

**MSC BUSINESS INTELLIGENCE & ANALYTICS**

# MODULE CODE : 7COSC012W

MODULE TITLE : **Msc Project**

MODULE LEADER : **Rolf Banziger**

Master Thesis Proposal

**Sentiment Analysis, Predicting Star Ratings and creating Machine Learning Pipeline**

**Using Amazon Data**

STUDENT NAME **: RAJ PRAVIN RAJENDRAN**

STUDENT ID **: W1795435**

SUBMISSION DATE **: 15-06-2022**

Abstract

In an extremely competitive world, it is critical for firms to listen to their consumers and seek to act on their comments. This has been aided by the rising trend of customers basing their purchase decisions on their personal opinions. The goal of this project is to enable Amazon.com to obtain a knowledge of their customers' attitude toward mobile phone products, service, and the organization as a whole, as well as to predict sentiment for data set. This will be accomplished by developing and deploying a sentiment analysis tool that is linked to the present CRM system and uses feedback and start rating data as input. I feel that by collecting this data, the company will be able to make better product selections and respond to negative comments more quickly, resulting in greater client relationships and retention.

Additionally, my deliverable is to find the sentiment score for certain brand and finally predict the star ratings whether the brand gets right reviews on the catalogue page or not and to predict what rating would the user give for a specific brand

Furthermore, creating a Machine learning pipeline in terms of automating the entire ML features using the library **Tkinter**.

**Introduction:**

Obtain insights into your consumers' perspectives has long been a tactic used by the most effective and forward-thinking firms in their pursuit of unmatched customer service. Bill Gates was famously cited as saying that in business, "your most unhappy clients are your greatest source of learning," which has helped Microsoft to achieve greatness.

Nevertheless, it is not only the most unhappy consumers that provide useful information for a business; a strong grasp of positive feedback is also crucial, helping a firm to understand what they are doing well and therefore what they should continue to do in order to remain competitive. A thorough comprehension of these "sentiments" enables:

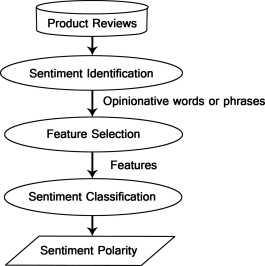
**1. Increased customer loyalty and retention** - It is considerably more cost efficient for a firm to keep customers rather than invest money to recruit new consumers. Listening to consumers will increase retention and decrease churn in a firm.

**2. Satisfied customers refer** - By utilising and acting on customer feedback, consumers are more likely to feel heard and hence more inclined to promote your services.

**3. Competitive advantage** - Understanding what your consumers enjoy and hate is the best predictor of what the rest of the market is looking for. Thus, a strong customer-focused approach allows a company to stay one step ahead of competition.

**Background:**

The computer examination of people's views, attitudes, and feelings about an item is known as sentiment analysis (SA) or opinion mining (OM). Individuals, events, or concepts can all be represented by the entity. Reviews are more likely to discuss these subjects. The terms SA and OM can be used interchangeably. They are expressing a same message. However, according to some experts, OM and SA have slightly distinct perspectives.



To do sentiment modelling, two things are required: an understanding of the sentiment information that words contain, and a thorough morphosyntactical examination of the text. Common patterns such as negation or coordinated phrases conveying opposing concepts are not accurately evaluated with only a list of which words are positive and negative. On the other hand, if we have a morphosyntactic analysis, you don't know what has a positive or negative connotation.

**Feature selection in sentiment classification:**

**Term presence and frequency:**

Individual words or word n-grams, as well as their frequency counts, are among these properties. It either utilises term frequency weights to highlight the relative relevance of characteristics or offers the terms binary weighting (zero if the word appears, one otherwise).

**Parts of Speech (Pos):**

Finding adjectives is crucial since they are significant indications of how people feel.

**Opinion words and phrases:**

These are terms that are widely used to indicate feelings such as good or terrible, like or dislike. Some sentences, on the other hand, offer opinions without employing opinion words. For instance: it set me back an arm and a leg.

**Negations:**

When negative words arise, it might cause a shift in perspective, as though not good is the same as awful.

A sentiment model is made up of entries, which are specified by a word or many words. In certain circumstances, we'll want to create more complicated scenarios than simply a single word or many words, thus we'll utilise subentries.  All entries and subentries are defined by their sentiment behaviour.

For example, the term "LOW." In general, it has a negative sentiment. Similar to that "Low pressure," "Low marks" are negative sentiment. Although if we look at pricing, you can easily find the word "low price", that is not a negative sentiment.

Sentiment analysis has progressed from a technique used to provide an insight into current opinion to a tool that may be used to make predictions. The automobile sector has quickly adopted this predictive technique, with businesses such as Buick using negative social media criticism to drive the construction of proposed car models.

Despite the fact that several research on sentiment analysis have focused on utilising social media as the major source of opinion, the issue of recruiting and its related products has received relatively little citation in social spaces. Searching the phrase reed.co.uk on Twitter yields a relatively small number of relevant tweets, showing the lack of potential that social media gives for gauging consumer attitude toward a firm of the sort of reed.co.uk. Based on this, the utilisation of call logs gives a solid option to gaining an understanding of consumer perspectives in a convenient format.

Problem Statement

Amazon.com is a massive Internet-based business that sells over 12 million items either directly or as a middleman between other businesses and Amazon.com's millions of consumers. However, Amazon's typical return rate ranges from 5% to 15%. Amazon, like many other online businesses, has a multitude of touchpoints during the customer's journey. One of the most common interactions is with the Mobile phone goods on the detail page.

Because the number of returns for Amazon electronic items is excessive, I propose developing a predictive analytic model that explains why customers are returning the products and ensures that Amazon suppliers are not pushed out of the sector.

My problem statement is to predict the **star rating** on the basis of customer feedbacks given to the particular product. Hence, my target variable is **ratings** in my dataset. Additionally, we can predict the star rating of a brand to extract user feedback from open-text reviews without reading them.

Objective

I propose to create a sentiment analyser for the downloaded dataset which would actually predict the sentiment and gives us the customer emotions for the particular product. Additionally, the project aims to find the sentiment of the customers using different mobile phones brands and strategically addresses the manufacturer/supplier churn (Vendors associated with Amazon). To perform this analysis holistically, the data needs to be analysed across various Machine Learning algorithms. Considering this, I aim to create the Machine learning Pipelines using the Tkinter library to automate the entire model deployment and analysis process.

The use of the sentiment analyser and Machine learning technique provide:

• A clearer idea of customer sentiment on an mobile phone brands with in the amazon where I should be able to establish a sentiment for each interaction.

• Looking more holistically, aggregating and allowing for grouping by specific brands, will make it possible for the brand owner to easily ascertain the sentiment of the customer to Amazon.com

• Highlights into some of the key themes of why customer has given the negative sentiment feedback, products and services, so that the business has an insight into which are being discussed.

It is my theory that by achieving these goals, Amazon would be able to improve customer satisfaction and retention rates. These are easily measurable KPIs for the company, since they presently track client churn. Furthermore, these goals allow me to improve my abilities in crucial tools for sentiment analysis, such as Python, while also challenging me to identify the best approaches to model for maximum accuracy.

The relevance of this endeavour to the business arises from the company's existing incapacity to analyse customer sentiment. Whereas past work in this field has concentrated on data gathered from kaggle website, my goal is to predict the, sentiment for the feedbacks given by the customer.

Deliverables

The following are the project's primary deliverables based on the aforementioned objectives:

1. A compilation of data from Amazon-affiliated mobile phone companies, with sentiment classifications

2. Star rating predictions using several classification model and freeze the model which has highest accuracy

3. Automation of the Machine learning pipelines

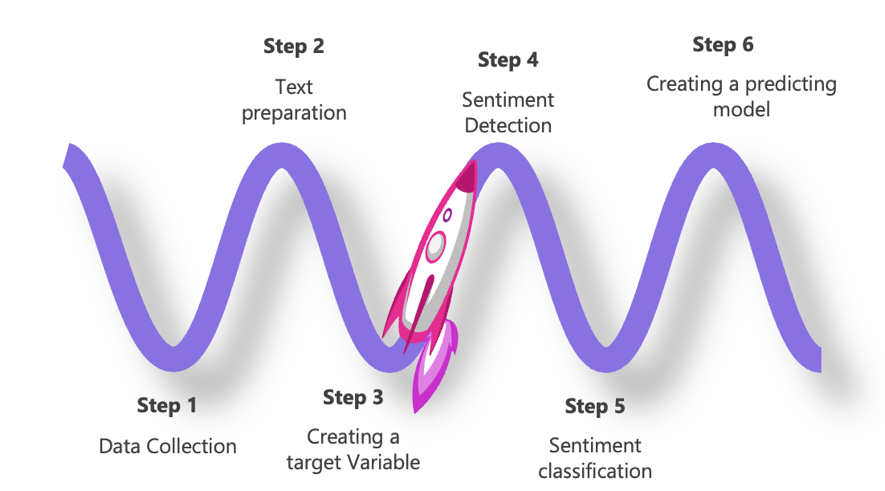
Amazon is able to get useful insights on their customers' opinions of the business, their products, and the services they give by providing the promised deliverables. This breathes new life into data that had previously been kept but had little to no purpose. They can therefore obtain a commercial edge, putting them in a strong position in their market. In addition, we can forecast star rating in our model for future data.

Further Applications

Sentiment analysis of Amazon's mobile brand data is only the tip of the ocean in terms of what is possible with the suggested sentiment analysis tool. As previously said, Amazon has several touch points with its consumers and can thus acquire a 360-degree view of them by merging various sources. The OU (Organisational Unit) ID issued to each client as a unique identification enables for smooth integration.

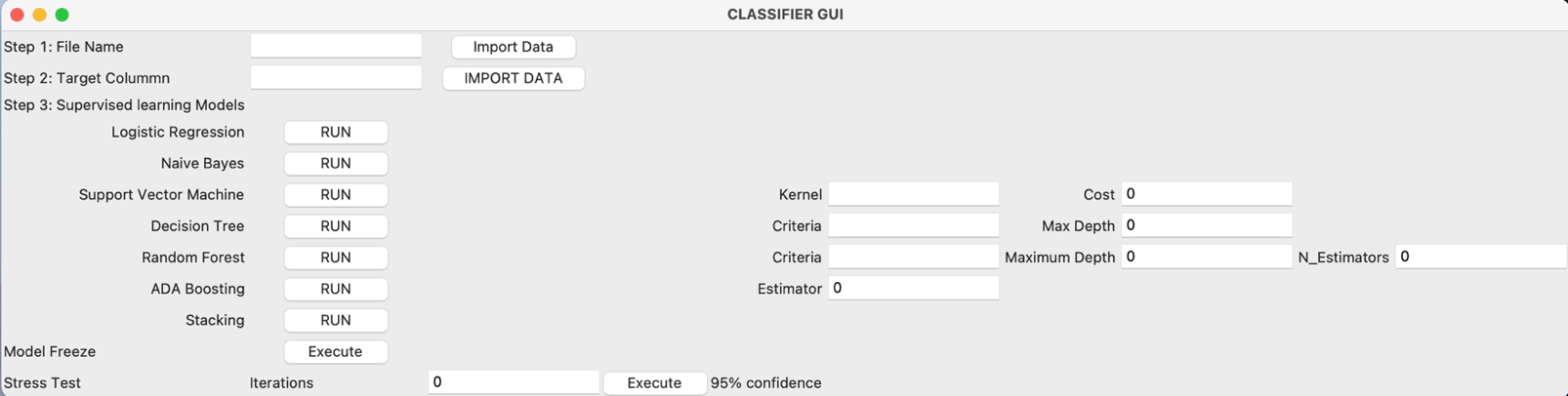
Approach

Building a comprehensive sentiment star rating prediction model is critical to the project's success, and I will complete six major stages:



The earliest phases require collecting of the data and pre-processing (text preparation) to guarantee that the sentiment analyser's inputs are ideal. Stage 3 involves creating a target variable using the review rating in order to determine the forecast. Stages 4 and 5 are concerned with the actual finding and categorization of sentiment from call log data. Stage 6 is creating a model and determining which model has the best accuracy. Finally, the model is centred on the display of the retrieved data, which will take various forms depending on the level selected.

Following the completion of the model, I am developing a Machine Learning pipeline GUI in which we can only see the accuracy score for all of the models in an automated tool. The image below is a dialogue window in which I enter my goal column and test the data using the associated models. We can finalise and freeze the model with the highest accuracy for deployment.



Resource & Tools

In order to develop the sentiment analyser, there are a number of tools that will be employed which have been identified as the appropriate tools for the scenario

**Python** - Open source object-orientated programming language. Making use of libraries including Pandas, NLTK, TextBlob, Numpy, Sci-ket learn, Seaborn, Matplotlib.

Additionally, I use “Natural language Processing” for text mining and “Tkinter” to automate the ML pipeline.

**Jupyter Notebook -** The Jupyter Notebook is an open source web application that you can use to create and share documents that contain live code, equations, visualizations, and text.

Risks & Potential Mitigations

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Potential Impact** | **Likelihood** | **Mitigation** |
| **Data Access** | High | Low | Data is freely available in Kaggle website |
| **Data Quality** | High | High | Usage of option sets and normalize your data |
| **Open software integration** –When merging open source software with licenced software, issues may develop. | High | Medium | Investigate integration approaches. A backup plan will be to create a stand-alone software. |
| **Data security** - Because the data involved is so sensitive, there is concern that it will be compromised throughout the project. | High | Low | Communicate with the "Head of data and integration" to learn best practises and how to ensure data security. |
| **Scope** - There may be a desire to keep adding additional features to the system, which might cause it to take longer to complete. | Medium | Low | This risk can be reduced by following a well-defined project timeframe and explicitly specifying the deliverables. |
| **Inadequate hardware/software** - The use of low-end hard/software will have an influence on system performance. | Medium | Low | Early run-time and performance testing to confirm that the final hard/software utilised is appropriate for the task |

Project plan

To ensure that the project runs successfully and on time, a detailed project plan (**Appendix 1**) has been created to emphasise the critical checkpoints.

|  |  |  |
| --- | --- | --- |
| **Day** | **Week** | **Step** |
| 10 May 2022 | 1 | Creation of project proposal |
| 17 May 2022 | 2 | Exploratory Data Analysis |
| 24 May 2022 | 3 | Exploratory Data Analysis |
| 31 May 2022 | 4 | Exploratory Data Analysis |
| 7 June 2022 | 5 | Introduction and Background |
| 14 June 2022 | 6 | Introduction and Background |
| 21 June 2022 | 7 | Literature review |
| 28 June 2022 | 8 | Literature review |
| 5 July 2022 | 9 | Data cleaning |
| 12 July 2022 | 10 | Data Visualization |
| 19 July 2022 | 11 | Splitting the data |
| 26 July 2022 | 12 | Model Creation and Testing |
| 2 August 2022 | 13 | Result interpretation |
| 9 August 2022 | 14 | Conclusion |
| 16 August 2022 | 15 | Final project review |
| 23 August 2022 | 16 | Final project review |
| 30 August 2022 | 17 | Presentation to the mentor |
| 6 September 2022 | 18 | Submission |

Bibliography

Gates, B. (1999), Business @ the speed of thought. New York, NY: Warner Books.

Beard, R. (2013), 5 Reasons You Need To Be Tracking Customer Satisfaction. [online] Client Heartbeat Blog. Available at: http://blog.clientheartbeat.com/tracking-customer-satisfaction/ [Accessed 8 May 2016].

Pang, B. & Lee, L., (2008), "Opinion Mining and Sentiment Analysis", FNT in Information Retrieval, vol. 2, no. 12, pp. 1-135

Rajeck, (2016), "Three steps to improve digital customer experience (CX) [APAC Case Studies]", Econsultancy. [Online]. Available: https://econsultancy.com/blog/67201-three-steps-to- improve-digital-customer-experience-cx-apac-case-studies/. [Accessed: 08- May- 2016].

Collomb, A., Costea, C., Joyeux, D., Hasan, O. and Brunie, L., (2014), A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation

https://www.sciencedirect.com/science/article/pii/S2090447914000550