative information diffused by third-party information sources such as NGOs." H. Hong, D. Xu, G.A. Wang, W. Fan states that "while altruism, enjoyment, and age factors show up as factors that stimulate the writing of negative eWOM just as they do for positive eWOM, attachment to the community is not a factor."

This affinity towards negative reviews is more pronounced in the older demographic than the younger demographic. B. von Helversen, K. Abramczuk, W. Kopeć, R. Nielek concludes that "younger demographics tend to lean towards higher-rated products based on average consumer ratings." In comparison, for older demographics, negative reviews play a significant role in deciding on a product purchase.

This aspect of negative review makes it hard to be ignored by the product owners since it is not just regular reviewers who are contributing but also ordinary people. There is also the problem of under-reporting bias. Noi Sian Koh, Nan Hu, Eric K. Clemons explains "under-reporting bias as a situation where consumers who are greatly satisfied or dissatisfied are more likely to report their reviews; correspondingly, those consumers with more moderate sentiments are less likely to post a review."

These factors highlight the significance of negative reviews and force product companies to work on the negative reviews promptly. This is also highlighted by K. Floyd, R. Freling, S. Alhoqail, H.Y. Cho, T. Freling (2014) where they conclude that "unresolved complaints are likely to motivate dissatisfied consumers to vent by posting negative (i.e., relatively lower) online product reviews that may deter legions of potential consumers from purchasing the offending brand in the future." Also, companies started utilizing this word of mouth reviews in their product planning. R.A. King, P. Racherla, V.D. Bush, states that "eWOM (electronic Word Of Mouth) has become more pervasive and mainstream, firms have devised strategies not only to use online feedback actively but also to engage eWOM providers actively."

The fake review issue faced by sites like Epinions continued to be a pain point for Amazon as well. Though it is not entirely possible to eliminate this issue, Amazon deployed solutions aimed to reduce this issue. Amazon tried to mitigate these issues by allowing reviews only to verified purchasers and by adding tags to indicate that the reviewer received the product for free in exchange for a review. They also deployed automatic fake review detection mechanisms.

A. Heydari, M. Tavakoli, N. Salim, Z. Heydari concludes that "there is a serious problem in the review spam detection area, which is the lack of gold standard datasets." This makes many automatic spam detection mechanisms to fail. This is further complicated by the way fake reviewers are writing content that blurs the line between honest ones and the fake ones. S. Banerjee, A.Y.K. Chua argues that "the process of writing fake reviews commences with extensive information gathering via common review websites such as TripAdvisor as well as search engines such as Google. The gathered information is then used as cues to write short, catchy, and succinct fake review titles, as well as informative and subjective fake review descriptions."

Hang Cui, Vibhu Mittal, Mayur Datar, states that "as the number of reviews available for any given product grows, it becomes harder and harder for people to understand and evaluate the prevailing/majority opinion about the product is." A lack of a 360-degree view of the product reviews may be the reason to blame for the scenario. With a consolidated review report, the impact of fake reviews is generally reduced since the customer ingests more review content than by manually scanning the review sites. When manually scanning, users tend to give more importance to detailed reviews. Fake reviews often exploit this tendency of customers and make their reviews more detailed and thereby attract more customer views. Under-reporting bias tends to be balanced or at least gets conveyed easily to the customer when a consolidated review reports are used.

# CHAPTER 3

# DATA COLLECTION AND ANALYSIS

## Data Collection Strategy

There are two options available when it comes to data collection strategy.

* Database based approach
* 'On the fly' data collection approach

The first option database based approach, collects data on periodic intervals, cleans it, and store them in a database. This option is the most popular option used by major big companies. Since the data is preprocessed and stored internally, the waiting time is significantly reduced for the end-user. However, on the downside, this option needs significant investment in terms of database infrastructure.

The use of database options necessitates the need to maintain a list of products that are being stored internally upon which the end-user can query. This will limit the set of products users can search on. After considering the scope and cost, it is decided not to pursue this database option for this project.

The next option is to collect the data on the fly when users submit their queries. This approach runs the dynamic data collection scripts once the user submits the query. This approach is customizable and can include a wide range of products. The problem with this approach is its query execution time.

User wait time now consists of the data collection time, analysis time, and dashboard creation time. To overcome the performance issues and for the limited scope of this paper, we set limits on the number of reviews being collected and limited the number of expert review pages being scraped. These limits can be easily revised based on the availability of high-performance computing systems in the future. These limits also enabled us to reduce the number of outdated reviews being pulled into the system. These limits also help to circumvent the review site's limitation on data scraping. We decided to proceed with 'on the fly' data collection approach to include more products and minimize cost.

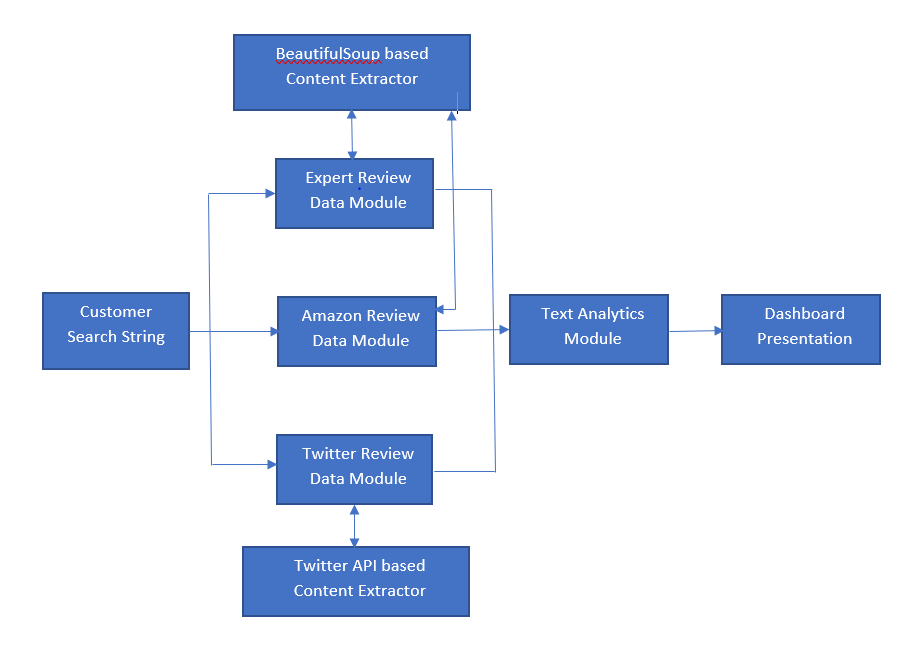
## Primary Data

The primary data for this project is customer reviews. Reviews are any text that is a formal assessment or critical appraisal of a product made by either the customer who previously used the product or by others who experienced the product. For the scope of the project, we do not differentiate whether the review being considered is spam or fake.

The primary data used in this analysis came from three sources: Expert Reviews, Twitter Reviews, Amazon Reviews. Data collection is performed using web scraping and APIs. The collected data is then cleansed before used in the analysis phase. We made sure not to collect any user information as it is not relevant when the reviews are consolidated.

The general data collection flow consists of the following modules:

* Data Extractor Module
* Expert Review Data Module
* Amazon Review Data Module
* Twitter Review Data Module
* Data Analysis Module



## Data Extractor Modules

As mentioned earlier, the project makes use of web scraping techniques and APIs to collect data. Web Scraping technique is used to collect Expert Reviews and Amazon Reviews. Twitter API is used to collected Twitter Reviews. Though Amazon provides API, it is challenging to retrieve individual reviews through API access. So, it is decided to use the web scraping technique for Amazon reviews as well.

Each webpage HTML has different named tags within which different elements of the site will be stored. To perform web scraping, we need to get familiarized with these tags for that particular website. We also need to make sure the HTML structure is relatively constant throughout the website entries.

Web scraping is performed using the python BeautifulSoup package. The HTML structure of the webpage from which content needs to be extracted is first analyzed to identify the selector element. The selector element is nothing but an HTML tag within which the review content is stored. Once the selector element is identified, it is then used within the module to retrieve the complete portion of the web content that is available within those tags. Let us call this as Soup. Once the Soup is extracted from the webpage, the content is then parsed to retrieve the reviews. This whole process is performed in the BeautifulSoup content extractor module. It accepts the webpage URL and the generic website name as inputs.

## Expert Review Data Module

This module extracts review data from a review website. We used TechRadar to extract review details for technology-related products. This is because TechRadar uses a defined HTML structure, which makes it easy to use in the code. For example, when two product reviews are extracted from the website, they happen to have the same selector element.

The module first picks the expert review website URL based on the product category and passes the URL for the specific product review within that website to the BeautifulSoup content extractor module, which then establishes a connection with the website and retrieves the content within the selector element tags. In the case of the TechRadar site, the selector element is 'pagination-numerical-list-item-link'. We will retrieve each paragraph within the selector tag and store them in a variable. We will also save the URL of the site for later use.

This extracted data comes up with a lot of embedded HTML tags like paragraph tags, hyperlink tags, and picture tags. These tags are of no use for our analysis and are removed using python. The cleaned data is stored in a python string variable and is passed to the analysis module for further refining and text analytics.

## Amazon Review Data Module

This module extracts review data from the Amazon website for the product searched by the customer. The module first constructs the Amazon URL for the specific product and pass it to the BeautifulSoup data extractor module. The data extractor module establishes a connection with the amazon URL and retrieves the page content using the selector element named 'data-hook': 'review-body'. We will extract only the review text and ignore any user-specific information.

Due to the restrictions imposed by Amazon to limit web scraping, we placed limits on the number of pages to traverse to get review content. We will sort the reviews by date before extracting to get only the latest reviews. This limitation restricts us to retrieve only a few hundreds of reviews.

For future work, the data can be cleansed to classify any reviews that include the words paid promotion. Even though these reviews may be genuine, they need to be classified to keep the user informed on the nature of the review. Also, verified purchase tagged reviews can be classified separately. These classifications are then passed along to the final dashboard so the user can make an informed decision.

## Twitter Review Data Module

Twitter is a rich source of data for mining consumer opinions and feedback. This module extracts review data from Twitter for the product searched by the customer. The module uses tweepy to interact with the twitter API and extract the tweets. Before starting with data collection, the first step is to get twitter developer access. Once the access is acquired, tweepy is used to connect with twitter using the acquired credentials from twitter developer id. Due to the limitation set by Twitter, we will collect tweets for a specified date frame. Tweepy query is customized to remove any retweets.

## Text Analytics Module

This module is used to perform sentiment analysis on the review text collected from various sources. Initially, we considered two leading packages in python NLTK and TextBlob for text analytics. The main criteria we set to choose one is whether it is easy to learn and implement.

NLTK is one of the most sophisticated and leading text analytics packages. It has many functionalities and very versatile. However, the same advantages add more complexity. Since we are interested only in sentiment analysis, it is better to deal with some other package that provides easy to use sentiment analysis option. Even though we decided not to use the NLTK package, we will still import it to make use of the stop words list provided by this package.

TextBlob, launched in 2013, is a relatively new entry to sentiment analysis. It uses rule-based sentiment classification, which provides Subjectivity and Polarity for the texts. Polarity gives a score between -1 to 1, which tells us whether the analyzed text is positive, negative, or neutral text. However, the main advantage is in terms of its ease of use and quick implementation. So for this project, we decided to proceed with TextBlob.

TextBlob expects data to passed in string format and converted to Blob data type. The cleaned review text from data modules is fed directly to the TextBlob package, and polarity is calculated for review as a whole and for each sentence within the review. The polarity is a score ranging between -1 and +1asentence. Any sentence with a score of 0 is considered neutral, any sentence with a score less than 0 is considered negative, and any sentence with a score above 0 is considered to be a positive sentiment text.

Once the sentiment analysis is completed, we will move to create the word cloud. It is essential we do sentiment analysis with complete data since, for word cloud, we will clean the data. Word cloud is a popular text analytics chart type. Word cloud is nothing but an image of words that are part of a textual data with the size of the text representing the frequency of the words. Word cloud provides an easy to consume chart that gives the user a view on the most frequented words of the reviews. This will provide a view of the product features.

Word Cloud needs a set of preprocessing to clean the words of any stop words, lemmatize words, remove punctuation, and numbers. Stop words are words (like the, is, was) that do not add value or meaning to the sentence and can be safely ignored before doing text analytics. We will use stop words from both the NLTK and TextBlob package. This stop word list is updated with any words we feel may be redundant or meaningless. We update the stop word list with these user-defined words and use them during the word cloud creation.

Lemmatization is the process of reducing different forms of words to their base word called a lemma. We perform lemmatization so as not to have different forms of words such as walking, walks, and other similar forms of a common base word walk within the word cloud. Lemmatization, unlike stemming, does not abruptly truncate the word, and hence it is preferred here over stemming.

The words are also converted to lower case since word cloud treats upper-case and lower-case words as unique words. We also removed any punctuation and numbers that are present in the text. Once the text is cleansed, Word Cloud is formed using the word cloud package. We also created charts to depict the sentiment values of the reviews.

We have established it is possible to consolidate reviews from multiple sources using python. The major limitation is due to the limits set by the social media platforms and amazon, which restricts the amount of data we collect. However, on the other side, it enabled us to collect data about more products. If this project is commercialized anytime in the future, then we can work with the companies to provide content for a fee and us them in the process. We are also able to do text analytics on the collected reviews in the form of sentiment analysis, and word frequency charts are computed in the form of word clouds.

# CHAPTER 4

# RESULTS

Let us examine the results by searching for a specific product. Let us take the example of the Canon 6D Mark II camera. When the customer searches for the product below, URLs are generated for each review module as follows:

Expert Review URL: <https://www.techradar.com/sg/reviews/canon-eos-6d-mark-ii-review>

Amazon URL: <https://www.amazon.in/Canon-EOS-6D-Mark-II/dp/B0749MNH83>

The expert review URL is appended with URL parameters to load all reviews content along with pagination information. The URL is then fed to the Expert review module, which paginates through the entire review content and extracts them into a python string variable. The review text is cleansed and passed on to the text analysis module. Sentiment analysis is performed and found to be 0.182. This suggests there is a weak positive sentiment about the product. A word cloud is created by the script using the expert review text as in Fig 1.

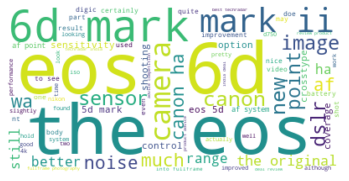


Fig 1: Expert Review Word Cloud

A frequency distribution plot of the sentiments of individual sentences in the review is plotted as shown in Fig 2

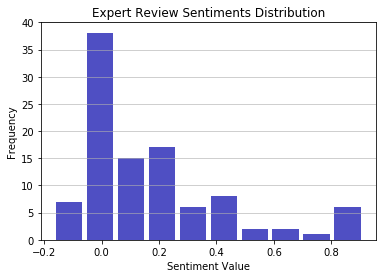


Fig 2: Expert Review Sentiment Distribution

The Amazon review URL is appended with URL parameters to load all reviews starting with the latest review. The URL is then fed to the Amazon review module, which paginates through the entire review content and extracts them into a python list. The review text is cleansed and passed on to the text analysis module. Sentiment analysis is performed and found to be 0.219. This suggests there is a weak positive sentiment about the product. A word cloud is created by the script using the amazon review text as in Fig 1.



Fig 3: Amazon Review Word Cloud

The frequency distribution plot of the individual reviews is plotted using the histogram as shown in Fig 4

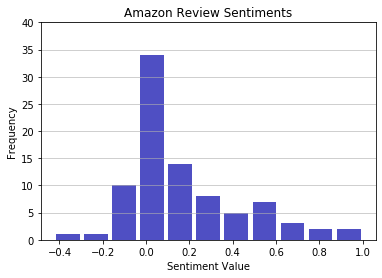


Fig 4: Expert Review Sentiment Distribution

We can see both the word clouds list some of the much-talked features of the camera like quiet operation, battery, and heaviness. It can still be refined by omitting the search string entered by the user.

From the frequency plots, we can the product enjoy fairs positive reviews though there is a considerable chunk of reviews clubbed under the 0-0.2 sentiment band. This needs further refinement in the future.

# CHAPTER 5

# CONCLUSION

The purpose of this project is to provide a way for consumers to view a customer aid tool that consolidates reviews from multiple platforms.Currently, there is a considerable demand for such a tool while there is nothing available in the market. Such a tool will not only benefit the consumers but also benefit the product owners.

We have established a way to consolidate reviews from various platforms. Consumers get a 360 view of the product reviews and can look at opinions from both experts and current owners on the same dashboard. Consumers also get an aggregated sentiment score, which lets them know how well the market talk is for that particular product. We also achieved one of the objectives of highlighting the features of the product using the word cloud.

By making use of the 'on the fly' data collection approach, we are able to make the product much more customizable for more products. However, this approach lacks in terms of performance since the data collection time also gets added to the user wait time. Options to minimize this issue are discussed in the recommendation and next steps section.

We also identified limitations in the form of restrictions imposed by social media platforms in data gathering through web scraping and twitter API. This limitation significantly reduced our ability to collect a vast volume of data, which can be rectified by subscribing to paid services from social media companies.

The aggregated sentiment score reduces the impact of fake reviews and under-reporting bias since the score considers all reviews equally and does not weight any one of the reviews higher, which is often the case with manual scanning.

## Recommendations/Next Steps

The current scope of the project, due to lack of sufficient time, did not include some of the advanced techniques that can enhance the quality of the reviews being considered. Some of these will significantly improve the data and thereby improve user trust on the dashboard. Let us see some of these improvements that can be applied in future

**Opinion Spam Detection:** The current project does not omit spams that are possible in the reviews. Though there are many techniques available, Noekhah, S., Nb, S., Zakaria, N. H., in their research paper, identified an efficient way using a novel approach called Multi-iterative Graph-based opinion Spam Detection. They conclude that the accuracy of 91.2% was achieved when the MGSD model was used.

**Emoji Classification:** Emojis are now an essential part of almost all social media platforms and text communications. Hence any sentiment analysis needs to consider them and should be able to classify them for better results. M. Rathan, V.R. Hulipalled, K.R. Venugopal, L.M. Patnaik in their research paper details an efficient way to classify emojis, which improves the accuracy of text classification. Considering more and more text interactions involve emojis, it becomes essential to consider them in our classification methods.

**Hybrid Data Collection:** For the limited scope of this paper, it made sense to use the 'on the fly' data collection approach. However, to make this into a product, we may need to invest in technologies that will improve performance significantly. One such option is to make use of hybrid data collection approach wherein both database options (for frequently searched product categories) and 'on the fly' option are used (for all other options).

**Bot Detection:** A significant percentage of twitter accounts are automated bot accounts. Often these accounts are paid up or provide misleading reviews and contents. Apart from spam detection, we need to remove bot accounts from our twitter data and analysis. These bot accounts often use sophisticated and novel methods to avoid detection, and hence an efficient detection mechanism is needed. Jorge Rodríguez-Ruiz, Javier Israel Mata-Sánchez, Raúl Monroy, Octavio Loyola-González, Armando López-Cuevas proposes a one-class classification bot detection which improves detection performance by 0.89 (measured using AUC).

**Verified Purchase Reviews:** Amazon started adding tags to the reviews to make it easy for the customers to understand the nature of the review. Paid Reviews are when the reviewer performs the review after accepting the product in exchange for a review. Though these reviews are useful, it needs to be flagged so the user can filter them out if needed from the dashboard. A verified purchase tag is added when amazon can ascertain that the reviewer indeed purchased the product. It is an easy way to convey to the customer that it is a genuine review. These tags can be scrapped the same way reviews are scraped and made visible in the dashboard, so the user can filter them to make a more informed decision. Though much of the required code is available already, this is not included due to the time constraint.

# References

1. Anjana, S.S. "A study on factors influencing cosmetic buying behavior of consumers," International Journal of Pure and Applied Mathematics, vol. 118, no. 9, pp. 453-459, 2018.
2. X. Li, L.M. Hitt, Price effects in online product reviews: an analytical model and empirical analysis, MIS Q., 34 (4) (2010), pp. 809-831
3. Ying-Ping Liang, The Relationship between Consumer Product Involvement, Product Knowledge and Impulsive Buying Behavior, Procedia - Social and Behavioral Sciences, Volume 57, 9 October 2012, Pages 325-330
4. Arezou Ghiassaleh, Bruno Kocher, Sandor Czellarc, Bestseller!? Unintended negative consequences of popularity signs on consumer choice behavior, International Journal of Research in Marketing, Volume 57, 5 June 2020
5. Ramanathan, U., Subramanian, N. and Parrott, G. (2017), "Role of social media in retail network operations and marketing to enhance customer satisfaction", International Journal of Operations & Production Management, Vol. 37 No. 1, pp. 105-123.
6. Zhu, F., Zhang, X., Impact of online consumer reviews on Sales: The moderating role of product and consumer characteristics, Journal of Marketing, Volume 74, Issue 2, March 2010, Pages 133-148
7. Yan Li, Ruijuan Wu, Dongjin Li, Hejun Fan, Can scarcity of products promote or restrain consumers' word-of-mouth in social networks? The moderating roles of products' social visibility and consumers' self-construal, Computers in Human Behavior, Volume 95, June 2019, Pages 14-23
8. Mudambi, S., & Schuff, D. (2010). Research Note: What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. MIS Quarterly, 34(1), 185-200.
9. Olivier Kraaijevel, Johannes De Smedt, The predictive power of public Twitter sentiment for forecasting cryptocurrency prices, Journal of International Financial Markets, Institutions and Money, Volume 65, March 2020, 101188
10. B. Bickart, R. Schindler, Internet Forums as Influential Sources of Consumer Information, Journal of Interactive Marketing, 15 (2001), pp. 31-40
11. T. Hennig-Thurau, K.P. Gwinner, G. Walsh, D.D. Gremler, Electronic word-of-mouth Via Consumer-Opinion Platforms: What Motivates Consumers to Articulate Themselves on the Internet?, Journal of Interactive Marketing, 18 (2004), pp. 38-52
12. E. Maslowska, E.C. Malthouse, V. Viswanathan, Do customer reviews drive purchase decisions? The moderating roles of review exposure and price, Decision Support Systems, 98 (2017), pp. 1-9
13. R. Iyer, M. Griffin, Modeling word-of-mouth usage: A replication, Journal of Business Research (2020)
14. Khalid Saleh (2015). The Importance of Online Customer Reviews [Infographic] [Blog post] Retrieved from: https://www.invespcro.com/blog/the-importance-of-online-customer-reviews-infographic/
15. J. Lee, D.h. Park, I. Han, The effect of negative online consumer reviews on product attitude: an information processing view, Electronic Commerce Research and Applications, 7 (2008), pp. 341-352
16. F.L. Weisstein, L. Song, P. Andersen, Y. Zhu, Examining impacts of negative reviews and purchase goals on consumer purchase decision, J. Retailing Consum. Serv., 39 (2017), pp. 201-207
17. Beverley A. Sparks, Victoria Browning, The Impact of Online Reviews on Hotel Booking Intentions and Perception of Trust, Tourism Management, 32 (6) (2011), pp. 1310-1323
18. Cesare Amatulli, Matteo De Angelis, Giovanni Pino, Gianluigi Guido, An investigation of unsustainable luxury: How guilt drives negative word-of-mouth, International Journal of Research in Marketing, In Press, Corrected Proof, Available online 13 April 2020
19. Jason Q.Zhang, Georgiana Craciuna, Dongwoo Shin, When does electronic word-of-mouth matter? A study of consumer product reviews, Journal of Business Research, 63 (2010), pp. 1336 - 1341
20. N. Purnawirawan, M. Eisend, P. De Pelsmacker, N. Dens (2015), A meta-analytic investigation of the role of valence in online reviews, Journal of Interactive Marketing, 31 (2015), pp. 17-27
21. B. von Helversen, K. Abramczuk, W. Kopeć, R. Nielek, Influence of consumer reviews on online purchasing decisions in older and younger adults, Decis. Support. Syst., 113 (2018), pp. 1-10
22. H. Hong, D. Xu, G.A. Wang, W. Fan, Understanding the determinants of online review helpfulness: a meta-analytic investigation, Decis. Support. Syst., 102 (2017), pp. 1-11
23. Noi Sian Koh, Nan Hu, Eric K. Clemons, Do Online Reviews Reflect a Product's True Perceived Quality? An Investigation of Online Movie Reviews Across Cultures, Electronic Commerce Research and Applications, 9 (5) (2010), pp. 374-385
24. K. Floyd, R. Freling, S. Alhoqail, H.Y. Cho, T. Freling, How online product reviews affect retail sales: A meta-analysis, Journal of Retailing, 90 (2) (2014), pp. 217-232
25. R.A. King, P. Racherla, V.D. Bush, What we know and don't know about online word-of-mouth: a review and synthesis of the literature, Journal of Interactive Marketing, 28 (3) (2014), pp. 167-183
26. A. Heydari, M. Tavakoli, N. Salim, Z. Heydari, detection of review spam: a survey, Expert Syst. Appl., 42 (7) (2015), pp. 3634-3642
27. S. Banerjee, A.Y.K. Chua, Understanding the process of writing fake online reviews, Ninth International Conference on Digital Information Management (ICDIM 2014) (2014), pp. 68-73
28. Hang Cui, Vibhu Mittal, Mayur Datar, Comparative Experiments on Sentiment Classification for Online Product Reviews, IEEE Journal, 2015
29. Noekhah, S., Nb, S., Zakaria, N. H. (2020). Opinion spam detection: Using multi-iterative graph-based model. Information Processing & Management, 57(1), 102140.
30. M. Rathan, V.R. Hulipalled, K.R. Venugopal, L.M. Patnaik, Consumer insight mining : Aspect based Twitter opinion mining of mobile phone reviews, Applied Soft Computing Journal, 68 (2018), pp. 765-773
31. Jorge Rodríguez-Ruiz, Javier Israel Mata-Sánchez, Raúl Monroy, Octavio Loyola-González, Armando López-Cuevas, A one-class classification approach for bot detection on Twitter, Computers & Security, Volume 91, April 2020, 101715

# Appendix

Code used for the data collection and analysis can be accessed in the below Google Drive links

<https://drive.google.com/file/d/1yqSgEwwoyRarzXOam4kJQKMd5gGb1SrV/view?usp=sharing>

<https://drive.google.com/file/d/1ZVzcwDHAScQbA8vYtkGZ9I_-v8F3LyHB/view?usp=sharing>