FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Customer Xperience – Using Social Media Data to Generate Insights for Retail

João Pedro Nogueira Ribeiro

WORKING VERSION

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

Supervisor: Pedro Alexandre Rodrigues João

June 25, 2018

Resumo

No mundo atual, as redes sociais são um meio de comunicação vastamente aceite e difundido com um número de utilizadores que tende a crescer todos os dias. No entanto, com o crescimento da popularidade ao longo dos anos, as plataformas das redes sociais começaram a ser muito mais do que isso; Estas páginas *web* começaram a funcionar como um púlpito de onde as principais entidades podem gritar e ser ouvidas por milhões: agências de notícias podem difundir as suas notícias, políticos e celebridades podem partilhar os seus pontos de vista e fazer declarações, e empresas podem representar a sua marca, fazer promoções e anúncios. Tendo isto em conta, as redes sociais também se tornaram um extenso repositório de informação sobre os pensamentos, sentimentos, opiniões e ações das pessoas; Isto pode, em teoria, permitir tirar conclusões significativas sobre o comportamento dos consumidores para que as empresas de retalho melhorem a sua perceção de mercado e tomem decisões mais informadas.

A vantagem óbvia deste tipo de informação é a sua relativamente fácil aquisição e a sua grande quantidade, no entanto, a informação recolhida de redes sociais é geralmente muito "ruidosa", com a grande maior parte dela sendo difícil de entender e até mesmo sem significado. Não obstante, através da filtragem da informação recolhida e de uma posterior análise sentimental, é possivel obter informação suficientemente relevante para ser usada numa análise de retalho extensiva, abordando áreas tais como previsão de procura, desempenho de loja, lealdade para com a marca, melhoria de produto, eficácia de promoções, entre outras. Apesar de todas as dificuldades e falhas expectáveis associadas ao processo, as possibilidades que este método providencia são entusiasmantes e representam um novo período emocionante para o retalho.

Abstract

In todays' world, social media platforms are a widely accepted and spread form of communication and social activity, with a tremendous number of users that tends to grow every day. However, with the growth of popularity over the years, social media platforms began to convey much more than that; These websites began to function as a pulpit from where major entities could shout and be heard by millions: news agencies can spread their news, politicians and celebrities can share their views and make statements, and companies can represent their brand and make promotions and announcements. Regarding this, the social media also became a massive repository of information regarding people's thoughts, feelings and actions that relate to a myriad of subjects, many of these relating directly to the most popular news and trends. Because of the way users utilize social media it is possible to draw a line between their sentiment, interests, opinions, and actions; This can, in theory, allow to draw meaningful conclusions about customer behavior in order for retail companies to improve their market perception and empower better informed decisions.

The obvious advantage of this data is its relatively easy acquisition and very large quantity, however, data collected from Social Media is generally very noisy, with the big majority of it being hard to understand or even meaningless. Nonetheless, by filtering the collected data and by putting it through a sentiment analysis it is possible to obtain data relevant enough to be used on an extensive retail analysis covering areas such as demand forecast, store performance, brand engagement, product improvement, promotion effectiveness, among others - This is the objective of the Customer Xperience project. Despite all the expected difficulties and short-comings associated to the process, the possibilities this method provides are exciting and represent a thrilling new period for retail.

Acknowledgements

Throughout this project several people had a direct or indirect impact on its success, and to those I wish to extend my sincerest gratitude.

Firstly I am thankful to God, for guiding me all the way until here; Since the beginning of my life until the present day He always gave me His invaluable help and guidance, without Him, none of this would be possible.

Secondly, I wish to thank my direct supervisor from Wipro, Susana Ribeiro, who despite all complications and setbacks always remained as available and as helpful as possible, and provided valuable guidance in the best and in the worst moments.

I thank Wipro, for giving me the opportunity of conducting my thesis in a great work environment, where everybody feels welcome and valued, and where there is always someone ready to help.

I would also like to thank Er. Pedro Alexandre Rodrigues João, my thesis supervisor from FEUP, who always showed availability to help with whatever guidance was needed, and helped me and my colleagues organize our time in the best way.

I'd like to thank to Profs. Maria Antónia Carravilla and José Fernando Oliveira, for their precious help in exploring the theme of demand forecast, which had a significant importance for this thesis.

It wouldn't be fair to not thank all of my colleagues that helped me throughout my course, specially the ones who conducted their thesis at Wipro with me; We often shared problems and challenges, and helped each other as much as we could, fostering a healthy work environment. For this I thank Filipe Pereira, Ana Cunha, Nuno Dias, and Gonçalo Martins

Finally I would like to extend my gratitude to my girlfriend and to my family who are the pillars of my life, with special gratitude to my parents, who always sacrificed themselves to provide me with a quality education, and always instructed me in the right path. Their effort unlocked many opportunities in my life that I otherwise wouldn't have, and their education made me the person I am today.

João Ribeiro

"Whatsoever thy hand findeth to do, do it with thy might;
for there is no work, nor device, nor knowledge, nor wisdom, in the grave, whither thou goest."

Ecclesiastes 9:10 - The Holy Bible

Contents

1 Introduction **1**
- 1.1 Company Presentation . . . 1
  - 1.1.1 The Spirit of Wipro . . . 1
  - 1.1.2 Wipro's History in Portugal . . . 2
- 1.2 Context . . . 3
- 1.3 Motivation . . . 4
- 1.4 Project Description . . . 5
- 1.5 Methodology . . . 6
- 1.6 Thesis Structure . . . 6

2 State of the Art **7**
- 2.1 Data Collection (Web Crawling) . . . 7
  - 2.1.1 Apache Nutch . . . 8
  - 2.1.2 Scrapy . . . 9
  - 2.1.3 Storm-Crawler . . . 9
  - 2.1.4 Wipro Holmes - Data Discovery Platform - Collection . . . 9
- 2.2 Sentiment Analysis . . . 9
  - 2.2.1 Amazon Comprehend . . . 10
  - 2.2.2 IBM Watson - Natural Language Understanding . . . 12
  - 2.2.3 Google Cloud Platform - Natural Language . . . 13
  - 2.2.4 Microsoft Azure Text Analytics . . . 15
  - 2.2.5 Solution Comparison for Sentiment Analysis . . . 16
- 2.3 Retail Analytics . . . 17
  - 2.3.1 SAS - Retail Omnichannel Analytics . . . 17
  - 2.3.2 Amazon - AWS for Retail . . . 19
  - 2.3.3 IBM Watson - Analytics for Social Media . . . 20
  - 2.3.4 Wipro Holmes - Data Discovery Platform - Analytics . . . 21
  - 2.3.5 Solution Comparison for Retail Analysis . . . 22

3 Demand Forecast **25**
- 3.1 Time Series Analysis . . . 26
  - 3.1.1 Linear Regression Analysis . . . 28
  - 3.1.2 Simple Moving Average . . . 28
  - 3.1.3 Weighted Moving Average . . . 28
  - 3.1.4 Exponential Smoothing . . . 29
- 3.2 Causal Relationship Forecasting . . . 30
  - 3.2.1 Multiple Regression Analysis . . . 30
- 3.3 Qualitative Techniques . . . 30

3.3.1 Market Research . . . 31
3.3.2 Historical Analogy . . . 31
3.3.3 Panel Consensus . . . 31
3.3.4 Delphi Method . . . 32
3.4 Forecasting Errors . . . 32
3.4.1 Sources of Error . . . 32
3.5 Forecasting Using Social Media . . . 33
3.5.1 Intricacies of Social Media Forecasting . . . 33
3.5.2 Social Media Forecasting in different areas . . . 34
3.5.3 Conclusions and Improvements . . . 35

4 Project Development **37**
4.1 Initial Project Outline and Research . . . 37
4.2 Final Project Outline and Research . . . 39
4.3 Incoming Batches of Data . . . 40
4.4 User Negativity Bias . . . 44
4.5 Preliminary Analysis of the Data . . . 45
4.6 Store Performance Drivers . . . 48
4.6.1 Store Performance KPI's . . . 48
4.6.2 Comment Selection . . . 49
4.6.3 Analysis and Outputs . . . 51
4.7 Demand Forecast and Demand Driven Replenishment . . . 52
4.7.1 Demand Forecast with Social Media - A Delphi Method Approach . . . 52
4.7.2 Comment Selection . . . 55
4.7.3 Analysis and Outputs . . . 56
4.8 Brand Engagement . . . 58
4.8.1 Brand Engagement KPI's . . . 58
4.8.2 Comment Selection . . . 59
4.8.3 Analysis and Outputs . . . 61

5 Conclusions and Future Improvements **63**
5.1 Conclusions . . . 63
5.2 Future Work and Improvements . . . 64
5.2.1 Improved Sentiment Analysis . . . 64
5.2.2 Remaining Use Cases . . . 64
5.2.3 Implementation and Integration with the Oracle Retail Suite . . . 65
5.2.4 Crossing Social Media Data with Historical Company Data . . . 65
5.2.5 Use of Machine Learning . . . 65
5.2.6 Dashboard . . . 66

A Appendix 1 - Customer Xperience Final Project Outline **67**

B Appendix 2 - Data Discovery Platform - Data Sample **77**

C Appendix 3 - Questionnaire 1 **79**

D Appendix 4 - Questionnaire 1 Answers **81**

E Appendix 5 - Questionnaire 2 **85**

F Appendix 6 - Questionnaire 2 Answers **89**

References **93**

List of Figures

1.1 Wipro's Official Logo [1] . . . 2
1.2 Enabler's Official Logo [2] . . . 3
1.3 The Work Subject of the Present Thesis . . . 5

2.1 Amazon Comprehend Interface - Entities [3] . . . 11
2.2 Amazon Comprehend Interface - Language and Sentiment [3] . . . 12
2.3 IBM Natural Language Understanding Demo - Sentiment Analysis [4] . . . 13
2.4 Google Cloud Storage Natural Language - Entity Identification [5] . . . 14
2.5 Google Cloud Storage Natural Language - Sentiment Analysis [5] . . . 15
2.6 Microsoft Azure Text Analytics - Sentiment Analysis [6] . . . 16
2.7 AWS for Retail - Use Case:Single View of the Customer [7] . . . 20
2.8 IBM Watson Analytics for Social Media - Conversation Clusters [8] . . . 21

3.1 Historical Product Demand Consisting of a Growth Trend and Seasonal Demand [9] 27

4.1 Wipro's Initial Project Definition . . . 37
4.2 Wipro's Second Project Definition . . . 38
4.3 Wipro's Customer Xperience Older Project Concept . . . 39
4.4 Customer Xperience Final Project Concept . . . 40
4.5 DDP Data Fields . . . 41
4.6 DDP Data - Incoherent Categories . . . 41
4.7 DDP Data - Incoherent Sentiment Analysis . . . 42
4.8 DDP Data - Comment Length Issue . . . 42
4.9 Distribution of Users Across Social Networks . . . 46
4.10 General Sentiment Trend of the Collected Data . . . 46
4.11 Monthly Comment Count for the Collected Data . . . 47
4.12 Category Count for the Previous Months . . . 47

A.1 Customer Xperience Final Project Outline - Pages 1,2 and 3 . . . 68
A.2 Customer Xperience Final Project Outline - Pages 4,5 and 6 . . . 69
A.3 Customer Xperience Final Project Outline - Pages 7,8 and 9 . . . 70
A.4 Customer Xperience Final Project Outline - Pages 10,11 and 12 . . . 71
A.5 Customer Xperience Final Project Outline - Pages 13,14 and 15 . . . 72
A.6 Customer Xperience Final Project Outline - Pages 16,17 and 18 . . . 73
A.7 Customer Xperience Final Project Outline - Pages 19,20 and 21 . . . 74
A.8 Customer Xperience Final Project Outline - Pages 22,23 and 24 . . . 75

B.1 Data Discovery Platform - Data Sample . . . 78

C.1 Questionnaire 1 - Page 1 79
C.2 Questionnaire 1 - Page 2 80

D.1 Questionnaire 1 Answers - Page 1 82
D.2 Questionnaire 1 Answers - Page 2 83

E.1 Questionnaire 2 - Page 1 86
E.2 Questionnaire 2 - Page 2 87

F.1 Questionnaire 2 Answers - Page 1 90
F.2 Questionnaire 2 Answers - Page 2 91

List of Tables

3.1 A Guide to Selecting an Appropriate Forecasting Method [9] 27

Abbreviations and Symbols

HOLMES	Heuristics Ontology-based Learning Machines Experimental Systems
AI	Artificial Intelligence
NLU	Natural Language Understanding
API	Application Programming Interface
ROA	Retail Omnichannel Analysis
POS	Point of Sale
IoT	Internet of Things
KPI	Key Performance Indicator
DDP	Data Discovery Platform
NLP	Natural Language Processing
SotA	State of the Art

Chapter 1

Introduction

This introductory chapter will present the company hosting this thesis and describe the surrounding context for its proposal. The project will also be briefly described along with its structure and the methodology chosen for its development.

1.1 Company Presentation

Wipro Limited (Western India Products Limited) is a global company that focuses on information technology, consulting and business process service, integrating cognitive computing, hyper-automation, robotics, cloud, analytics and emerging technologies in their solutions to provide their customers the ability to adapt to the digital world and be successful in their businesses. The following figure 1.1 displays Wipro's logo. Wipro's headquarters are located in Bangalore, India but it is a worldwide company recognized for its considerable portfolio, commitment to sustainability and corporate citizenship. Wipro's domain has a wide reach over the world operating in 6 continents, 62 different countries and having a total workforce that rises above 160,000 people and growing. Wipro has grasp on a broad range of competencies, respectively: Programming, Project Management, Solution Architecture, Business Consultancy, Analytics/DW/BI, Merchandising, Supply Chain Execution, Planning, Optimization, Warehouse and Store Operations, Finance, Data, Testing, Integration, Technical Services, and IT Service Management. The company has worked for several renowned international brands such as Makro, Debenhams, Morrisons and Primark. For its remarkable achievements, Wipro has been recognized with many awards over the years, including the award for being one of the world's most ethical companies, by the US-based Ethisphere Institute in 2017 for the 6th time in a row, and also the award for 2nd rank in the Newsweek 2012 Global 500 Green Companies, among many others.

1.1.1 The Spirit of Wipro

The Spirit of Wipro is the core of values of the company. It reflects the character that all the company's employees should nourish. These values are reflected in the company's behavior and are deeply rooted in the unchanging essence of Wipro. Moreover, it is also what the company

Figure 1.1: Wipro's Official Logo [1]

aspires to be, since reaching these ideals is a continuous and challenging journey. The spirit of Wipro is, then, a beacon, guiding and providing a clear sense of purpose to everyone, uniting people on a common goal.

The Spirit of Wipro is established on four values [10]:

- **Be passionate about clients' success**

  "We succeed when we make our clients successful. We collaborate to sharpen our insights and amplify this success. We execute with excellence. Always."

- **Be global and responsible**

  "We will be global in our thinking and our actions. We are responsible citizens of the world. We are energized by deep connectedness between people, ideas, communities and the environment."

- **Treat each person with respect**

  "We treat every human being with respect. We nurture an open environment where people are encouraged to learn, share and grow. We embrace diversity of thought, of cultures, and of people"

- **Unyielding integrity in everything we do**

  "Integrity is our core and is the basis of everything. It is about following the law, but it's more. It is about delivering on our commitments. It is about honesty and fairness in action. It is about being ethical beyond any doubt, in the toughest of circumstances."

1.1.2 Wipro's History in Portugal

Wipro's entrance in Portugal came through the acquisition of the company previously named Enabler. Enabler's logo can be seen in figure 1.2. Enabler was created in 1997 through a spin-off of Sonae/Continente, a leading Portuguese retailer. Two years later Enabler started co-developing, along with Sonae, the capabilities of Oracle ERP, and won its first international implementation

in Italy, for DeSpar. In 2000 the company began to make implementations across Latin America, the United States and Europe. The company furthered its internationalization with the opening of new offices in Braga, Reading, Paris, Dusseldorf, Curitiba and Minneapolis, having their Human Resources and Finance teams managing the processes for each correspondent country.

Figure 1.2: Enabler's Official Logo [2]

It was then in 2006 that Wipro decided to acquire Enabler, recognizing it for having a solid domain knowledge in retail and for being a leading implementer of the Oracle Retail ERP package. The primary office for Wipro in Portugal was inaugurated in 2008 at Maia, and is currently the only office the company has in Portugal. Since then, the workforce in Portugal has risen above 180 people, harboring consultants that have participated in more than 10 years experience in IT and solution implementation, and being able to speak up to 6 different languages. The Maia Office was, since then, classified as an Oracle Retail Center of Excellence due to its notorious implementations with the software suite over the years, being one of the few with this title on the company globally.

1.2 Context

In a growing digital world, companies have never felt more the need to acquire information such as today. Information is vital for companies to be able to make the best and most informed decisions regarding their business strategies. More and more companies are actively engaged in integrating Information Systems and Enterprise Resource Planning Softwares in their businesses in order to most effectively collect, store and manage their assets, including information. The present state of digital technology allows companies to collect and store great amounts of data almost effortlessly, specially if the company has a big customer pool. Companies that lag behind on these advances often tend to lose market to others that are ahead of them in this area. No one wants to be left behind, and as such, companies are always looking for new solutions that can provide an edge against their competitors. Since technological improvements are highly expensive and can only do so much to make the business progress, many companies are now focused on utilizing their data in the most efficient way possible, investing heavily in data mining processes, and also in data collection from various sources. [11]

The main way companies acquire their data is internally: from their customers, the transactions they make, their feedback, from their suppliers, from their processes; everything that happens inside the company can be monitored. Nonetheless, that isn't enough, because it's also important to get data from the outside, from surveys, questionnaires, public outreach, or even from buying

pools of data from other companies. One interesting and very appealing opportunity that cropped up in the later years comes from the popular, and always growing, online social networks. These platforms, with their massive number of users, hold an incredible amount of data that could potentially be harnessed by companies (within the legal and ethical boundaries, naturally) to improve their perception of the public trends and be able to respond to them more effectively. Social networks are widely used around the world by virtually any individual and have active users every day, posting, liking, sharing and commenting on a wide assortment of subjects. According to an article about Twitter activity [12], users in 2013 were posting every day, on average, over 500 million tweets. It is clear then that information that can be gathered in the social networks is plenty, diversified, and frequently renewed. This information comes with some drawbacks however; A large quantity of user written text can be considered "noise" or irrelevant, sometimes even impossible to understand. Furthermore, the large majority of useful text that can be collected is written in natural language, often containing abbreviations or slang, making it difficult for a software to crack their true meaning. These drawbacks provide a challenge for analyzing data extracted from social networks, considering this is a process that demands automation, for it is not viable to do it manually.

How can we overcome these difficulties? What technologies can aid us in collecting and classifying this information? Furthermore: How do we analyze the collected data? What information is considered relevant for retail companies and how to extract it? And finally: What is the best way to deliver this information to the retailers?

These are the question that this project aims to answer.

1.3 Motivation

Understanding the patterns of people's behavior is and will always be a difficult challenge. The ever changing trends of the world affect each particular individual in a different way, according to its personality and interests. It's fascinating because it goes to show how diverse people are and how each different choice affects the world in a small but significant way. Nonetheless, while most believe it will never be possible to get a fully accurate prediction of human behavior, it is thrilling to think that we can get pretty close to it in certain areas. Over the years this aspect has been studied and some have reached significant results in predicting patterns, particularly in the retail area, in the form of Demand Forecast. With the rise of the social networks people have been given the ability to freely express their sentiment over a wide assortment of subjects in a public environment, creating a continuous feed of potentially valuable information. Studying that information may provide insight to better understand the public's segmentation, their habits, trends and opinions, and how strongly they can be influenced. This will, most likely, bring the retailer one step closer to accurately predicting customer's actions and to having a greater understanding of their perception about the market, which is an exciting thought.

1.4 Project Description

The objective of this project, Customer Xperience, in its total length, is to collect information from social networks, analyze the data and extract useful information for retail, and finally, deliver that information to the retailer through a dashboard, in an integrated way with the Oracle Retail Suite. However, this project had the potential to be massively time consuming and laborious, and as such, not suitable for an individual thesis that is to be carried out in such a short length of time. Therefore, the project was divided into these three main stages: data collection, data analysis, and data delivery (dashboard). Each stage was considered as a separate thesis, and there was an effort to make each of these as independent as possible from each other in order to not entangle their work together and minimize dependencies. The first stage, data collection from social networks, was given to my colleague Ana Cunha; The second stage, data analysis and refining, was given to me, João Ribeiro; And the third stage, data presentation and dashboard, was given to my colleague Pedro Martins.

The present thesis will then focus on the second stage of this project, data analysis and refining. The data collected from the social networks will reach this stage slightly transformed, having already gone through a simple sentiment analysis in stage one. The objective in this second stage is to analyze this data and observe patterns and trends, helping the retailer to extract meaningful conclusions and insights to make strategic decisions in its business. Though the initial objective of this project was to comprise up to sixteen areas of retail analysis, due to problems with the tools' availability and effective work time (further discussed in the Project Development chapter), this thesis will focus on four areas of analysis: Store Performance Drivers, Demand Forecast, Demand Driven Replenishment, and Brand Engagement. For a better understanding, Figure 1.3 demonstrates the positioning of this thesis in the Customer Xperience project.

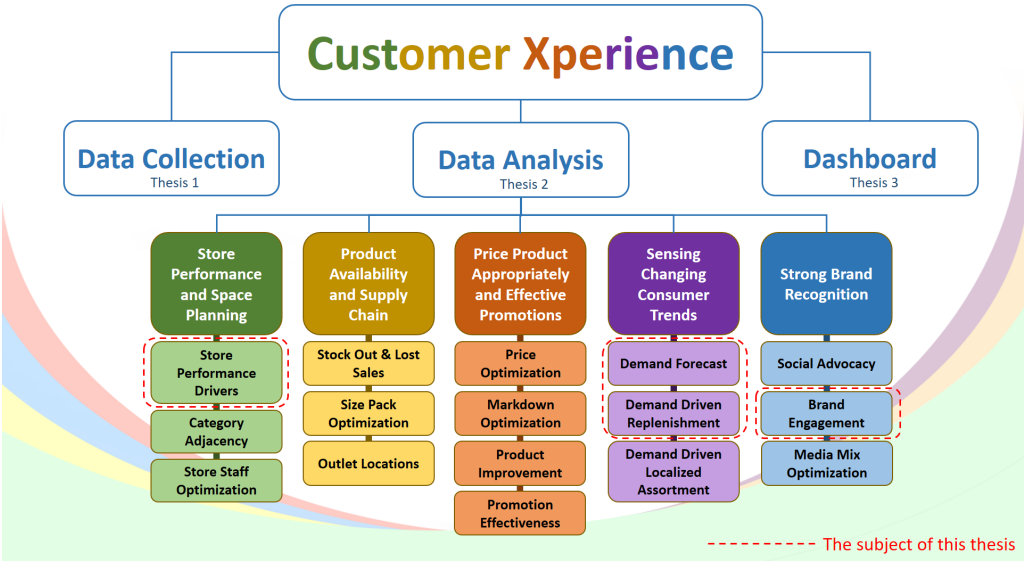


Figure 1.3: The Work Subject of the Present Thesis

1.5 Methodology

The methodology used in this assignment mostly consisted on theoretical research; This research comprised consulting trustworthy websites about the several different project subjects and studied tools, as well as the examination of some articles and books on the contents hereby approached. There were also different instances where help and knowledge from Wipro's personnel was sought through small meetings or presentations that provided better understanding of the project and the tools. Other than that, the Delphi Method was used once to acquire insight on the demand forecast subject, harnessing the knowledge and experience of a small group of company members.

1.6 Thesis Structure

The present dissertation follows the ensuing structure:

Chapter 1 - "Introduction" - This chapter introduces the company, the problem, the motivation, and the project's objectives.

Chapter 2 – "State of the Art" – In this chapter, the main technologies to be used in the project are introduced, as well as some of the available tools for each of them on the market.

Chapter 3 – "Demand Forecast" - This chapter develops the topic of demand forecast, exploring not only the traditional methods of forecasting but also the possibility of forecasting with social media data.

Chapter 4 - "Project Development" - This chapter comprises the project development stage since its beginning until its end, detailing how the analysis should be made for each specific use case.

Chapter 5 - "Conclusions and Future Work" - In this chapter, conclusions are drawn about the work done as well as the current state of the project and perspectives for future improvements.

Chapter 2

State of the Art

A state of the art research is an essential component of any project of considerable size such as this one. Being aware of the existing technologies and methods that are directly related to the project field is pivotal and will have a direct impact on its development. The present chapter will describe existing solutions and methods and deliver insight on already existing software that can perform those functions. Since most, if not all, of the presented solutions are paid, existing demos and tutorials are going to be used as basis to draw conclusions about them.

Although the focus of this thesis is centered in Retail Analytics, for the sake of context and in order to have a more complete overview of the available solutions, this state of the art will cover solutions for the three initial stages of the complete project: Data Collection, Sentiment Analysis, and Retail Analytics. This will provide a clear view of how data is collected, how it is treated, and then how it is analyzed and transformed into useful knowledge for the retailer, by the different solutions.

2.1 Data Collection (Web Crawling)

Data collection is a key element of this project. Since the data is meant to come from the social media, means must be provided for that extraction to occur in the most efficient and seamless way possible. The extraction of data from the web is commonly referred to as web crawling. Web crawling or data crawling is not to be confused with data scraping which is retrieving information from various sources that aren't exclusively the web.

A web crawler is a program that systematically browses through the web with an automated script. In the pages it visits, the crawler will identify keywords, the nature of the content, and the respective links. These web crawlers can be known by different names: bots, automatic indexers, and even robots. [13]

Web crawlers are particularly crucial for search engines. The crawler automatically gathers pages from the web and indexes them methodically to support the search engine queries. They can also help in validating HTML codes and links. When the user types a search query, the crawlers scan all the relevant web pages containing these words and turn it into an index. Web crawlers

will also visit the hyper-links present in the websites they visit, effectively fetching all the layers of websites related to the subject that's being searched. To maintain their indexes up to date, web crawlers usually monitor the web regularly, collecting new entries for each topic.[13]

De-duplication is an integral part of data crawling; it consists in detecting and removing all duplicates of collected data from the database. Its importance relies on keeping the servers as free of irrelevant data as possible and improving the quality of the data stored. [14]

Though web crawlers often relate to search engines, web crawlers aren't only capable to index web page links and references but can also download entire or partial pages [14], which is particularly useful to collect information (from social media pages, in the case of this project) to be analyzed.

For a web crawler to crawl on a social media page it is important to define what kind of data it is wished to extract, because the design of a crawler depends on it. This also affects the stopping condition during the crawling process. As an example, if only the topology of the social network is needed, the crawler simply needs to grasp the friend list in a page. If however, it needs to crawl the pages of a particular user domain, then it may need to extract information related to it such as country, age, or interests of the user present in the page. Considering the massive size of a social network, it is necessary to define an objective so that the crawler doesn't search through the whole thing. Usually, the process stops when a sufficient number of results is met. [15]

An extraction that looks for specific content requires format analysis and content analysis in the HTML content, because of the inherent complexity of the extraction pattern. Web crawlers need to be tailored for the pages they are trying to crawl on because of page layout and restrictions. For instance, in Facebook the amount of post that are displayed in a page directly depend on the size of the screen viewing it, the bigger the screen the more posts it can show; the crawler can then bypass this problem by simulating a much bigger screen size that's large enough to contain all posts it needs to read. On the other hand, the web crawler needs to understand what is the location of the content it seeks in the layout, so if it's looking for comments it needs to know what is the location of the comments and what may surround them. [15]

Most of the existing web crawlers don't provide an extensive description of their product features because they mostly do the same thing and are also highly customizable by the user, making the features variable from user to user. With this in mind, this section will mention some of the existing free solutions for web crawling.

2.1.1 Apache Nutch

Apache Nutch is an open source web crawler licensed by the Apache Software Foundation. This software can be used to aggregate data from the web, being usually utilized together with other Apache tools such as Hadoop. Though Nutch can run on a single machine, it gains significant strength from running in a Hadoop cluster. [13]

This software utilizes multiple web crawling algorithms to collect and save data. Apache provides a massive repository for data collected from several different sources through a framework called Apache Solr. [13]

2.1.2 Scrapy

Scrapy is a Python open source collaborative application framework for extracting structured data from websites which can be used for a wide range of useful applications, like data mining, information processing or historical archival. Though Scrapy was initially design to perform web scraping, it is also able to extract data using API's or as a regular web crawler.

As Web Medium [13] states, Scrapy is ideal for medium-sized scraping jobs. Its web crawling framework is quick, simple, collaborative, and open source. It also has the advantage of being able to accept new functionalities introduced by the user without its core being affected.

2.1.3 Storm-Crawler

Storm-Crawler is an open source software development kit for building web crawlers that runs on Apache Storm. Storm-crawler allows to build web crawlers that are scalable, resilient, low latency, easy to extend, and polite but yet efficient. [16]

This solution consists in a library and collection of resources that developers can use to assemble their own crawlers in a simple and straightforward way, often reusing components provided by the project, but allowing for customization and tweaking. In conjunction with this, Storm-Crawler also provides some external resources that can be reused in a project. [16]

Storm-Crawler is well suited to fetch URL's that come as streams but can also be perfectly used for large scale recursive crawls, specially if low latency is a requirement. [16] As pointed out by Web Medium [13] Storm-Crawler is the ultimate choice for a low-latency scalable web crawler. In comparison to Nutch, Storm-Crawler is able to fetch URL's as per the configurations of the user, but is not ready-to-use, as Nutch is.

2.1.4 Wipro Holmes - Data Discovery Platform - Collection

The Data Discovery Platform is a tool by Wipro that, amongst other things, is able to collect data from the web and effectively function as a web crawler. Its data collection component is built over a set of API's that is able to extract data from the web. In terms of social media, the DDP is able to extract comments from Facebook and Twitter, if provided with specific links, and return the data in the form of a .csv file containing all available information. Besides collecting the data, the DDP also performs a simple sentiment analysis in the comments collected, extracting general sentiment, categories and keywords.

2.2 Sentiment Analysis

Sentiment Analysis is the process of quantifying, qualifying and studying highly subjective information such as opinions and sentiments. Sentiment Analysis aims to combine various technologies, such as Natural Language Processing, Text Analysis, Computational Linguistics and Machine Learning in order to convert ambiguous and subjective data into intelligible and objective data. In its core, sentiment analysis is able to process a user written text and classify it into

positive or negative feedback. To do this, the text is pre-processed after collection, going through a "tokenization" process that aims to separately classify words into positive or negative, regarding different weights. Most of these systems utilize machine learning techniques, and, as such, they require to be trained with supervision beforehand in order to attain more accurate results. [17] But Sentiment analysis can be more than just classifying something into positive and negative; it could/should also be able to identify topics, entities, categories, syntax, or even identify features of an item/entity that the user is reviewing and classify their specific sentiment towards that feature. This brings more depth to the analysis which in turn will allow to draw more consistent conclusions by cross-checking its different aspects and relate them to each other.

Most of the information on the Internet is highly subjective, ambiguous and sometimes even meaningless, making it a borderline impossible task for an automated system to "decrypt" its meaning. To make it worse, online users often utilize natural language, abbreviations and slang when expressing themselves on the web, which makes the process of understanding the sentiment of their content much more challenging. The lack of an associated face expression when reading a text is another thing that can easily throw off humans when trying to understand underlying emotions on the text, let alone for an algorithm. Sentiment analysis aims to resolve these issues, by trying to consistently classify sentiment from virtually any text, in a way that can be automated.

In this section some of the existing solutions for sentiment analysis will be reviewed based on their demos and product presentations, since all of these solutions are paid and their tools aren't available for free testing.

2.2.1 Amazon Comprehend

Amazon Comprehend is Amazon's natural language processing service. Being a machine learning algorithm, it utilizes a pre-trained model to examine a given UTF-8 text document or set of documents to gather insights on its contents. Comprehend is being continually trained by teams of engineers and data scientists that aim to make the service more accurate and broad in its knowledge, and as such, users don't need to provide their own training data. [18]

Comprehend has the ability to identify up to 98 different languages, although its full functionalities of analysis only apply for English and Spanish languages. For those two, Comprehend is able to identify entities (people, places, brands, products, etc.) key phrases, and, of course, sentiment (positive, negative, mixed, neutral). This analysis is usually very quick, with process times rounding the hundreds of milliseconds. Besides this, Comprehend is also able to extract topics from large sets of documents for analysis or topic-based grouping. For these, the processing times are proportional to the size of the documents. [3]

Amazon suggests three different use cases for Comprehend [19]:

- Voice of Customer Analytics: Analysis of customer feedback in multiple channels such as emails, online comments, social media posts, and telephone transcriptions, in order to identify positive and negative experiences and use this data to improve products and services.

- Semantic Search: Obtain a better search experience by enabling the search engine to index key phrases, entities, and sentiment, allowing the search to focus on intent and context, rather than basic keywords.

- Knowledge Management and Discovery: Organize and categorize documents by topic for easy discovery and personalize content recommendations by associating other articles related to the same topic.

In figure 2.1 it is shown Comprehend's interface when processing a given text.

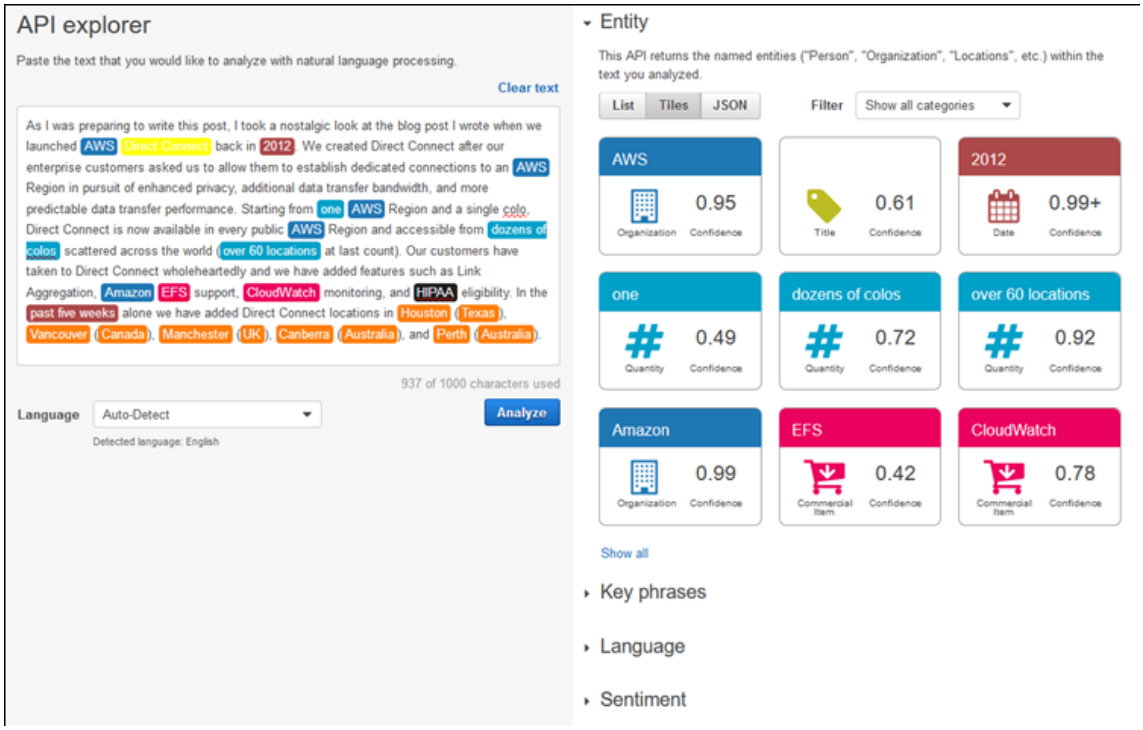


Figure 2.1: Amazon Comprehend Interface - Entities [3]

The algorithm detects the language of the text and then analyses it, identifying and highlighting the entities found on the text and categorizing them accordingly to their type. For every analysis, Comprehend assigns a confidence level, letting the user know how accurate the algorithm believes its answer is. The software can also identify key phrases that can range from a single word to five words or more. The key phrases have an appearance count on the document and a confidence level. Finally and foremost, Comprehend classifies the document on its overall sentiment as Neutral, Positive, Negative or Mixed, giving to each one a confidence level that functions as a percentage, indicating how much of each sentiment can be found in the document, as shown in figure 2.2. [3]

Although this interactive interface provides a better visual representation of the process, Amazon's Comprehend can perform the same functions on large batches of information; It can accept S3 buckets of data, which are objects created by Amazon's S3 cloud storage, and can perform the same functions mentioned before. Besides that, in this format, it can also create topic modeling

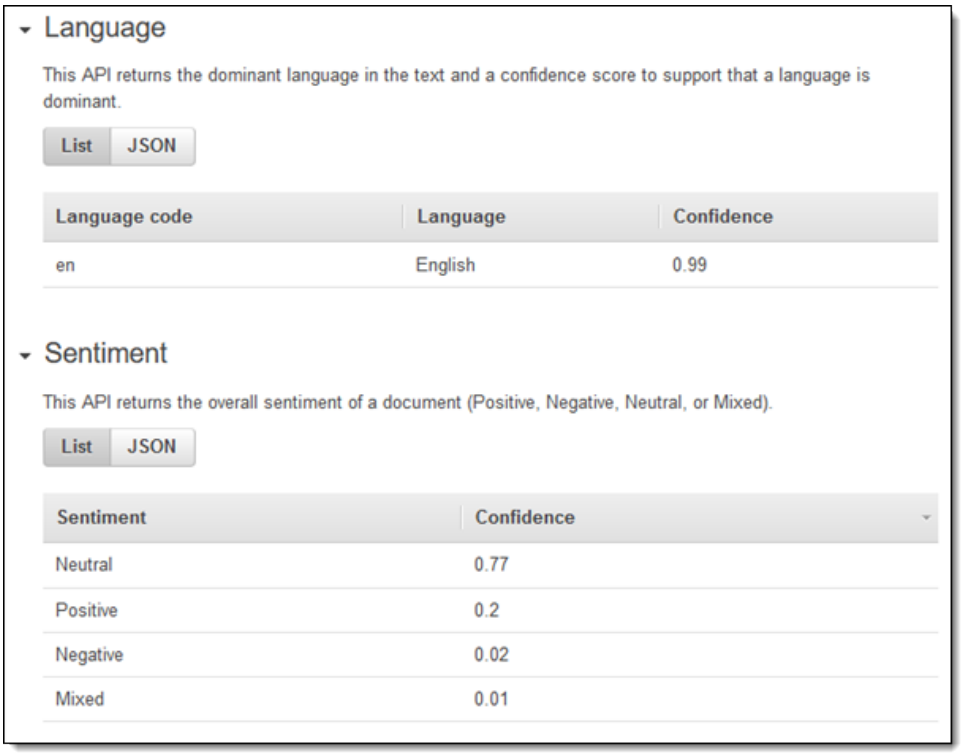


Figure 2.2: Amazon Comprehend Interface - Language and Sentiment [3]

jobs, organizing the most common topics in the batch and separating the documents and phrases in groups.

2.2.2 IBM Watson - Natural Language Understanding

Natural Language Understanding is IBM's tool for sentiment analysis which is integrated on their machine learning software Watson. NLU is built over a set of API's that can be used together to analyze semantic features of text input including categories , concepts, emotions, entities, keywords, meta-data, relations, semantic roles and sentiment. The API accepts text, HTML or even public URL's as inputs to perform the analysis. IBM's Natural Language Understanding provides features for thirteen different languages but currently only supports full functionalities for English. [20]

NLU's functionalities can be enhanced through customization. With IBM's Watson Knowledge Studio, users are able to create custom models that can identify custom entities and relations unique to their domain. [21]

IBM provides a demo of the API that can be tried for free with a few limitations. Figure 2.3 demonstrates NLU analyzing a small written text. Sentiment is measured on a scale from -1 to 1, where -1 is extremely negative sentiment, and 1 is extremely positive sentiment. It is also possible to isolate a phrase from the input text and perform a targeted sentiment analysis, by writing it on the text box bellow. [4]

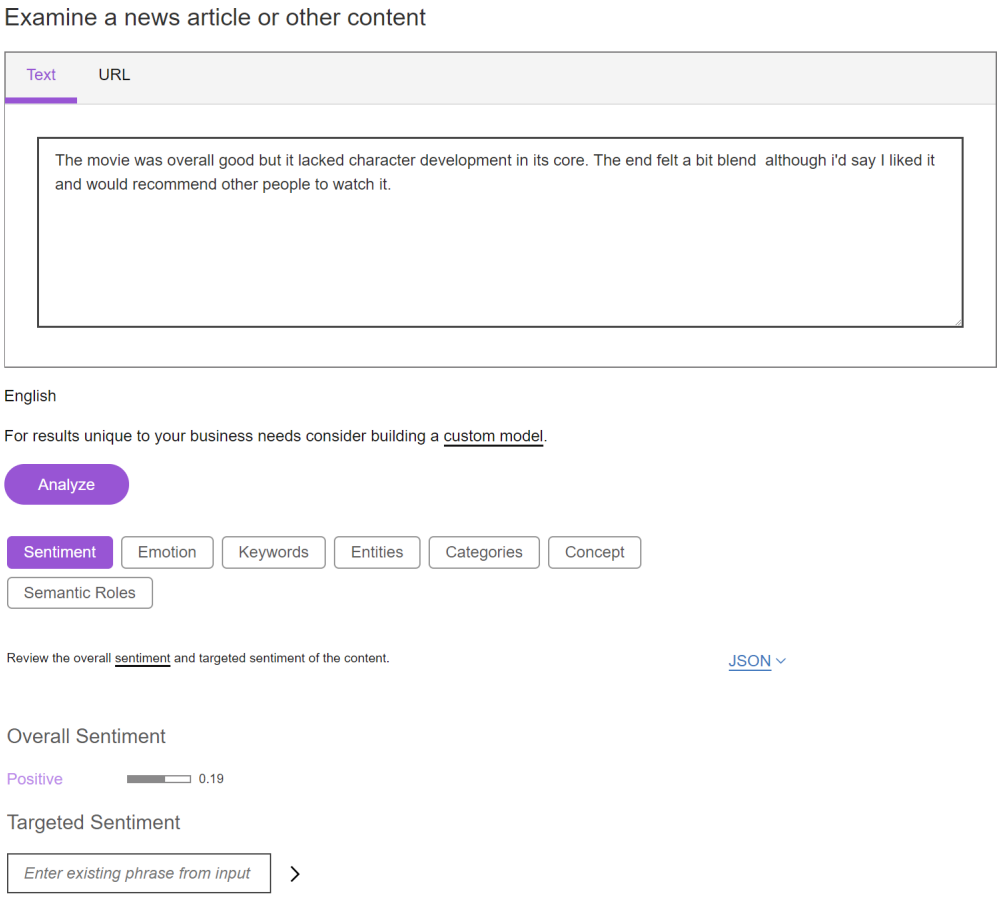


Figure 2.3: IBM Natural Language Understanding Demo - Sentiment Analysis [4]

NLU has a particularly interesting feature that is different from the other solutions: it is able to identify emotions present in the document. The algorithm distinguishes 5 distinct emotions: joy, anger, disgust, sadness and fear, then it gives to each a percentage, indicating how much of each emotion is present in the text. Just as sentiment, it is possible to do a targeted analysis for emotion. [4]

IBM's Natural Language Understanding is also able to perform other standard functions such as retrieve Keywords, Entities, Categories and Concepts from the document. It can also analyze the semantic roles of words in sentences, parsing them into subject, action and object form. [4]

All of the functionalities described above can be equally utilized when analyzing the content located in a specific URL, input by the user.

2.2.3 Google Cloud Platform - Natural Language

Natural Language is Google's Cloud Platform solution for sentiment analysis. It consists on an easy to use API that reveals the structure and meaning of text through powerful machine learning models. Among many other uses, Google's Natural Language claims to be particularly useful

on understanding sentiment about a given product on social media or parse intent from customer conversations, which is highly relevant for the target of this project. As would be expected, this tool is integrated on Google Cloud Platform making it easier to be used together with the other Cloud tools and data. [5]

Natural Language supports ten different languages, respectively: Chinese (Simplified and Traditional), English, French, German, Italian, Japanese, Korean, Portuguese, and Spanish. For all of these languages the tool appears to be able to do a complete analysis, utilizing all its features with more or less success, although it produces the best results for English language. Nevertheless, it is remarkable that it has such a wide range of supported languages for full functionalities compared to most other solutions. [22]

Google Cloud Natural Language has a demo available to test its functionalities on small user input texts. The API analyses the text returning entities found, sentiment, sentence syntax, and categories. Figure 2.4 demonstrates the demo interface, upon analyzing a text and recovering its entities.

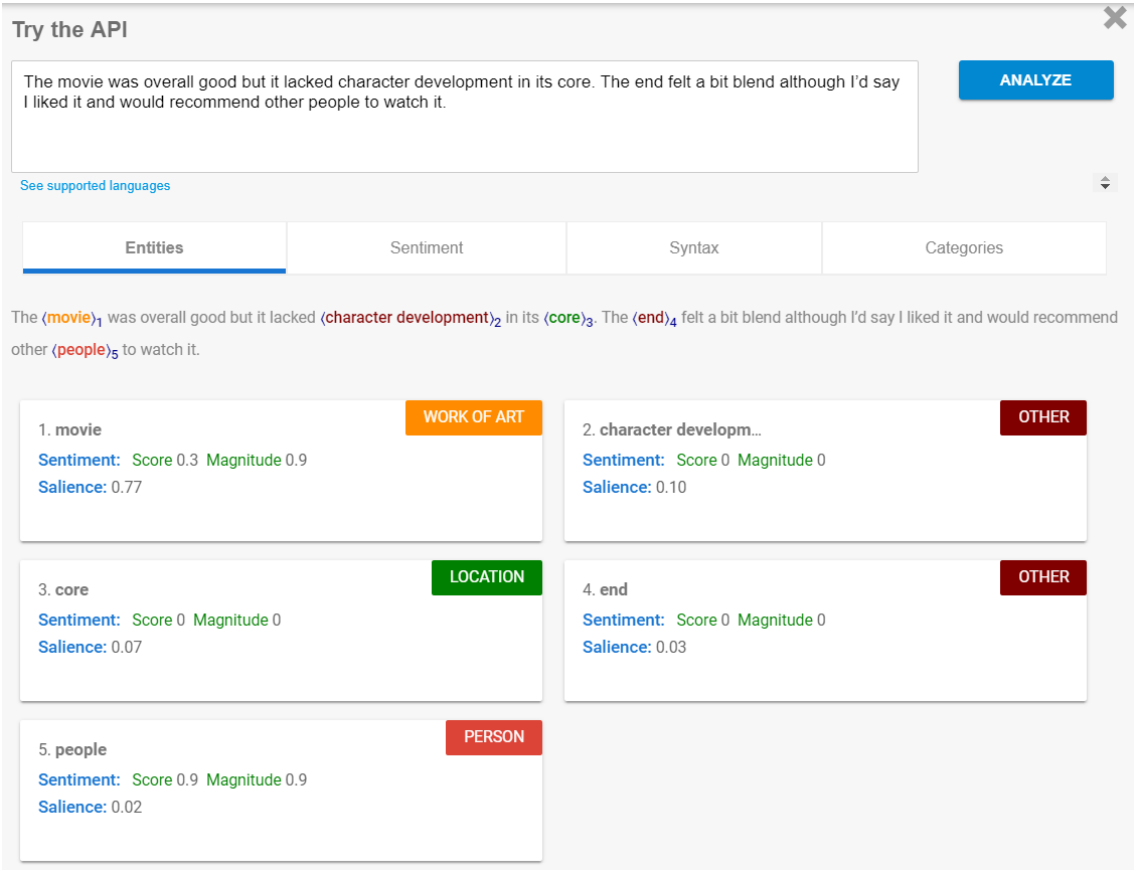


Figure 2.4: Google Cloud Storage Natural Language - Entity Identification [5]

The entities retrieved are classified on their category (ex: Work of Art, Location, Person, etc.) and a given sentiment is associated with them that has an associated magnitude, representing the algorithm's confidence in its classification. A salience level is also given to the entity, representing

its relevance in the text, and, furthermore, if found relevant, a Wikipedia article about the entity is associated with it for user convenience. [5]

Figure 2.5 shows the results of a the sentiment analysis made by the API. First, an overall sentiment score is given to the entire document. Each score is accompanied by a magnitude that represents the confidence level of the classification, as mentioned before. Then the algorithm separates the document in its individual sentences and performs the same sentiment analysis for each sentence.

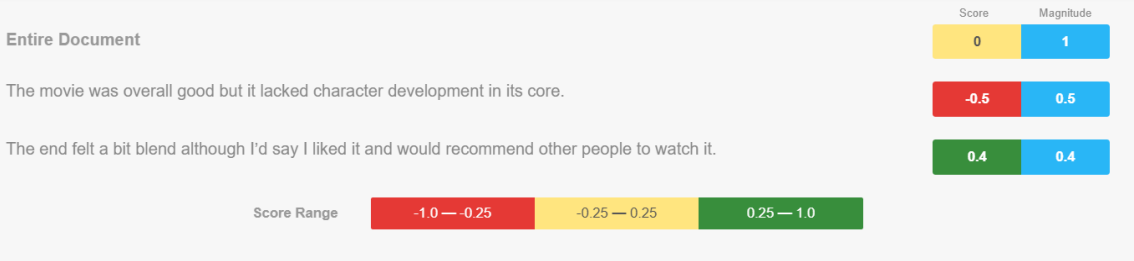


Figure 2.5: Google Cloud Storage Natural Language - Sentiment Analysis [5]

Also, categories can also be identified in the text, these are picked from a comprehensive list of hundreds of known content categories put together by Google's Cloud Platform, and for each category found in the text a confidence level is assigned. [23]

Lastly, Natural Language has the ability of analyzing with remarkable detail the syntax of sentences. It distinguishes Dependencies, Parce Label, Part of Speech, Lemma, and Morphology, providing a clear graphical view of the syntactical construction of the sentence with great detail. This feature is fairly unique (at least with this kind of detail) to Google's Natural Language, although for this particular project it doesn't convey any real advantage. [5]

Each one of these different kinds of analysis are counted as separate services and can be paid individually, though Google provides the features for free if a monthly quota of five thousand units of data is not exceeded.

2.2.4 Microsoft Azure Text Analytics

Text Analytics is a cloud based Microsoft Azure tool for sentiment analysis with three main functionalities: Sentiment Analysis, Key Phrase Extraction and Language Detection. It's a simple and limited tool, providing a small range of features for currently twelve different languages (English and Portuguese included). [24]

Microsoft Azure provides a demo of the tool's functionalities, as shown in figure 2.6. The API analyses the text, detecting its language with a confidence level given in percentage. It also recovers key phrases present in the document and lastly it calculates the general sentiment of the document. The sentiment score ranges from zero percent to one-hundred percent, from most negative to most positive, respectively.[6]

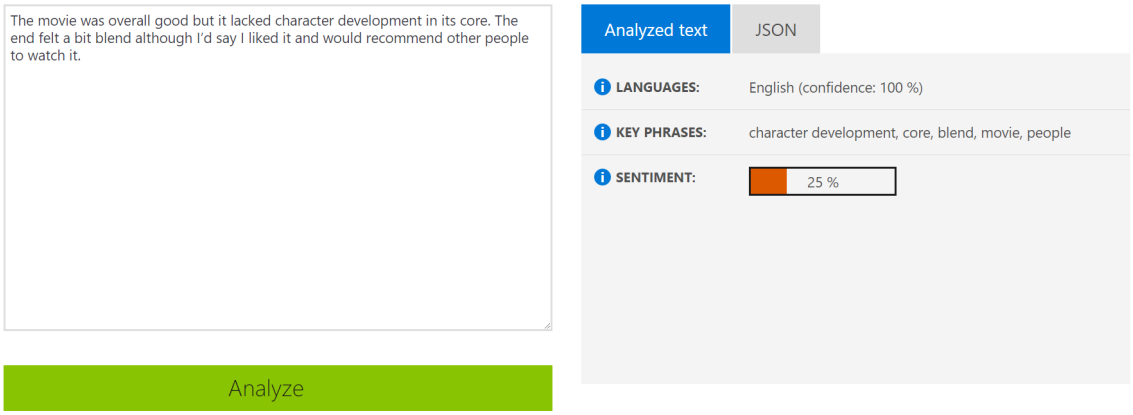


Figure 2.6: Microsoft Azure Text Analytics - Sentiment Analysis [6]

Microsoft Azure Text Analytics overall is a very limited tool, providing few features and shallow analysis, standing behind all other solutions reviewed so far. The lack of features such as entity and category recognition and the impossibility of a targeted sentiment analysis renders the viability of the tool very unlikely in comparison to the others.

2.2.5 Solution Comparison for Sentiment Analysis

The solutions reviewed are only but a sample of what is out in the market, but it wouldn't be time-viable for this project to review all of the existing solutions; As such, only some of the most known were chosen for this purpose.

It's possible to conclude that most solutions have a set of standard features shared between them, which are: sentiment analysis, key phrase/word identification, entity identification, and language identification. Although each does it in different ways, with more or less detail, these features are overall common to them. Google and IBM possess sentence syntax features, allowing for a more detailed analysis on semantics and morphology of the text, though that is fairly irrelevant for this project since the focus is to study customer sentiment and not to study how to extract it from text. Both Google and IBM's solutions are also able to identify categories from documents, which is extremely helpful for the retail analysis in understanding the overall content of a text without having to analyze it individually.

IBM's Natural Language Understanding brings out two interesting features, extracting emotions and concepts from text, which can be valuable data for the posterior retail analysis; Emotions can help to deepen the sentiment analysis and further understand customer opinion while the concepts can help in contextualization.

In conclusion, the choice for the best solution would be between Google's and IBM's. Amazon Comprehend is a good software but only conveys the standard features and no category identification. Azures Text Analytics is even worse being extremely basic and with no features at all. Google's Natural Language and IBM's Natural Language Understanding are fairly on par, with

the advantage being given to IBM for having the best features while its other standard functionalities are at about the same level as Google's.

2.3 Retail Analytics

Retail analytics are a series of processes that aim to analyze pools of data and extract from them valuable information for the retail business. The analysis made is different according to the type of insight they try to provide which can be diverse in the area of retail. So, generally, a retail analytics solution tries to help the retailer: forecasting demand, planning supply, evaluating promotions, optimizing inventory, understanding customer sentiment, deciding product assortment, amongst other things. Retail analysis goes beyond superficial data analysis such as data mining or data discovery, instead it refines the information so that it can be used as actionable insights for the retail industry. [25]

No matter how the analysis is made, it will always rely in pools of data to extract its conclusions. Traditionally, this data is historical company data, like purchases made by the customers, complaints, returns, markdowns, balance sheets, market survey reports, etc. However, Customer Xperience intends to innovate by only using data collected from the social media websites on its analysis, and therefore not relying on historical company data to generate its insights.

This project refrains from the traditional approach of retail analysis in the sense that it only utilizes customer feedback collected from the web to generate insights for the retailer. In chapter 4 there will be presented a series of use cases that were defined to be the focus of retail analysis for the project; However, due to time limitations, it was decided to approach in this thesis only four of those use cases of retail analysis:

- Demand Forecast - Predict, with agility, customer demand by product, location and time.
- Demand Driven Replenishment – Use demand forecast to optimize store replenishment strategies and maximize gross income.
- Store Performance Drivers - Use customer feedback to identify drivers and improve store performance in several areas.
- Brand Engagement – Understand public sentiment towards the brand in various aspects to improve general brand perception.

This section will serve to review some of the existing solutions for retail analysis. The reviews will have their basis, mostly, on the brands' presentation of their product features, given that none of these solutions are free nor they provide a free demo.

2.3.1 SAS - Retail Omnichannel Analytics

SAS Retail Omnichannel Analytics is a solution for retail analysis that is particularly oriented for Omnichannel, nonetheless it still provides insights in many areas that are common to other

retail analytics solutions. ROA is versed in five main areas of analysis: Omnichannel Analytics, Merchandising Intelligence, Customer Experience & Insight, Supply & Demand Planning, and IoT for Retail.

Omnichannel Analytics aims to help the retailer being agile at responding to customer solicitations across multiple channels of communication, through the offering of relevant offers, competitive prices, and the right merchandise; Correctly harnessing the data potential of all communication channels can lead to a better understanding of customer segmentation and demand. It can also help define communities of customers, allowing to tailor efforts, pricing and marketing in each store. [26]

Merchandising Intelligence allows the retailer to follow the evolution of customer preferences with relevant changes in the stores and merchandising. With the aim of aiding the retailer becoming more customer-centric and less product-centric and through the use of analytics, Merchandising Intelligence optimizes the assortment of merchandise and identifies missed opportunities. Another aspect of this analysis is the identification of customer demand by channel and by location, allowing to optimize inventory management and boost customer satisfaction. Through analysis of historical sales, Merchandising Intelligence can drill-down demand down to the store/size level, recommending ideal sized pack configurations. Lastly, Merchandising Intelligence can understand competitor's pricing through analytics and optimize revenue by devising optimal price strategies. [27]

Customer Experience & Insight's objective is to help the retailer rethink the way it uses customer data insights and analytics in order to meet the needs of a more mobile generation of customers that have more control over their purchasing process. This enables the retailer to respond to the customer needs instantly through the use of contextual listening and advanced data integration. In the end, retailers will be selling more by understanding what conveys value to the customer and the best way to deliver it; Also they will be spending less in customer acquisition with an Omnichannel marketing approach that utilizes predictive modeling. [28]

Supply & Demand Planning focuses specifically on customer demand and inventory optimization. It's advanced analysis allows the retailer to make small, medium or large-scale forecasts, and also to use proven what-if analysis to help answering big business questions. With the right forecasting and a correct analysis of sales and inventory data, it is also able to correctly manage inventory across suppliers, warehouses and stores, avoiding excesses and markdowns. [29]

IoT for Retail allows for a global, real-time view of customers across all devices, that can be used to adapt and anticipate customer needs. The tool will provide real-time insights of customer behavior, collecting and analyzing data from both online and offline channel. Combining all of this data with past customer experiences grants the ability to anticipate future needs with more precision, and act before the competition. [30]

Other solutions that SAS ROA conveys are Advanced Analytics, Customer Intelligence, Data Management Software, Integrating Merchandise Planning, and many others. [31]

[1]

2.3.2 Amazon - AWS for Retail

Amazon Web Services For Retail is a platform that provides a cloud computing infrastructure and several tools that are required for retailers. As Amazon states, with this modern and intelligent platform retailers are able to be agile and responsive towards customer requests and transform data into valuable and actionable insights for the business. [32]

AWS for Retail promotes four main benefits for retailers that use their software:

- ***"Investing in what Matters*** *- With AWS you invest less time and money on infrastructure. Instead, you can invest your scarce resources to drive innovation for your customers and efficiencies for your business."* [32]
- ***"Scale on Demand*** *- With AWS you can add and scale computing capacity quickly and efficiently. Spin up thousands of servers for peak seasons or major promotions then take them down just as quickly and easily."* [32]
- ***"Be Agile and Insightful*** *- With AWS you can quickly and cost effectively develop, test and launch new websites, mobile apps, and digital campaigns, evaluate their impact and course correct in real time."* [32]
- ***"Go Global in Minutes*** *- With AWS's multiple availability zones and geographic regions, deploying large-scale campaigns or expanding into new markets becomes easier and more cost efficient."* [32]

AWS for Retail incorporates solutions for three main areas:

Unified Commerce: With nowadays customers being "channel-agnostic", interacting with the retailer/store through different communication platforms, the company must be equipped with the means to process and store all of this information. AWS provides a wide range of reliable, scalable and secure data storage services that optimize the way this data can be stored in order to be most efficiently browsed and analyzed. [7] Figure 2.7 showcases a use case for this tool.

Customer Engagement: AWS for Retail tries to cater unique customer experiences to customers by using machine learning, facial recognition, and other AI capabilities, in order to improve brand loyalty. It aims to empower the retailer's ability to interact with customers, offering them relevant, personalized and differentiated experiences that will ultimately improve their engagement. [33]

Data Analytics: AWS for Retail comes equipped with the means to quickly and costlessly process and analyze great volumes of data from customers, businesses, and transactions. The application is able to analyze data from POS systems, replenishment and fulfillment models, loyalty programs, and customer databases, and from them extract business insights to empower decision making. [34]

[1] Omnichannel - The unification of all the communication channels and platforms through which the customers interact with the company in order to create a seamless customer experience.

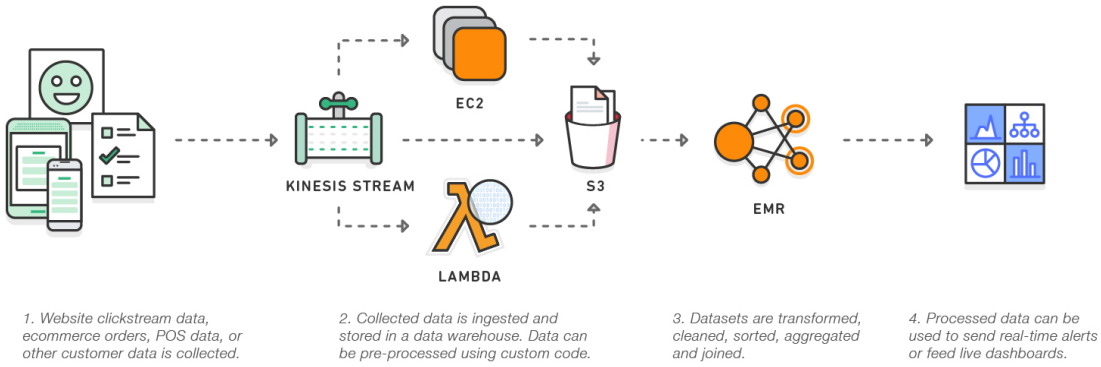


Figure 2.7: AWS for Retail - Use Case:Single View of the Customer [7]

2.3.3 IBM Watson - Analytics for Social Media

IBM Watson offers a very similar solution to the one proposed by the Customer Xperience project with Analytics for Social Media (IBM-ASM). This tool aims to help the user achieve a deep understanding of the customers and the market by analyzing data collected from a myriad of social media sources. [8]

IBM-ASM can extract customer customer sentiment from several sources, acquiring relevant insights about the brand, products, customers and competitors. The software automatically identifies topics, themes, relationships, patterns and trends on the data extracted, that can be valuable means for comparisons and decision making in the business. All of these insights are presented to the user in rich but comprehensible dashboards. [8]

With this tool, product research and development is made easier; By understanding what customers are saying about the products in the web the company is able to design products and services to directly match customer needs. [8]

IBM-ASM also focuses another interesting aspect of social media analytics which is acquiring information from competition. It can track consumer reaction to competitor's products and campaigns and provide insights for intelligent market positioning, improving the company's decision making. [8]

Overall, the main six features of IBM-ASM are:

- **Conversation Clustering** - By identifying relationships between words and pairing them, IBM-ASM clusters conversation topics that can give insights on common relationships, providing information for potential new opportunities. Figure 2.8 shows this functionality.
- **Advanced Text Analysis** - This functionality learns how users are behaving, what are they're interests, opinions and family structure. Knowing the target audience is essential in order to cater to their needs.
- **Guided Topic Conversation** - Helps removing irrelevant content from collected data and customize the search patterns to improve results.

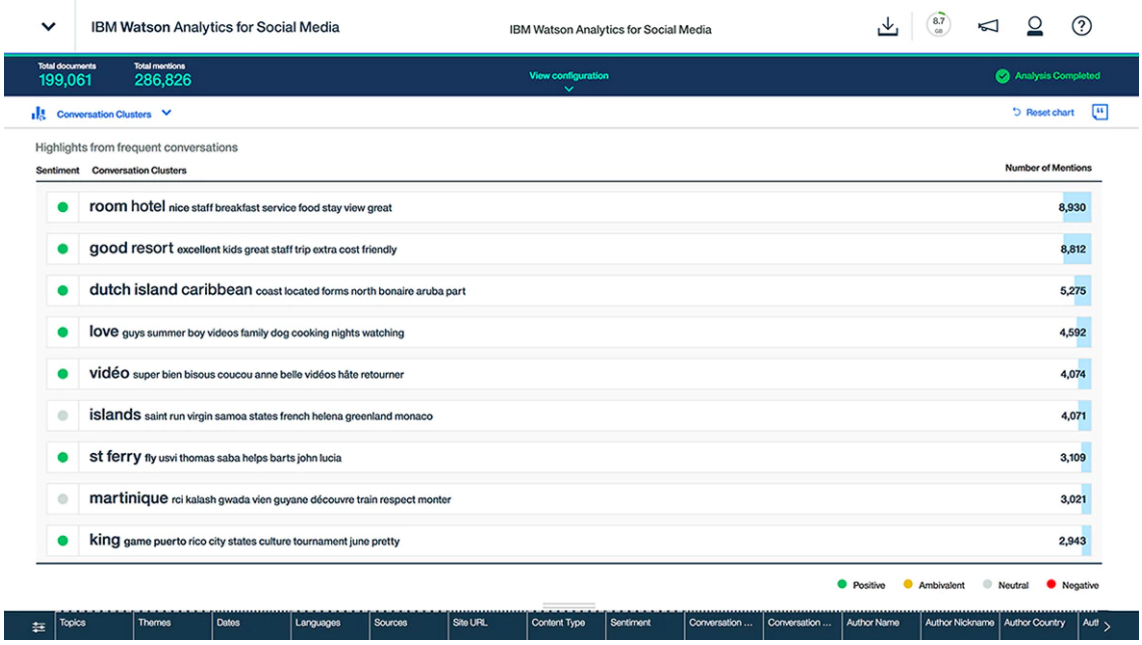


Figure 2.8: IBM Watson Analytics for Social Media - Conversation Clusters [8]

- **Influencer Analysis** - This analysis is meant to identify the most relevant influencers (i.e. the individuals that can influence the most public online) and study their actions, to help choose to leverage them in the company's favor or how to most effectively compete with them.

- **Context-based Sentiment Analysis** - This tool guides sentiment analysis by context, allowing for more targeted and meaningful results. The sentiment analysis is also refined to be able to identify nuances such as both a positive and negative sentiment in a sentences, rather than simply deeming them as neutral.

- **Pairing social data info with the user's own data** - Being able to mix the data collected with other sources of data from the company is crucial to obtain more accurate and meaningful results. IBM-ASM allows exactly that, for a full view of the product/business.

[8]

2.3.4 Wipro Holmes - Data Discovery Platform - Analytics

Data Discovery Platform (DDP) is Wipro's solution for data collection and data analysis based on Wipro Holmes, though this section will only cover the data analysis aspect of the platform. DDP has a vast range of tools to be used in various business domains such as finance, retail, health care, manufacturing, energy, and others. For this comparison, however, only its retail use cases are relevant.

The DDP use cases for retail comprise areas such as Market Basket Analysis, Store Performance, Customer Behavioral Engagement, Promotion Effectiveness Analysis, Targeted Campaigns, Customer Channel Preference, Assortment Assist, and Market Mix Model.

Market Basket Analysis focuses on determining products that customers are most likely to purchase and on determining relationships between product sales, to design better promotion or offers in the future. Store Performance categorizes and analyzes internal and external factors to weigh their relevance to each store's performance and better allocate resources. Customer Behavioral Engagement groups customers based on interest similarities in order to segment customers and better understand the profit potential for each market area. Targeted Campaigns observes customer trends to predict their behavior and creates promotional activities to appropriately answer that behavior. Customer Channel Preference tries to understand customer channel and product preference to improve interactions and distribute offers appropriately across all channels. Assortment Assist determines the key attributes that drive the sales of each product and utilizes that information for new product design. Promotion Effectiveness Analysis analyzes the historical performance of promotions over time to optimize promotion effectiveness based on past trends. Finally, Market Mix Model quantifies the sales impact of various marketing activities to understand the effectiveness of each marketing element for sales volume and to optimize the use of marketing resources.

Many of these cases are, overall, very similar to the ones being studied in this thesis. This is understandable since this project is meant to be an adapted version of this already existing tool, but relying only on social media data and with the intention of being integrated with the Oracle Retail suite.

2.3.5 Solution Comparison for Retail Analysis

As it was for web crawling and sentiment analysis, the solutions reviewed for retail analysis were selected from many that are available in the market. This selection was made considering that these were the most popular solutions and were also the ones that provided the most information on their tools. Other solutions like Google for Retail or SAP's POS Retail Analytics were considered but not included for lack of depth in their product descriptions. For this product comparison Wipro's DDP will be disregarded, since the opinion given on the company's tool could be, at the very least, considered bias.

As for the solutions reviewed, Amazon's AWS for Retail seemed to be the less optimal and also the less detailed in its description from the selection. The product presentation claimed that it had a solid data storage and organization, an optimized interaction between customer and retailer, and demand & supply insights. However it doesn't go into any detail on how it plans to deliver that, and the use cases showcased are not enough to formulate a sustained opinion.

SAS's Retail Omnichannel Analytics showed potential in the functionalities it proposed. The fact that it is oriented towards Omnichannel doesn't wander too much off the topic of this thesis since Omnichannel includes all communication platforms, including, in theory, the social media.

Also, the tools it disposes of aren't, at all, far from the Customer Xperience's goal, providing insight for demand forecast, demand-driven localized assortment, lost sales, size-pack optimization, price optimization, and brand engagement, all with a great focus on positive customer engagement through all platforms.

IBM's Analytics for Social Media showed the most potential from the solutions reviewed. In it's core, IBM-ASM is very similar to what Customer Xperience is trying to be, focusing on utilizing data from the social media to generate insights for retail. Tools that it provides like conversation clustering, advanced text analysis, guided topic conversation, and context-based sentiment analysis, can be phenomenal for a first stage of data analysis, right before it can be converted into insights. IBM-ASM also appears to have the means to engage on some specific use cases such as brand engagement, demand forecast, product improvement, price optimization, and marketing optimization; However it is never clear if it provides such specific analysis. Showing great strength and potential on the advanced analysis tools it disposes of, IBM-ASM's weakness appears to lay on capitalizing on those tools on specific and meaningful use cases (or at least it seems so, in their product presentation).

Chapter 3

Demand Forecast

Though this chapter could be considered as a part of the state of the art, it was made separate to provide more clarity, since the demand forecast subject is so complex and required a bigger study.

Demand forecast is the science of estimating future customer demand in order to optimize the supply, replenishment and pricing of products effectively. Forecasts are vital for every business organization as they are a crucial component to make management decisions, specially if these are for the long run. Forecasts provide the basis for any budgetary planning, as controlling business costs is essential to allow the company to make financial decisions. Moreover, forecasting is an important part of planning periodic production and operations decisions like selecting the right supplier, the right processes, planning capacity, facility layout, purchasing, production planning, scheduling and inventory management. [9] Choosing not to forecast customer demand may put a company in a dangerous spot, by wrongly producing or ordering more/less than they need which will inevitably lead to money loss. To supply its stores with the right amount of each product, the company needs to expect a certain customer demand so that it won't have less products than it needs (leading to loss by lost sales) nor it won't have more than it needs (leading to loss by stationary inventory and product markdown). The company can also regulate its prices appropriately by knowing customer demand; The more customers are looking for a given product, the more the store can exploit its price. On the other hand, if demand for a product is low, it is usually the right answer to reduce its price. These rules don't apply to every case, of course. [35]

Considering their main purpose, it is possible to divide forecasts into two main groups: strategic forecasts and tactical forecasts. Strategic forecasts are tailored for high-level demand analysis and are ideal for predicting demand in the longer run. One example could be: What is the demand expected for shampoo in the company's stores next year? These forecasts help setting strategy on how to meet demand in an aggregate sense. Tactical forecasts, on the other hand, are mostly adapted to help the firm operate its processes in a day-to-day basis. These forecasts generally help to make decisions for the next days or weeks, or months, providing an estimate for relative short term. The importance of tactical forecasts is that of being able to meet customer demand in the short term while maintaining an acceptable lead time and availability rate for the customer. [9]

It should be noted that there are no such thing as perfect forecasts. The massive amount of

variables and subjective factors in the business environment (which in turn are connected to more factors regarding the world that directly affect the business and the customers) make predicting demand with certainty virtually impossible. It is then more reasonable to learn how to use inaccurate forecasts to generate the best results, rather than trying to find the perfect forecasts; This doesn't mean, however, that efforts shouldn't be made to improve forecasting; These forecasts should be constantly reviewed to try to improve the forecasting models and methodologies and, thus, reduce uncertainty. [9]

To make their predictions, most of the traditional forecasting models rely on past data regarding past sales. This project, however, aims to only utilize social media data to produce the forecasts, and that comes as a challenge for there is little data on the matter. Nonetheless, it is important to know what are the existing forecasting methods and how they work in order to make an informed judgment on the matter. Traditionally, there are 4 different types of forecasting: [9]

- Time Series Analysis
- Causal Relationship Forecasting
- Qualitative Techniques
- Simulation Methods

Qualitative techniques are subjective estimates based on judgments and opinions, generally from experts. Time series analysis rely on the premise that future demand is directly related to data from past demand. Causal forecasting assumes that demand is directly affected by underlying factors in the environment. Finally, simulation models allow the forecaster to run through a range of assumptions about the condition of the forecast. Forecasts can also be classified into short, medium or long-term forecasts, depending on how far ahead they intend to predict demand. In business forecasting, "short" term usually refers to three months; "medium" term means from three months up to two years; and "long" term refers to more than to years. Short-term forecast would usually be used for tactical decisions such as replenishing inventory or scheduling employees in the near term and medium-term forecasts to plan a strategy for meeting demand over the next six months to a year and a half. Short-term forecasts are usually better for compensating random variation and quick variability in demand, allowing to set safety stock levels and to estimate peak loads in a service setting. Medium-term forecasts are particularly good at capturing seasonal effects, while long-term models detect general trends and are central to identify major turning points. [9]

This chapter will cover the existing forecasting methods as well as discuss typical forecasting errors and their sources. It will also approach some of the existing studies on forecasting using social media.

3.1 Time Series Analysis

Time Series forecasting techniques attempt to predict future demand based on past data. As an example, with the demand data from the 6 previous years, it would be possible to predict the demand

for the 7th year utilizing these methods. There are four models for time series forecasting, each should be used according to the situation regarding time horizon to forecast, data availability, accuracy required, budget, and availability of qualified personnel. [9] The four models are presented in table 3.1.

Table 3.1: A Guide to Selecting an Appropriate Forecasting Method [9]

Forecasting Method	Amount of Historical Data	Data Pattern	Forecast Horizon
Linear Regression	10 to 20 observations for seasonallyat least 5 observations per season	Stationary, trend, and seasonality	Short to Medium
Simple Moving Average	6 to 12 months, weekly data is often used	Data should be stationary (i.e. no trend or seasonality)	Short
Weighted Moving Average and Simple Exponential Smothing	5 to 10 observations needed to start	Data should be stationary	Short
Exponential Smoothing with Trend	5 to 10 observations needed to start	Stationary and trend	Short

To understand time series analysis it is also important to understand how a time series is constituted. A time series is a set of chronologically ordered data that can be decomposed in some or all of these demand components: trend, seasonal, cyclical, autocorrelation and random components. While it is, usually, relatively easy to identify the trend and the seasonal component, even without mathematical analysis (as shown in Figure 3.1), it is notably more difficult to identify cycles, autocorrelation and random components. To decompose time series methods such as the Least Squares Regression are used, on which this sub-chapter wont go into detail; [9] Instead it will focus on showcasing the different methods of time series analysis.

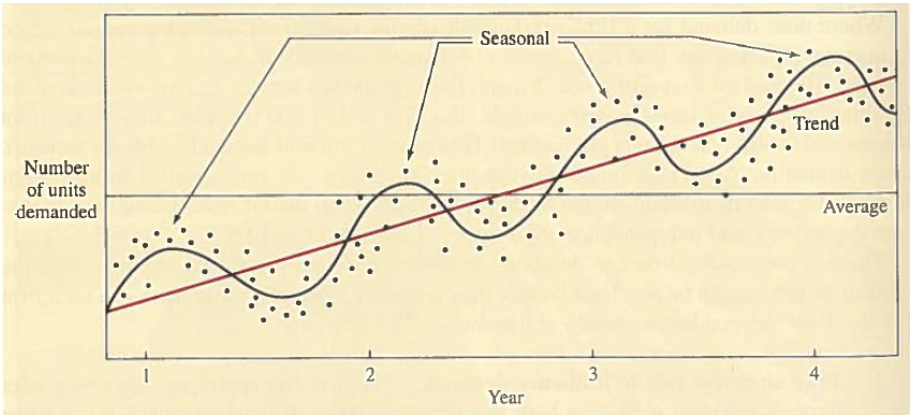


Figure 3.1: Historical Product Demand Consisting of a Growth Trend and Seasonal Demand [9]

3.1.1 Linear Regression Analysis

Regression is a functional relationship between two or more correlated variables. If this relationship forms a straight line when plotted, then it is a special class of regression called linear regression. Regression is used to predict a variable, given the other, considering their relationship. Linear regression lines are of the form Y = a + bX, given that 'X' is the independent variable, 'Y' is the dependent one, 'a' is the Y intercept, and 'b' is the slope. Linear regression is ideal for long-term forecasting of major events and planning but it has a major restriction; as its name implies, it assumes that past data and all future predictions follow a straight line. There is a workaround that's usually used with this method which is using shorter periods of time that can be approximately linear, reducing inaccuracy. [9]

Despite being presented within the theme of time series analysis, linear regression analysis is used for both time series and causal relationship methods; When the dependent value changes as a result of time, it is time series analysis, however, if it changes due to the change of another variable, then it is causal relationship. [9]

3.1.2 Simple Moving Average

The simple moving average is a method that can be used to remove the random fluctuations for forecasting when demand for a product isn't growing or declining rapidly and also doesn't have seasonal characteristics. The simple moving average consists in picking existing data from a period of time and calculating its average as a forecast for the following time segment. For example, if the period chosen is from January to May, the forecast for June will be the average of the months from January to May. However, once the data from the actual demand of June is acquired, the average will "move" including the sales from June into its calculation. [9]

As any other, this method has advantages and disadvantages: When choosing a longer period of time for the average, the effects of random elements are smoothed, which is generally a desired outcome; however if a trend exists in the data, the moving average will have the negative effect of lagging it. In conclusion while shorter time spans have more oscillation, they follow the trend more closely. On the other hand, longer time spans provides a smoother response but tend to lag the trend. [9]

3.1.3 Weighted Moving Average

Similar to the simple moving average, the weighted moving average calculates the moving average of several components but allows for weights to be placed on each element, providing that the sum of all weights is 1. [9]

Choosing the weights to assign to each element can be somewhat challenging and the simplest way to deal with it is by trial and error. As a general rule of thumb, the most recent period of time is the most relevant indicator of what comes ahead, and as such it should be assigned the most weight. It's easy to understand that the sales made on a store on the last month are going to

more truthfully help infer what is to be expected in the next month. If, however, there is a seasonal pattern in the data, weights should be given according to that pattern. [9]

The weighted moving average is overall superior to the simple moving average for having the advantage of being able to alter the influence of the different segments of past data. This gives more freedom to customize the forecast and gradually improve it, but at the same time it's more inconvenient and costly than other methods, such as exponential smoothing. [9]

3.1.4 Exponential Smoothing

In the forecast methods previously approached, as data moved forward in time and new pieces of data were added, older observations would be dropped for the calculation of the new forecast. Furthermore, the emphasis was put onto the newest pieces of data, as they have the most significant influence on the forecast than those on the distant past. Using this premise as a backbone for forecasting, it is concluded that data loses relevance as it becomes more distant in the past; This is where exponential smoothing comes in as a pertinent method of forecasting. [9]

Exponential smoothing gains its name for the way it calculates the weight of data on the time series. For each period, the oldest piece of data loses its relevance by $(1-\alpha)$, where 'α' is the smoothing variable defined between 0 and 1. So, for instance, if there are 'n' periods of time, the weight of the older piece of data will be calculated as $\alpha(1-\alpha)^n$. The smoothing constant can also be called the response rate because it regulates the speed of reaction to the differences between forecasts and the actual occurrences for the respective periods of time. The value of this constant is determined by the user considering the nature of the product or the stability of the business. For instance, if the demand of a product is overall stable, a small response rate is better, but if high variance in demand is expected, the response rate should be high. This means that α can only be optimally defined once it is known what type of demand variation the company is facing, which can take time to be observed. Also, demand does usually vary with time, and as such, the α should be revised continually. To track the α there are two methods: predetermined values of 0.2 or 0.8 are used for a small degree of error or a big degree of error respectively; computed values for α where an application keeps pace of the changes in demand to adjust it. [9]

Because of its overall good performance, exponential smoothing is the most used forecasting technique of all, and it is an integral part of the majority of all computerized forecasting softwares, being used to order inventory in retail, wholesale, and service companies. The success of this method relies on six major reasons [9]:

- Exponential models are very accurate.
- It is very easy to formulate an exponential model.
- It is easy for the user to understand how it works.
- It is easy to test the accuracy performance of the model.
- Computation needed is minimal.

- Digital storage is small due to low use of historical data.

One aspect that makes the model so easy to use is the fact that it only requires three pieces of data to formulate the next forecast: the most recent forecast, the actual demand that was verified on that period, and the smoothing constant alpha (α). The equation for a single exponential smoothing forecast is given by the following expression: $F_t = F_{t-1} + \alpha(A_{t-1} - F_{t-1})$, where F_t is the exponentially smoothed forecast period t, F_{t-1} is the exponentially smoothen forecast made for the previous period, A_{t-1} is the actual demand for the previous period, and α is the smoothing constant. [9]

3.2 Causal Relationship Forecasting

Causal relationship forecasting aims to use independent variables other than time to predict future demand. These independent variables should be leading (the opposite of lagging) indicators for the events that are being tried to predict to have any value for the forecast. As an example, the approach of Halloween will increase the number of costumes sold: This is a form of causal relationship, one event causes the other. In this case, it is possible to consistently know in advance when Halloween is happening in order to prepare for it, making it a good basis for forecasting. It is important to note that not all leading indicators are causal relationships but they often suggest the occurrence of other events, and as such should be taken into account. Taking the last example into account, an increase on candy sales can be a leading indicator that Halloween is approaching but it isn't a causal relationship. Finding the causal relationships for the events that one is trying to predict is the most important step to engage on a causal relationship forecast. [9]

3.2.1 Multiple Regression Analysis

Another, more complex, method of causal forecasting is the multiple regression analysis, which considers a number of variables along with their respective impacts in the forecast. This method of forecasting is particularly useful when a number of factors influence a variable of interest. [9] As the name suggests, multiple regression analysis is an extension of simple linear regressing. The dependent variable is the one that we are trying to forecast, and it is dependable of the independent variables, the ones used to produce the prediction. [36]

Although it can be quite difficult in terms of mathematical computation, the availability of computer software that can perform multiple regression analysis is a great convenience, relieving the need for monotonous and difficult manual calculations. [9]

3.3 Qualitative Techniques

Qualitative forecasting techniques usually utilize the expertize and experience of specialists, requiring a great deal of judgment. Although they rely on human intervention, and as such are

susceptible to some subjectivity, the processes generally used are well defined. The methods usually rely on the opinions of not one but several individuals, often from different hierarchy levels and departments, that are guided in order to reach an unanimous opinion. With this in mind, qualitative techniques are not even close to being just wild guesses or hunches coming from a group of experts, but are rather a structured and organized method of thought and decision-making. [9]

These methods are particularly useful in situations where information is scarce (for instance, a new product, a new costumer segment, a new region). When venturing into these new areas of business all information can be potentially useful to improve the accuracy of the forecast: similar products, customer segment habits, or even competitor's data in the field. [9]

3.3.1 Market Research

Market Researches are studies usually made by outside companies that mainly utilize surveys, questionnaires and interviews to understand customer product preferences, habits, and willingness/possibility to pay. These types of researches are mostly used to probe the public opinion about new product ideas, already existing products, or preference within a product range. [9]

3.3.2 Historical Analogy

The historical analogy method aims to use the data on existing similar product as a model to forecast. The comparison made is dependent on the type of product being it complementary, substitutable or competitive. The difference between an analogy and a causal relationship needs to be addressed: a causal relationship would be that the demand for keyboards is caused by the demand of desktop computers, while an analogy would be that the historical demand for desktop computers can influence the demand for laptop computers. Products that are in the same category and offer a similar kind of utility can be used to make analogies between them. [9]

3.3.3 Panel Consensus

Panel Consensus is a method based on the idea that various people can think better together than a single one; A group of individuals from different positions are then put together to develop a forecast that's accurate and reliable. In practice, the forecasts should be conducted through open meetings with free exchange of ideas from all the individuals involved, however, since there are employees from different hierarchy levels present in the room, the lower level employees will feel intimidated by the higher levels management, preventing them from expressing an opposing opinion, that may possibly be right. The Delphi method, which will be discussed next, was created to work around this issue. In some situations however, where forecasting decisions are set at a higher level of business, the panel is constituted solely by higher level of management personnel, in a decision process often called *executive judgment*. [9]

3.3.4 Delphi Method

As stated above, the panel consensus method has the underlying problem of having the opinions of higher-level personnel being more weighed than those of lower-level personnel, and the inherent fear of these last to express their true opinions for feeling threatened by higher level individuals. The Delphi method comes as a clever solution for the problem by concealing the identity of the individuals that are participating in the study, collecting all the answers anonymously and giving them back to all the participants along with a new set of questions each time. [9]

In its integrity, the Delphi method is composed by five main steps:

1. Select a group of knowledgeable participants from different areas.
2. Gather their forecasts and insights through anonymous questionnaires.
3. Summarize the results and redistribute them to all the participants along with the new set of questions.
4. Re-summarize, refining the forecasts in the process and developing new questions for the participants.
5. Repeat step 4 as many times as necessary, always providing the participants with the end results.

This method is generally able to provide good results after three rounds of questionnaires. The time each round takes to be completed depends on the difficulty of the questionnaire and on the response speed of the participants. [9]

3.4 Forecasting Errors

Demand forecasting is a complex task that has an inherent tendency to harbor a significant amount of errors. The amount of factors that influence the forecast and their intricacy in the system make an accurate modeling of the process an impossible task. It is then necessary to accept that all forecasts will contain a certain amount of error.

A forecasting error is the difference between the forecast value and the value of the actual occurrence. But an error isn't really an error as long as the forecast value is within the established confidence level, although common usage may define a difference as an error. [9]

3.4.1 Sources of Error

As mentioned above, in forecasting there are several possible sources for error. One of the most common and overlooked source for error ir mistakenly projecting past trends into the future. As an example, in regression analysis, statistical errors are deviations of observations from the regression line. To better limit the amount of unexplained error, a confidence band is usually attached to the regression line, providing a range on which the error is acceptable and usually explainable.

However, when the regression line is projected into the future, and the respective confidence band with it, the error may not be correctly defined by that confidence interval; This happens because this interval is being based on past data and may not hold for the projected data points. It is actually proven that these errors tend to be worse than those predicted from forecast models.

There are two classifications for errors in their source: bias or random. Bias errors result on the consistent repetition of a mistake made by the forecaster; Random errors are errors that can't be explained by the forecast model being used.

3.5 Forecasting Using Social Media

Only recently being an area of study, and with rising popularity, the hypothesis of harnessing the social media to predict the future is something that has already been explored in previous occasions. In fact, there are cases of institutions that already tried to explore the social media in their favor in this way and obtained significant results. In 2009, the United States Geological Survey began studying the potential inherent to predicting earthquakes in real time using social media[37]. This was based on the premise that word about the event spreads more quickly through the social networks than it spreads through the crust of the earth. Another similar study resulted in a system called EMBERS; this system was deployed in 2012 to monitor civilian dissatisfaction and turmoil in order to forecast events such as protests and riots [38]. By combining the data from social media and other non-social media data, EMBERS has been able to predict when, where, and even why a protest will take place.

Successful cases like these validate the possibility of forecasting using social media, and are important baselines for future works. Furthermore, an article from 2017 called "Using Social Media To Predict the Future: A Systematic Literature Review" [39], which is an extensive and thorough study on the validity of using social media to predict future events, finds strong evidence that supports the notion that social media can, in fact, be used to make accurate forecasts into the future.

3.5.1 Intricacies of Social Media Forecasting

To forecast any kind of event, there needs to be a direct causal link that can be identified beforehand. As explained in Causal Relationship Forecasting, an independent variable such as the appearance of rainclouds, will affect a dependent variable such as the number of umbrella sales. The issue with the social media scenario is that often causality between user actions and physical world events doesn't translate into causality between user actions and the social media, or vice-versa. This is explained by the fact that what is disclosed on the social media usually doesn't impact the real world directly and the real world doesn't impact the social media directly. What happens is that the real world interacts with the users and these last interact with one another and the social media. This leads to a three-way interaction that is not necessarily causal and that makes prediction more difficult as the gap between any two of the three elements increases. To be able to assess the possibility of forecasting, there would be a need to grasp the underlying mechanisms

that produce actual causal relationships between the real world and the social media and understand the discrepancy between how users interact with the real world and with the social media. [39]

Particular attention needs to be given, however, to big event manifestations on social media that may not be representative of a reliable causal relationship for a forecast. This is mostly because in some cases users represent an opinion or manifestation towards something while merely being spectators and having no practical influence on it. While their input can provide a good understanding of an ongoing event, it doesn't always reflect a direct impact on it.

The matter is further complicated by how users interact with each other. It may be rather difficult to obtain a true assessment of something that's being described by users on a social network because of how they influence each other. [39] An individual that may be completely ignorant on a certain matter may express an opinion on it based on the opinions of others with whom he directly relates. On the other hand there are certain people that have the habit of contradicting the common trend, despite of its basis. This bears the question as who to base a forecasting on: a single individual may have been influenced by a group, but the group itself can be composed of people that may have been influenced themselves, incurring on the common phenomena of group-polarization. The rule of thumb is that in most scenarios the bigger crowd is closer to the truth; Group-polarization is something to have in mind, though is something that's not controllable and that it's difficult to work around; Instead, one should try to use it in its favor, by understanding the mechanics of the phenomena to predict behaviors.

3.5.2 Social Media Forecasting in different areas

Research made in the article before mentioned [39], covered the five ares where social media forecasting has been most explored: Elections and politics, stocks and marketing, public health, threat detection, and user characteristics. Research on some of these areas provided relevant insight for the project at hand.

Study in the area of marketing shows that an accurate prediction (through social media) of the future success of a certain product in terms of sales is directly dependent on how well the social media is reflecting real world events. The same study also proves that advertisements that target users according to their personality traits improved click and follow rates by 66% and 87% respectively, which is highly significant. [39]

User characteristics is also something that is extremely relevant for any kind of forecast. Leveraging demographic information about users could be particularly important to be able to "de-bias" forecasting models by weighting each element according to known demographic relationships. Demographics is also a very important component to predict user habits since this highly correlates with them purchasing certain products. [39]

One of the most indicative demographic property is geo-location. However, the article "From Interest to Function: Location Estimation in Social Media" [40] indicates that less than 1% of tweets in their study contained geo-location tags, and that of US Facebook users in 2010, only 4% introduced their home address in a way that could be converted into latitude and longitude

coordinates. This is problematic because any use of geo-location as a basis for forecast is going to be biased since not all the users make this information available. The same study [40] tries to relate friendship connections on Facebook to geo-location based on the premise that friendships are less likely the bigger the distance is between the users. This opens the possibility to predict an otherwise unknown user location based on its connections; The more friends the user has, the more the hypothesis can be proven, and the accuracy of the forecast increased. Besides geo-location, two of the most studied demographic properties are age and gender [41], which are usually easy to infer on the social media. While age is something that most users want to hide or mislead, specially older users, it is something that is not hard to deduce through statistical algorithms.

3.5.3 Conclusions and Improvements

Though this presents an extremely abridged overview of the studies made around the subject, it demonstrates the validity of the social media as a mean to conduct forecasting. In almost all areas of study positive results were found, though the degree of success was heavily affected by a number of negative factors such as noisy data, data bias, lack of "generalizability", and difficulty incorporating domain-specific knowledge and theory. Major problem solving may require specific modeling for user biases, applying complex data-driven models, training on assorted data sources, and using domain-specific knowledge in the modeling process.[39]

An Internet study by the name of "Understanding the predictive power of social media" [42] states that using advanced techniques that filter social media data based on keywords is usually more successful in predictive tasks than using basic filters. The way these keywords are selected is also important: Of the studies this article reviewed, papers which selected keywords manually to filter data had a predictive accuracy of only 50%, on the other hand, every paper that utilized statistical algorithms to select keywords in an automated way, found consistently positive predictive results. This is not to say that simple keyword or hash-tag filtering could not be sufficient in some cases, but, overall, filtering is best done in a principled fashion. On a separate note, the same study infers that sentiment analysis techniques present somewhat bipolar results; In some cases they provide no benefit on the forecast, while on others they prove to be extremely valuable.

Chapter 4

Project Development

4.1 Initial Project Outline and Research

The Customer Xperience project, that originally was meant to be addressed by a single thesis, was divided into three thesis by Wipro. This decision came out of necessity to distribute the workload in a fair and realistic way, since there wasn't enough time to explore all the aspects of the project in the time of a single dissertation.

The initial concept for the three project blocks, as defined by Wipro, was the one shown in figure 4.1. These three blocks can be commonly called as the data collection block, the data analysis block, and the dashboard block, respectively. The present thesis addresses the second block of the project.

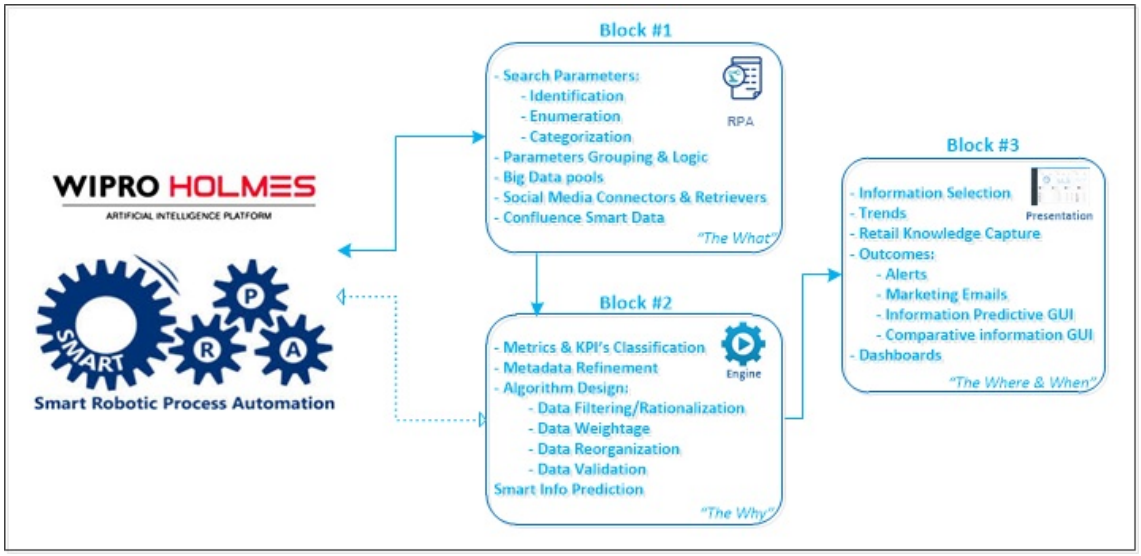


Figure 4.1: Wipro's Initial Project Definition

The descriptions for the blocks themselves were just a placeholder as the actual work plan was far from being defined yet; Since the beginning of the project it was strongly believed that

the data analysis would be highly dependent of HOLMES, Wipro's artificial intelligence platform. One part of the analysis that was thought to be crucial for the project was sentiment analysis, which was to be conducted by HOLMES in some way. In this initial phase, a strong emphasis was then put into researching about sentiment analysis, artificial intelligence, and to try to find any available data on HOLMES, which wasn't an easy task. It's important to note that despite HOLMES being property of Wipro, like many other tools, it was under the control of the Indian branch of the company (the main branch). Communication between branches isn't always easy due to time zone, and language barrier issues; this proved to be a problem with acquiring both the documentation on HOLMES and the access to the tool. The research on HOLMES was then halted until further notice, with no information on its functionalities and capabilities; this continued for one month. With this obstacle, the course of research was flawed; not having information on HOLMES meant there was no way of knowing what it could do both in terms of data collection and data analysis. A series of assumptions had to be made to proceed with the research: It was assumed that HOLMES was capable of collecting data from any social network; It was also assumed that HOLMES sentiment analysis was developed enough to include functionalities such as feature identification. As this project could potentially be integrated on the Oracle Retail Suite, used by Wipro in the majority of its projects, some time was also put into studying its overall architecture.

About half a month into the project another meeting was made to present a second concept defined by wipro for the first two project blocks. This concept is shown in figure 4.2

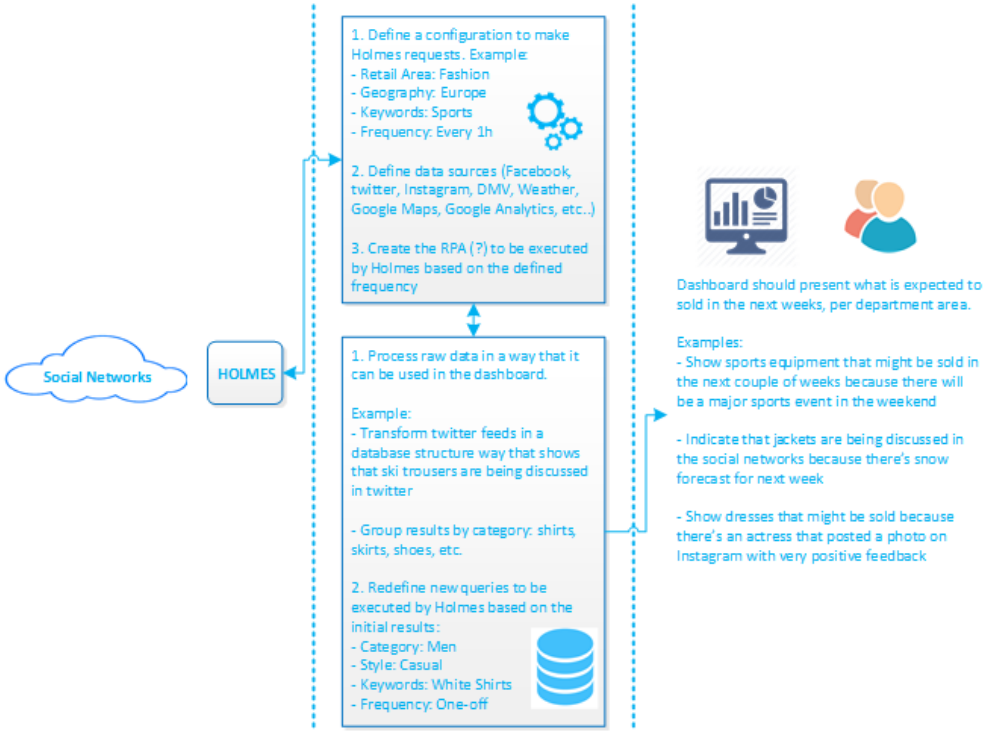


Figure 4.2: Wipro's Second Project Definition

With this second concept the work plan was also slightly altered. The focus was placed into

researching existing solutions for the technologies that were going to be used, specially data collection and sentiment analysis. Sentiment analysis and natural language processing continued to be a big focus of the research because it was thought to be an essential part of the data analysis process and there was no real notion of retail analysis yet. Since there was still no access to HOLMES or its documentation, the scope of the search was very limited.

One month into the project, a call was finally setup with representatives of the indian branch to assess the available tools and understand their functionalities. It was only then that it was made clear that HOLMES was a massive software with multiple functionalities, and that for this project only one of its tools was going to be used: the Data Discovery Platform (DDP). At that time not much information was disclosed about the DDP besides that it was going to be our main tool to collect data from the social networks; On the other hand an old project concept (figure 4.3) for Customer Xperience was brought to attention, which proved to be a valuable baseline to define a work line. This concept presented clearly all the pretended use cases for the analysis in each respective area of analysis.

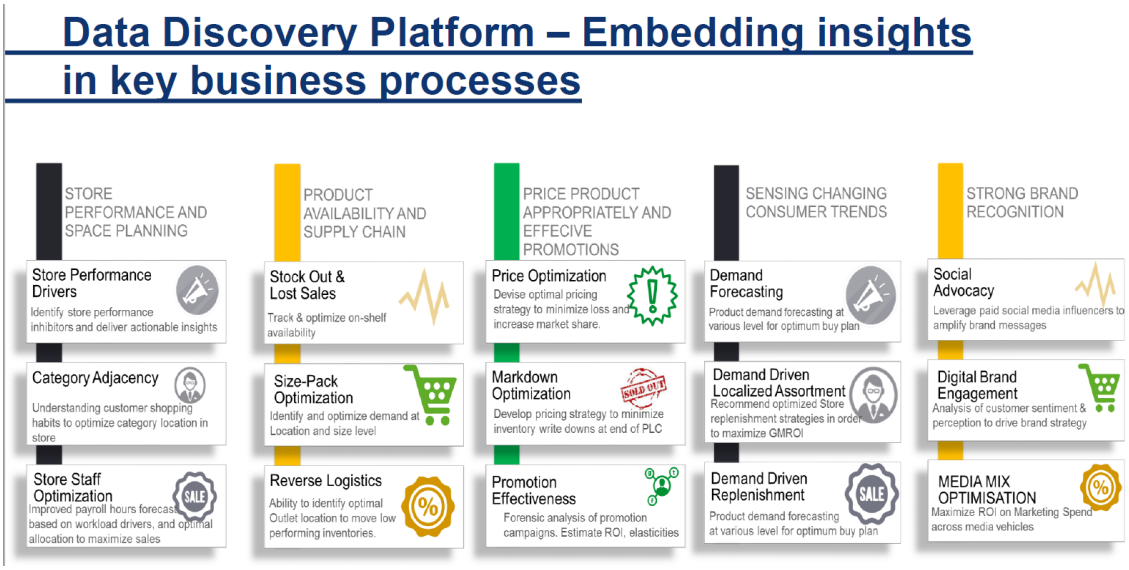


Figure 4.3: Wipro's Customer Xperience Older Project Concept

With this information, a new and more detailed project concept started to be defined, and a higher focus was put into researching about retail analysis since now it was clear where the project was headed. However it was only a month after this that a new call was setup with the Indian branch to have a proper presentation of the Data Discovery Platform. Only then the capabilities of the tool were disclosed as well as some of its documentation.

4.2 Final Project Outline and Research

With all the obstacles presented since the beginning of the project, the final project outline, as seen in figure 4.4, could only be defined two months into it. The full final project concept can be

viewed in appendix A.

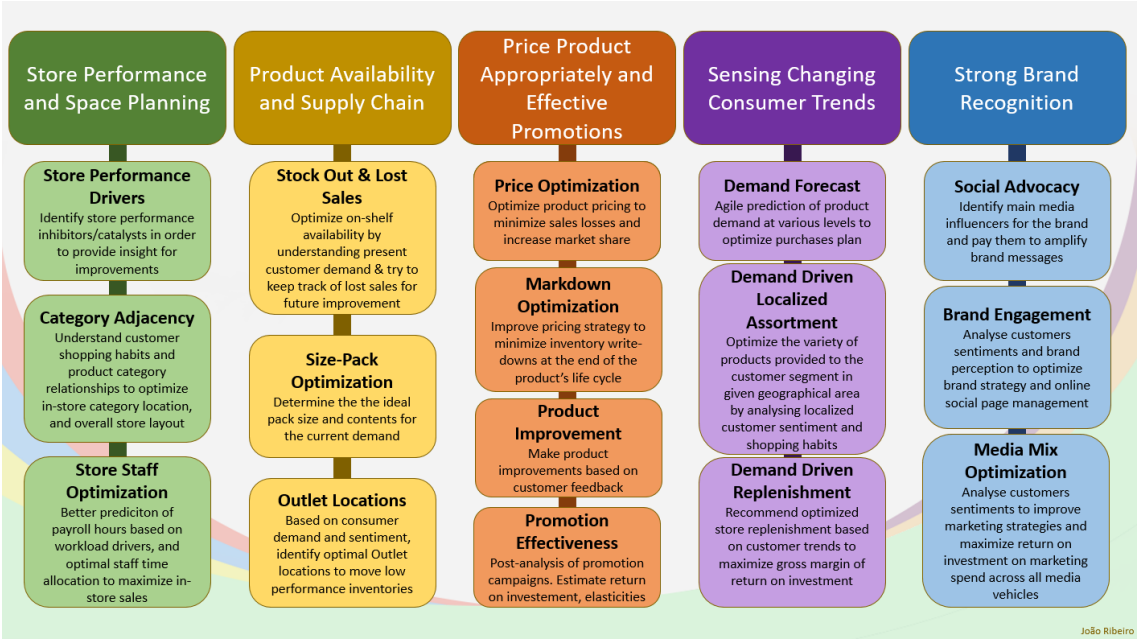


Figure 4.4: Customer Xperience Final Project Concept

At this time it was clear that there wasn't nearly enough time to finish it in its entirety during the thesis period; As such, it was defined that only a group of 4 use cases was to be chosen from the 15 to be developed at this time, and that implementation would be left to second plan for now. Four use cases were chosen: Demand Forecast, Demand Driven Replenishment, Store Performance Drivers, and Brand Engagement.

4.3 Incoming Batches of Data

To be able to analyze the data, first there was the need to extract it. This was of the responsibility of the first block of the project but it influenced the second block directly because of the necessity for data to work. To acquire the required data from the social media, it would only make sense to use Wipro's tool, the Data Discovery Platform (this was required upon the project). The DDP should not only be responsible for the collection of the data from the social media but also for a preliminary sentiment analysis that was to be performed in each instance of data collected.

After the presentation call of the DDP, connection was made with a representative of the Indian branch that was directly involved with the tool so that documentation could be shared and access could be granted to it. It was soon disclosed that it would be impossible to gain direct access to the tool for the project because of the sensitivity of the software and also due to confidentiality issues; In this sense, the Indian Branch would collect data weekly and send it in the form of a .csv file, avoiding all direct contact from the Maia branch with the tool. To demonstrate the concept practically, data would need to be collected and analyzed from the point of view of a

specific company. It was agreed that the data would be collected from the point of view of a company with which Wipro had worked in the past. This company, which will be henceforth called Company-X to protect its identity, is a big fashion retailer with more than 200 years of history. With headquarters in the United Kingdom, Company-X has stores operating in more than 26 different countries.

The first batch of data was received almost three months into the start of the project; A sample of that data file can be observed in appendix B. The Data Discovery Platform soon revealed some inherent issues; One of those is the way it conducts its data search; As an input to make its search, the DDP only accepted social page links, and collects all available data from those links. There was no possibility to search by keywords or categories, limiting the search in its scope. As such, the data was collected from the main social page links of Company-X on both Facebook and Twitter. The data came in large batches with more than 10000 lines, mostly composed by comments and tweets from users. Each instance of data (each comment) had associated with it several fields of data, as shown in figure 4.5.

Bank Name	Message Source	Friends(Twitter)	Listed (Twitter)	Like Count	Share Count
XXXX	XXXX	XXXX	XXXX	XXXX	XXXX

No of Comments (Facebook)	Gender	Parenthood Status	Marital Status	Keywords	Handle
XXXX	XXXX	XXXX	XXXX	XXXX	XXXX

Date_Created	Dimension	Category	Media	Sentiment	Popularity	Date	Social ID
XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX

Subscriber Name	Subscriber Location	State	Latitude	Longitude	Comment	Spell Corrected Text (Twitter)
XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX

Figure 4.5: DDP Data Fields

As data began arriving, a series of issues were detected. The first issue observed was that the "category" field sometimes returned incoherent categories for comments. Some examples can be seen in figure 4.6. Though this wasn't ideal it was expected that categorization wouldn't be perfect, and manual observation deemed that this didn't happen too often.

Category	Comment
Commendation	That's still not good enough. Your staff member was ill informed which prompted me to purchase a non vegan item. Veganism is my life, it's not a joke. I don't think I will be shopping at your store again.
Features	Hi Mikey. Yes we were dealing with the furniture and lighting team.
Finance	Hi, I was just wondering if you could tell me if a particular item is coming back in stock please? It's the "Tripp - Magenta II 'Holiday 5' large 4 wheel suitcase" thanks in advance :)
Grievances	No problem thank you

Figure 4.6: DDP Data - Incoherent Categories

Sentiment analysis was something more problematic, however; To begin with, the analysis was only able to classify text as positive, negative or neutral, with no kind of score or confidence level associated with it. Classifying the comments like this meant there was a lot less clarity concerning the true sentiment of the comments. A slightly positive or negative comment is considered as

100% positive or negative respectively, and a comment that has both positive and negative features is most likely considered as neutral; This type of analysis loses depth and value.

The most troublesome aspect of the sentiment analysis was that it frequently misclassified comments. It is normal to expect that the algorithm is not going to get it right 100% of the times, but in this case, manual observation indicated that misclassification was happing way more often than it should. Figure 4.7 shows a sample of comments being wrongly classified by the sentiment analysis.

Sentiment	Comment
Negative	Hi Grant I've DM'd the Debenhams account x
Positive	Your customer service is shocking ... told we'll refund you on your card but they actually charge you the money again and are in no way bothered or sorry and pretty much refuse to help you get your money back. I want this sorting and my money back. Extremely unhappy.
Neutral	Again. 2 tills open. Both agents flogging store cards. Queue out the door. Get a grip.
Negative	Thanks I have DM'd the card numbers to you!
Positive	Please can you give me a number to speak to an actual human being
Negative	Where can I find the ticket number? The invoice I have only has a returns number or do you mean the order number?Thanks
Negative	I accidentally ordered my order to collect in the wrong store by any chance could I change this? Thank you

Figure 4.7: DDP Data - Incoherent Sentiment Analysis

Another shortcoming inherent to the DDP is that there isn't a field that reefers comments to their original source. It is not possible to know what was being commented: a product post, a promotion post, an event post, or just the main page. Knowing this information would help make the analysis more accurate since some comments only make sense knowing what they are being directed at.

The biggest problem found in the data received from the DDP was comment length. If the comment collected for some reason exceeds a limit of 115 characters it will be cut off and the hyper-link for the comment will be put in front of it. This fault completely dismisses the possibility of automation of the analysis process as is, since many comments exceed that length and won't be completely revealed in the file. Without the whole comment there is no way of conducting a sensible analysis of it. Figure 4.8 showcases some comments where this issue happens (links were blurred for confidentiality). This problem doesn't affect however the sentiment analysis, since the algorithm analyses the whole comments before being cut.

Comment
Just the general number for travel money they've taken the money from my account but I got an error mes... https://t[redacted]
I just tried to set up a completely new account with a different email address and it would not let me a... https://t[redacted]
please help! I need to change my delivery address on my account to place a big order but your website do... https://[redacted]
The name on the order is my husband's who isn't on twitter I paid on my card from our joint account. Al... https://t[redacted]

Figure 4.8: DDP Data - Comment Length Issue

Some other minor issues where also found in the data received such as a lot of comments not having the gender and/or location of the user associated. This is comprehensible; As briefly discussed in Chapter 3, many users choose not to disclose that information. Also, this happens more often because of restrictions imposed by the social media page itself, not letting crawlers extract all of the information. Though this might be somewhat problematic because of potentially biasing the results, there really isn't a straightforward way to solve it without engaging on more complex methods of analysis.

Spell corrected text is another data field that shows signs of malfunctioning, often eliminating words it doesn't understand, wrongly correcting brands and people's names, and eliminating punctuation from sentences. This will be a much bigger issue once an effort is put into automating the analysis process.

Something that was noticed on the data batch was that several comments weren't coming from customers but actually from employees that were assisting customers on the brand's page. Since the focus of the analysis was to be centered on customer opinion, these comments had to be removed. A quick analysis proved that, fortunately, all of the comments made by employees were targeting specific customers (such as: @BenClarke). With a simple sorting and selection these comments were removed; These comments represented about 6% of the total sample.

Although there are several negative points associated with the Data Discovery Platform, it is worth mentioning the positive points as well. For one, the DDP is well able to collect great amounts of data from the links provided, effectively gathering all available data from comments on the pages. The platform is also able to identify the time the date that the comment was posted on as well as the hour, which is very useful for the analysis. The ability to convert the literal location into geographical coordinates is something that may also be useful when trying to use comments to localize demand. Another thing that can be valuable is the number of friends of the user commenting; This can give a rough estimate of how many people are viewing the comment. The same goes for the like/share count, which can allow to understand how the comments are getting visualizations and popularity. The keyword feature is very useful too; Though the DDP only provides single keywords for each comment, which inevitably affects its consistency, its still a fairly reliable way of understanding the main comment context.

These issues are nothing but normal on a first approach at collecting the raw data from the social media, and the treatment and cleaning of this data should be of the responsibility of the first block of the project; It won't be possible, however, for an already treated data to be used in this second block of the project due to both thesis for blocks 1 and 2 of the project being developed at the same time, and as such not being able to sync their work in a sequential way. Therefore, for this thesis, the data used is going to be the one directly collected from the Data Discovery Platform; This will inevitably lead to some changes in the final project outline, and to some assumptions being made whenever needed.

4.4 User Negativity Bias

When analyzing the data, it is important to take into account common crowd tendencies that can directly influence its results. Negativity bias is the human predisposition to give more weight or attention to negative experiences than to neutral or positive ones, even if the negative experiences are negligible. This phenomena has been verified in several studies. As an example, in behavioral economics, individuals gravitate towards options that avoid loss of money because losing money is always going to feel worse than winning is going to feel good, even if it is the same amount. [43] In the same sense, a customer will feel much more negative feelings for having a faulty product from a brand, than the positive feelings for the other several products purchased from that brand that have worked just fine.

Social psychology states that human impression is shown to form in a way that affects negative traits more heavily, implying humans focus a lot more on negative information than on positive information. This can translate to the web and even to retail, with users giving far more importance to negative details than to the overall good experience. It is common for people to say nothing when things run smoothly, according to their expectations. When users are engaged on a fluid experience they often don't even notice or comment on it, despite of all the work put into making it so by the company. On the other hand, if the experience doesn't match their expectations, they will immediately lash at it and memorize the incident long after [43]. As services get better, one would expect that customers would become more satisfied, but this is not the case; Researchers from the Nielsen Norman Group [43] observed that as websites improved over the years, users' satisfaction ratings remain the same. They hypothesize this happens for two reasons: for one, user experience failures weigh more than user experience successes; and two, people judge things in comparison to other things they used, in this case comparing to other sites they've visited, always finding something that's better than the other in a certain area.

An article [44] that comprises an extensive study on negativity bias, theorizes that this phenomena occurs based on four major factors: Negative entities are stronger than equivalent positive entities; Negativity of events grows more rapidly with the approach to them in space or time than the positivity of events does; Entities that combine positive and negative facets are more negative than what they should be according to the algebraic sum of the individual facets; Negative entities are more varied than positive ones and urge a wider range of responses and reactions.

The same study [44], highlights the importance of empathic interactions in this area. It is far easier for an individual to feel empathy for someone experiencing something negative than for someone that is experiencing something positive. Linguistic terminology further supports this claim as there are many expressions that denote an empathic response to other people's distress like pity, sympathy, and compassion, but there are no English terms that denote a positive empathic response such as "I'm feeling well at the sight of another's happiness". As R.Thompson stated in his book[45]:

> "There is reason to believe ... that adults as well as young children are more likely to respond empathetically to salient expressions of negative emotions in others.

> The hypothesized functions of empathic arousal in human adaptation enlist empathy primarily in response to others' distress cues." (p.139)

These researches imply that not only people are generally more susceptible to negative experiences and therefore more willing to complain about it, but also that they can influence each other through negative empathy. This explains why the sentiment analysis indicates that 60% of all the DDP collected comments from Twitter and Facebook are negative, only 8% are positive comments, and the remaining 32% are neutral. People are more prone to comment on a brand's page to complain or ask for help on a negative product experience than to comment to commend the brand in any way. Considering that this can be "contagious", double attention should be given to these percentages, for they may not represent the reality of user satisfaction.

One suggestion made is to include "likes" into the equation to obtain a more accurate view of user sentiment. With the current data collection method, the amount of likes directed at the original post/comment or even to the page itself are not extracted. This information could be valuable in understanding true customer sentiment. For instance, a post about a certain product "Y" can have 70% of its 50 comments being negative, but then have 450 likes. What does this tell us? That potentially 450 or more people enjoyed the product silently, left a like and moved on, while 35 had problems with it and complained in the comment section. As an anonymous customer from Company-X's wrote on their social page: *"(...) People are too quick to criticize."*

The existence of a ratio that would compare likes and positive comments with negative comments would greatly help mitigate the negativity bias in the data and obtain truer results. Other suggestions may be the use of different weights to classify negative data, giving more weight to complaints that are more frequent and less weight to isolated cases; alongside with that, negative data could be overall weighted less than positive data, though it would be difficult to find the right numbers to enforce that.

4.5 Preliminary Analysis of the Data

As data arrives from the DDP a first analysis can be made to identify general trends, even before dividing it by the different KPI's. First, to understand the distribution of users across both channels, the total comments coming from each of the social networks should be measured, as shown in figure 4.9.

This revealed that a crushing 95% of the comments were coming from Twitter, and only 5% from Facebook. This information can be valuable to help the brand know what social networks are the best to focus its marketing on; In this case, there is little value on investing resource to represent the brand on Facebook, since the amount of the brand's customers that use it is so small.

It might also be interesting to see how the general user sentiment progresses in time; this can give the retailer insight on the existence of underlying problems with the brand at a certain time; it can also help identify specific customer sentiment trends in certain periods of the year/month/week. For this, the time stamp on each comment is used to create a progression. Figure 4.10 shows a graph for the general sentiment progression in time for the data collected.

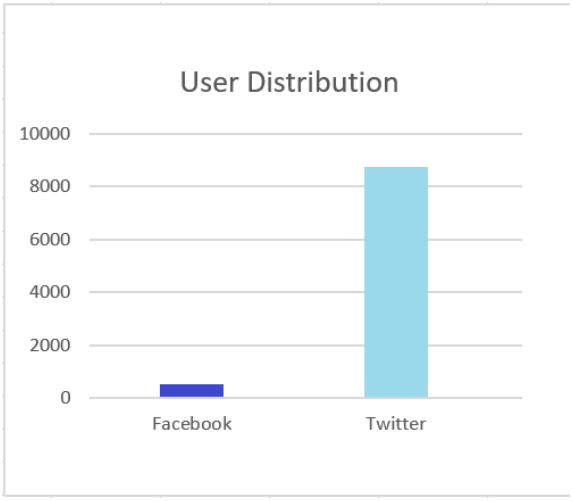


Figure 4.9: Distribution of Users Across Social Networks

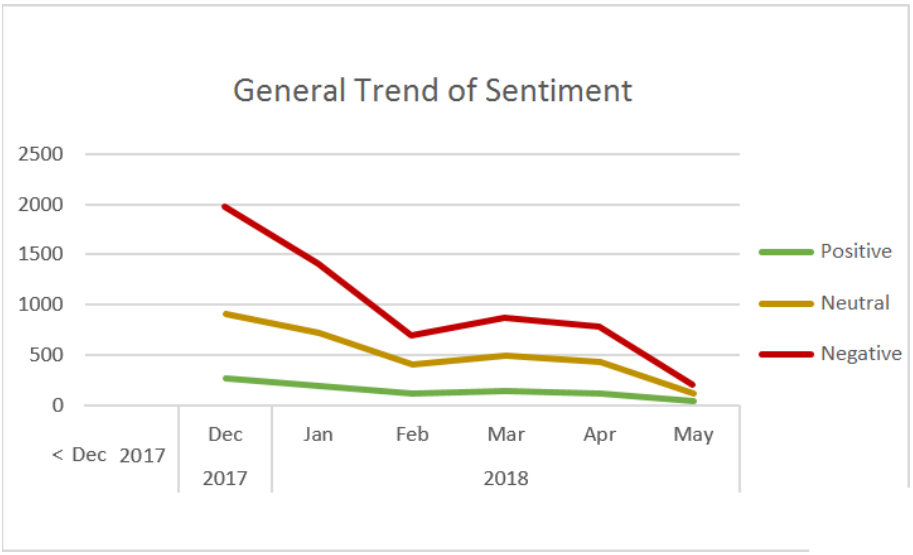


Figure 4.10: General Sentiment Trend of the Collected Data

Observing how the total amount of comments varies monthly can provide not only a sense of how engaged the users are with the brand and how well the social network is functioning as a communication channel, but also help identify peak periods of activity in the year. The figure 4.11 shows the number of comments over the last months.

Also beneficial is to observe how a series of factors behave weekly. For one, noticing how much certain categories are cropping up is key to understanding what is the subject of the user's comments lately. Figure 4.12 is the graph for total category count in the last months. In the same sense, total weekly count for keywords indicates what are the main topics of opinion for users in a more detailed sense, potentially indicating specific products or problems being mentioned by them.

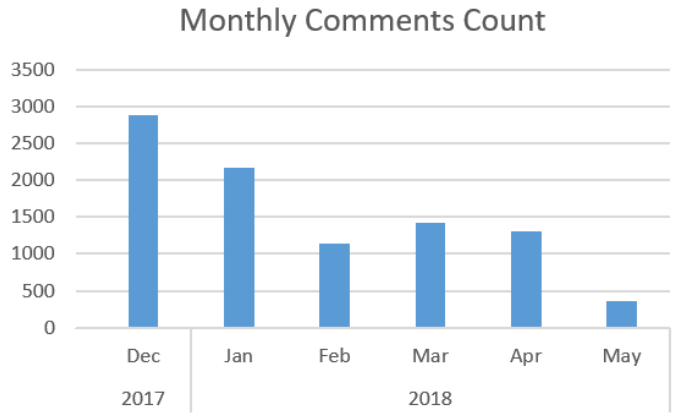


Figure 4.11: Monthly Comment Count for the Collected Data

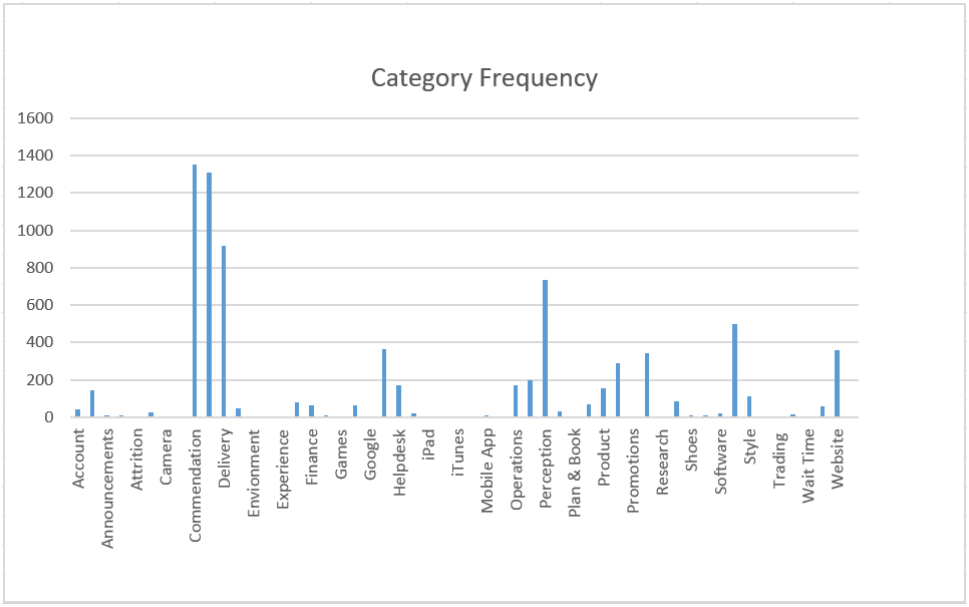


Figure 4.12: Category Count for the Previous Months

More demographical data should be followed in a continuous sense such as customer sex and location. As discussed before, users frequently hide this data or purposefully falsify it making the process of analysis more difficult. In the case of location, besides the users that don't provide their location at all, there are also those who just write nonsense or humorous text in their location, which calls for a filter that can remove these entries. Besides that, it was observed in the collected data that users often write the same location in different ways making it difficult to group them automatically (i.e. some users wrote Aberdeen, while others wrote Aberdeen Scotland, even though its the same location). To successfully work around these issues, either we refrain of using the

written location and only use latitude and longitude coordinates (which makes data far harder to read and understand on a dashboard), or a filter needs to be applied that can remove the gibberish and group locations that are the same but written in different ways.

Theoretically, all of these elements could be measured over time: yearly, monthly, weekly, or even daily. The data could then be drilled down accordingly to the needs of the retailer, observing data in the scale it considers to be more relevant for the given situation.

4.6 Store Performance Drivers

Store Performance Drivers is the use case for Customer Xperience with the objective of identifying factors that directly influence store performance. Using the social media to improve store performance means that it is only possible to measure things with each the customers interact directly. Therefore, the focus will be in anything that is directly associated with the store front itself and that can influence users in their opinion about the store. Store presentation, organization, layout, and ambiance, or employee presentation, sympathy and helpfulness would be factors that can directly influence store performance to the customer. The goal is to utilize user comments to understand what is their opinion about a specific store or stores and provide insights that aim to optimize a series of KPI's related to the store performance.

4.6.1 Store Performance KPI's

In order to define how the data analysis is going to be conducted it is important to define what KPI's are trying to be measured for this particular use case. Regarding store performance, traditional KPI's revolve around profit margins, store traffic and customer satisfaction; With the limitation of using social media data only, these KPI's need to be adapted to what is possible to extract from user comments. With this in mind, a sensible group of KPI's was chosen to measure the overall store performance. Three major KPI's were selected: store aspect, employee quality, cash-out quality.

NOTE: The chosen KPI's are of general use, and serve as a standard template; Each retailer is different, with distinct needs, and might choose different KPI's depending on its business strategy.

Store aspect measures the customer opinion about the overall store looks. Even before the customer enters the store, the first thing it notices is its aspect and design; this can be a compelling factor to attract them in. Once inside the store there is other factors to take into account. Store presentation is one that is most obviously visible; If there are unfolded clothes everywhere, empty hangers laying around or even a floor swarmed with clothing tags the customer is going to feel uncomfortable and won't want to stay in the store. Finally, the last factor to take into account is store layout; It's important for the customer that the store layout is coherent and simple to understand, making the look for the desired articles easy and effortless. Overall, the major store aspect KPI can be drilled down into four smaller ones: design, presentation, and store layout.

The **Employee quality** KPI aims to evaluate customer opinion regarding the employees of the store. This can be drilled down into five smaller KPI's: sympathy, helpfulness, availability

and nuisance. When it comes to employees, presentation is an important factor to take into account, The first thing a customer sees in an employee is its appearance; A presentable employee is more approachable and provides more assurance. While still in the category of presentation, an emphasis should be put onto clarity; Sometimes customers have a hard time identifying who's an employee in a crowded store because their outfit just blends too much into the surroundings or because it doesn't distinguishes enough from regular customer clothing. This should never happen, an employee should be easily identifiable and never mistaken by a customer. Sympathy is also a quality that every employee should convey. Customers that feel welcome and appreciated are more likely to make more questions and eventually make a purchase. But as presentable and nice an employee might be, it needs to be helpful for the customer needs; The employees needs to have knowledge of the store's products and prices and know how to help the customer in every situation. This has to be paired with availability; There needs to be enough employees available to assist customers in their needs around the store, otherwise they might feel unappreciated and leave. Finally, one factor that is commonly overlooked is how employees can become a nuisance for customers. Contrarily to having too little employee availability, some stores have too much in the sense that the employee just becomes annoying to the customer, always following it around and making unwanted suggestions. This is an express way of driving the customer out of the store and should be downright avoided.

Cash-out quality is the last of the store performance major KPI's, and it measures the customer's evaluation on its cash-out experience. This major KPI can be simply drilled-down into two smaller KPI's: simplicity and speed. There's not much to explain here: cashing out should be a quick and simple experience; The customer shouldn't need to go through massive and tedious queues to pay for its product, nor it should need to deal with complicated procedures regarding its payment. If a customer knows that the cash-out in a certain store is slow/complicated it will avoid shopping in that store in the first place.

4.6.2 Comment Selection

To be able to calculate the selected KPI's, specific user comments that regard the topics discussed need to be selected. First, it's essential to figure out what type of user comments can be used to make the analysis, and then, based on those stereotypical comments, identify what are the primary keywords to search by in order to automate the process of finding these comments. The inherent problem with this method is that customers won't always define why they liked or disliked about something specifically, which will difficult the drill-down of specific KPI's; Sometimes it is only possible to obtain a global overview on a broader subject than to focus on a specific one.

The store appearance KPI's relate directly to the store and as such the search should contain the word "store". However, an initial search using only the keyword "store" revealed that many comments that had nothing to do with store performance were being selected, such as people asking if a certain product was available in store. This only meant that "store" had to be paired with other words to refine the search. When searching about store layout, the objective is to find comments similar to this model:

- *"I enjoyed the store and the layout is easy to understand!"*

Nothing seemed more fit than the exact word, "layout", as a keyword for the search. A search through the data revealed that if "layout" and "store" where in the same comment, it was usually referring to the store layout. This is a comment found by searching by "store" and "layout".

- *"Was great. Recommend it. Although the layout of your store is such that for a while I believed that I may have to camp there overnight. Had grown a beard by the time I made it out."*

As for store design, though no comments could be found in the data collected, the combination of the words "store" and "design" or "style" or "model" are believed to be good for finding comments related to that subject. Store presentation is something that users don't usually mention to commend, so instead of keeping the search focused on key words such as "store"+"presentation", it may be wise to combine words that are often used in negative manner such as "organization", "disorganization", "mess" or "messy", "unfolded" referring to clothes, "dusty", "chaos" or "chaotic", and "sloppy".

Employee quality is something that may be easier to measure through customer feedback since its something that people mention more frequently, even to commend. To designate employees, words like "employee", "clerk", "assistant", "staff", "personnel", "service" or "customer service" should be used, and combined with "store" related words to guarantee they mention events happening within the store.

Employee quality KPI's are particularly difficult to drill down into more specific categories because of how people usually phrase their commendations/complaints in a way that doesn't blatantly focuses anything in particular. To identify these subtleties a more advanced tool of natural language processing would have to be used in order to extract deeper meaning from the text in an automated way. Sometimes it is obvious what the comment is mentioning, as these examples show.

- *"It was the sofa area by the ladies fitting room. The assistant was really helpful."* - This comment clearly commends the assistant's helpfulness.

- *"Top marks to your assistants in the ladies dept at Lakeside everyone I came across was pleasant friendly and very helpful"* - This comment is complimenting both the assistant's sympathy and helpfulness in a clear way.

However most situations are much less obvious, as the following comments exemplify.

- *"Dreadful customer service in Derry store today. 1st. The assistant chewed gum while serving me yuck."* - This comment about an assistant chewing gum at work relates directly to the employee presentation KPI, however that conclusion is not at all obvious for a regular natural language processing algorithm.

- *"Fantastic service from the lovely sales assistant at Warrington store today for a no quibble exchange of a faulty school bag when we couldn't find our receipt... nothing was too much*

trouble" - This comment probably fits into the employee helpfulness kpi, but it is so unclear that having a NLP algorithm extract the right meaning for it would be very difficult.

As for cash out quality, words such as "cash out" or "payment" should be combined with "store" to convey payments done inside the store. The feature people will most be commenting is speed, since queues are what mostly drives costumers angry. As such, to evaluate the cash out speed KPI, the previous words should be combined with "queue", "time", "forever", "fast", and "quick". As an example, here are some comments found by searching with these words.

- *"I've just stolen some things from your store served my prison sentence and it was still quicker than waiting in line"*
- *"Again. 2 tills open. Both agents flogging store cards. Queue out the door. Get a grip."*
- *"you must address your customer service - long queues for 20 minutes at Rugby store this afternoon and only one assistant serving. No wonder your results are poor"*

4.6.3 Analysis and Outputs

To engage on the analysis, first the general sentiment should be extracted from all comments, indicating what is the global opinion of the users about the brand's stores. This will be one of the major outputs of the analysis. Then the global comments should be drilled-down into the major KPI's and so on, applying the same method of analysis. In the end the retailer will have not only a global overview of the customer's sentiment about its stores, but will also understand the sentiment towards specific aspects such as employee helpfulness, or cash out speed.

The final outputs should be a reliable representation of reality; As such each aspect evaluated should be weighed by the number of comments mentioning it, and by time of posting. The number of comments mentioning something specific is an important indicator of reliability: if only one person is complaining about a faulty shirt, its probably just an isolated case, but if there are 20 people complaining, than that's probably a problem. Time of posting is also a crucial indicator of relevance. As studied in Chapter 3, data loses relevance as time passes: a complaint about long queues in December may not be relevant one month later, when Christmas has already past.

All of the KPI's are meant to be calculated for specific store locations, in order to give the retailer actionable information about its stores. However, in order to do that, the users commenting must direct their comments towards a specific store location, otherwise it's impossible to know what store they are talking about. In these situations there are two options: Assume that the user's opinion applies to all the brand's stores in general, or use the user's location (if available) to estimate to which stores he may be referring, regarding the brand's stores that are closest around the user.

To analyze many of these comments and to more accurately drill down the information on why the customers are commending or complaining about the store, an advanced NLP algorithm should be used for the processing of the text. Searching by keywords is just the first step of the

analysis; if detailed conclusions are to be extracted from the text, a natural language processing algorithm needs to be used.

Finally, it would also be of interest to detect trends in the collected data, for each KPI, identifying what are users most talking about and what are the most common keywords or expressions. This might give the retailer an edge to understand the rest of the data (i.e. a trend of the words "long queues" might help explain a rising negative sentiment in user comments about cash out quality).

Summarizing, with the collected comments about each specific KPI drill-down, outputs for each of them should be generated comprising trends and totals of sentiment and keywords, weighed by quantity and time, and grouped by store location. The same analysis should be done for the whole data collected for store performance drivers, to present a global overview of the use-case. As a suggestion, each KPI drill-down could have a threshold, defined by the retailer, that could make a warning go off whenever reached, making it easier for the retailer to identify abnormal events.

4.7 Demand Forecast and Demand Driven Replenishment

Demand Forecast is a use case for Customer Xperience that aims to predict customer demand only by using data from the social media: the user's opinions about products, complaints, wishes, and recommendations to friends, all provide a basis to infer the probability of a future purchase. Being the most complex and perhaps innovative of the use cases, this subject was studied in detail in chapter 3 where not only traditional solutions were reviewed, but also the hypothesis of forecasting with social media was explored in detail. Now, using that information, an attempt will be made to realistically devise a method for selecting and analyzing user comments in a way that most benefits the formation of a forecast in a practical way.

Deeply connected with Demand Forecast is the use case of Demand Driven Replenishment, as it depends directly from the previous. Demand Driven Replenishment aims to recommend optimized store replenishment to the retailer based on the demand forecast to help maximize the margin of return on investment. Because of this dependency, it only makes sense that both of these use cases are explored within the same section.

4.7.1 Demand Forecast with Social Media - A Delphi Method Approach

Forecasting demand using social media data is an innovative, and certainly challenging, concept, and not one that has been very explored yet. Though some articles provided invaluable information about the subject and its intricacies, it still was left unclear how to practically accomplish a forecast using a set of collected data from the social media. It wouldn't, however, feel right to not at least attempt to reach an outline for a forecasting method that could be applied on the available data.

Since none of the traditional quantitative methods are applicable in this case, it was thought of using a qualitative method to try to attain some insight on how to forecast, and then hopefully be able to translate that knowledge into something that can be automated. From all the qualitative

forecasting methods revised in chapter 3, the Delphi method appeared to be the most effective; as such this method was chosen to be conducted in this situation. The objective was to harness the knowledge of people that had more experience in the retail area and that could potentially provide enlightenment in this subject, in a controlled way.

The method was to be conducted within Wipro Maia, selecting a group of personnel members to participate through the answering of questionnaires. A relatively small group of people was chosen in order to not make the process of data analysis to extensive. Initially, 24 members were selected, though the objective was to have at least 15 of them answering (a greater number was selected because it was already known that not all would respond). Knowing that the time of people in the company was limited, and also trying simplify the process of analyzing the data collected, this exercise was constituted by only two questionnaires. The exercise was initially explained to all participants and all the response periods and deadlines were defined; Each questionnaire had a response period of 5 days, but the actual time needed to answer it was around 10 to 15 minutes.

4.7.1.1 First Questionnaire

The first questionnaire conducted (that can be observed in Appendix C) had four main sections: participant background on the social media, social network distinction, keywords for forecasting, and mutual user influence. The first section had the objective of understanding what is the involvement of the participant with the two social media networks that are being studied: Facebook and Twitter (this is because the DDP only collects data from these two networks). The second section was meant to observe how each participant differentiates the two social networks and to take their opinion on if the forecasting should be different for each network based on their differences. The third section asked the participants to indicate the most significant keywords for forecasting, in their opinion. Finally the fourth section was focused on how the participants weigh the influence that users have on one another when it comes to purchase intent.

Surprisingly, and unfortunately, only 9 people of the selected group responded to this questionnaire, though their answers were still valuable for the study. The results from the first questionnaire provided some interesting insights. The results of the first questionnaire can also be observed in Appendix D. The results indicated that almost none of the participants used Twitter at all, with 7 out of 9 answering that they never use Twitter, while most of them use Facebook regularly, with only one of them never using Facebook. Despite of most users not utilizing Twitter, most of them said that the networks are somewhat similar and indicated that the main distinguishable characteristics between the two are contents and user interactions. When asked about how the differences between the two networks affected the process of forecasting, all of them agreed that a relevant influence exists. The following is the list of keywords selected by all the participants that were deemed relevant when searching for comments that indicate purchase intent: Want, Looking, Buy, Need, Wish, Like, Bought, Interesting, Useful, Good Deal, Cheap, Must Have, Trendy, Good Looking, Opportunity, Promotion, Good Product, Price, Cost, How Much, Recommend, Love, Excellent, Saving, Bargain, Fast Delivery, Discount, Coupon, Free Shipping, Top, Wear, Nice, Amazing, Handy, Great Experience, Reliable, As Advertised, Easy Checkout, Great Customer

Experience. A rough approximation of the relevance given by the participants to the keywords chosen provides the following five most important keywords:

- 1 - Buy
- 2 - Must have
- 3 - Like
- 4 - Need
- 5 - Cheap

Finally, under the subject of mutual user influence, participants users agree that users exert a significant influence on each others opinions, and that comments about products can greatly influence their viewers' purchase intent, even more if these comments are liked or shared.

This information already provided a solid baseline for devising a forecasting method, however a second questionnaire would present a more concrete challenge to the participants, and get closer to achieving an actual forecast.

4.7.1.2 Second Questionnaire

The second questionnaire conducted (which can be viewed in Appendix E) consisted only in 4 questions. Despite being smaller it required its participants to think their answers through more thoroughly. The first question allowed the users to rethink their answer on the first questionnaire about the five most important keywords, now that they've seen the answers of the other participants. In the second question, a group of eight comments selected from the DDP data, based on the keywords chosen by the participants in the previous questionnaire, is shown to them, and it is asked of them to classify these comments on how much they indicate an intent of purchase from their authors. In the third question, six made up comments about a fictitious dress are presented, along with the number of likes, shares, and the number of friends of the user that commented; It is then asked to the participants to make a qualitative forecast on how many of the people that visualized the comments are going to buy the dress. Finally, the last question requests the participants to make the same forecast of that dress, but now quantitatively, as a rough estimate.

The only people qualified to participate in this second questionnaire were the 9 people that participated in the first one, since their answers were going to affect the way they made their decisions in this one. From these 9 people that responded the first questionnaire only 7 were available to participate in the second. The results from this second questionnaire can be consulted in Appendix F.

Results of this questionnaire showed that most participants would not change their previous answers on keywords, except for two of them who tailored their answers more in line with the general trend from the previous questionnaire. The second answer indicates that most participants give more value to a user's direct intention in buying something than a positive remark on the product; This can be observed in comments 1 & 2; While in the comment 1, the user is stating

that it is going to buy the product with no kind of remark about it, in comment 2 the user makes a very positive remark but doesn't confirm a future purchase, instead just shows a strong intent to. It is also notable that participants find comment sentiment not as relevant, as long as the user is showing any intent of buying; this can be seen particularly on comments 5 & 6 where the comment sentiment is clearly negative and yet the participants classify them, on average, as indicating a moderate intent of buying. The third question had a variable range of answers, which is comprehensible, since a qualitative answer will always depend on the perspective of the one who's answering. On average, the participants classified the forecast as a 3.1/10 in terms of how much people would buy the product. However, in the fourth question, answers diverged greatly from their qualitative counterparts: The same qualitative score given by different people translated to very different quantitative forecast numbers. Nonetheless, most participants gave a similar number of sales as forecast, averaging around 13 units sold. The number of 13 sales is a very safe number because it conveys the summation of the positive comments plus their likes and shares, which totals 17; considering that not all of them will buy, due to different states of sentiment, 13 is a sensible number to choose. However this number doesn't comprise the friends of the user that will potentially see its the comment and be influenced by it, though this might be speculating too much into the participants' reasoning.

4.7.2 Comment Selection

Comments chosen to forecast demand must indicate an intent to buy a given product. Though negative comments about a certain product may affect demand as they indicate flaws in the products to other potential customers, this approach won't take negative comments into consideration because they would further complicate the already unknown and complex process of forecasting. As such, for selecting the comments, keywords will be used, and these keywords will be selected from the ones suggested in the questionnaires.

Starting with the five most relevant keywords, in the perspective of the participants: Buy, Must Have, Like, Need, Cheap. Searches conducted with the keyword "Buy" seem to present consistent results indicating intent to purchase with comments such as the following:

- *"when will the Urban Decay Backtalk palette be available to buy in store? Thanks"*
- *"I just wanna buy some foundation and it won't let me delete this number to be able to proceed."*
- *"Is there a problem with your website? Trying to buy some items but can't get passed the address page at checkout..."*

When using the keywords "Must Have" to conduct a search for comments the results were not satisfying; the keywords were not being used in the intended way in any of the comments found. Instead of referring to products the users "must have" or need or want, the words "must have" were being used as the following collected comments show:

- *"Must have been a tight fit. The only item sent in this enormous box!!"*
- *"Must have been a return? Does the returns not get inspected? I'm really disappointed I was going to use this bag but not anymore."*

The keyword "Like" didn't prove to bring any particularly interesting results either. Most comments found did not show any intent of purchase, instead the keyword was being use in phrases such as "i would like to" or "like this", just as the comments collected show:

- *"I would like to use 3 gift cards to buy make up online but it isn't letting me! Please help! Thanks!"*
- *"Would anyone like to advise me further on this please??"*
- *"Poor quality and uneven ends on this mac product. Should it look like this?"*

The "Need" keyword also didn't provide many useful results as the comments were rarely indicative of an intent to purchase. Below are some of the comments collected with this keyword:

- *"Maybe you need to re-evaluate who you are trusting with your Brand"*
- *"You need to learn how to address you customers politely. Your aggressive tone is not appreciated"*

The same goes for the keyword "Cheap", as the few comments found were referring to anything but purchases. Some comments are shown below:

- *"Is it legal to sell cheaper imitation cola in your cafes? ?? when advertised as Pepsi ??it's disgusting"*
- *"Ordered my wife a Gilet for Christmas got some cheap men's socks instead..."*

As most of these keywords didn't provide any valid results, searches were made with all the other keywords suggested by the participants of the questionnaires. Surprisingly, only the keywords "want", "love" and "amazing" provided some usable comments, and were far from being optimal. This analysis demonstrates that the comment selection for forecasting cannot be based on simple keywords but instead will require a more complex search, possibly utilizing a machine learning algorithm to perfect different combinations of words. Nonetheless, the reviewed keywords will prove as a baseline for the work to be developed ahead.

4.7.3 Analysis and Outputs

The study made so far allows to delineate the outline of what might become a forecasting method for social media data. The analysis made in this section will not go so far as to try to reach a method of forecasting, but instead will cover the important aspects that such method should convey, and provide a theoretical basis for its development.

Firstly, any forecast made can only be based on comments referent to a specific product. Many times users comment that they will buy but are not specific as to what; they might say its a dress or some shoes but won't specify which. This is a real problem that is difficult to work around. A solution for a retailer that wishes to perfect its forecasts using the social media is to increase the quantity of posts it makes on its social pages relative to specific products; That will guarantee that most comments made on those posts are related to that product, and as such the target of the comments will be known. To make this viable, however, the data collecting software needs to be able to identify what post the comment is targeting. Also worth mentioning is the fact that all comments that indicate purchase intent are relevant for the forecast, even ones that refer to online purchases; This further validates the concept of Omnichannel that wishes to unify the platforms through which customers interact with the company. No matter if the customer acquires a product in a store or has it delivered to its home, the product is being sold and stock needs to be allocated.

Equally relevant is the location of the user. Though some users will state the name of the store on which they plan to buy, most won't. In this case, it is advised, as mentioned before, that the location of the user's residence found in its social profile is used to estimate which store in its vicinity it will visit. Though this might have a significant margin for error, it is still better than not having an estimation for location at all; The need for a demand forecast is mostly to understand how to allocate and move inventory in space and time to supply all the customer needs, location is a key factor of that process.

As like the rest of this project, sentiment analysis also has an important part to play in demand forecasting. The weight a comment has for a specific forecast should be proportional to the score of its sentiment. Though, as seen early, most questionnaire participants disregarded sentiment as a very important factor in showing intent to buy, it still had some effect and should be taken into account. Even within the negative, neutral or positive sentiment range, a comment could be more or less positive, or more or less negative, with varying sentiment scores that can ultimately affect the comment's weight for forecasting.

A particularly important factor to take into consideration when forecasting is the mutual user influence in social media. This has been covered before in Chapter 3; User interaction definitely fosters mutual influence that has to be taken into account; The friend list of a user may say a lot about how many people it is reaching, though that count may probably be more accurate if the number of people liking and sharing/retweeting its comments is taken into account. The questionnaire done shows that participants value this factor and that, in fact, think it is highly relevant for the forecast. The real challenge is how to weigh these factors in the forecast. Common sense (as well as the results from the questionnaires) dictate that shares/retweets mean more than likes, and likes mean more than a user's number of friends. This is because sharing/retweeting means the user agreed so much with the post that it decided to show it to its own friend pool; Liking is a simple way of the user showing mild sympathy for the post's content while not taking any significant stance; And finally, a friend list is something passive that doesn't accurately translates into anything tangible: from 2000 friends, only 50 may have seen a comment written by the user and if they don't react to it it's not even clear if they liked it or agreed with it in any way.

Last but not least, it is once again remembered the importance of weighing data according to its date. A comment is as relevant as its proximity to the present time. This is evidenced in the exponential smoothing forecasting method, that exponentially reduces the weight of past data as time passes and as new data is acquired. It's fairly easy to understand that if a customer stated intent to buy a pair of boots 15 days ago, that statement doesn't hold up as much as the one of a user that said the same thing 2 hours ago. This also bears the question as to how long a user takes to act after declaring intent of purchase; It's obviously not an instantaneous thing, otherwise the user wouldn't even bother stating it online, but it's also not a very delayed act, because if the user craves the product it will try to buy it as soon as possible. Finding the ideal timing might be tricky, though it might be made easier considering known facts such as people making most of their clothing purchases on weekends, or such as the occurrence of specific limited time events.

In the end, the outputs of this use case should provide the retailer with a quantitative forecast for sales referring to specific products and locations. Taking all of the discussed factors into consideration, a forecasting method should be devised. Being an innovation, it will probably require a long testing period, forecasting with real data for retail companies, and comparing results to perfect the algorithm and reach an accurate solution.

As for Demand Driven Replenishment, its outputs are directly connected to the Demand Forecast's outputs. Knowing what the demand forecast is for each location, it is then possible to create an algorithm that best allocates the inventory in a way that optimizes availability and, consequentially, maximizes return of investment. This can only be done by connecting with the internal company data and knowing what is the inventory on each store and warehouse, and create schedules to move it according to demand and safety stocks.

[1]

4.8 Brand Engagement

Brand Engagement is the use case of Customer Xperience that is responsible for tracking the general public sentiment towards the brand itself and how to improve that. By analyzing customer comments it is possible to understand what is their opinion towards the brand, its products, services and stores; This can be used to adapt the brand strategy in order to become more likable, approachable and iconic. The insights gathered will also be beneficial to improve how the brand represents itself in the social media and how it conducts its social media page management.

4.8.1 Brand Engagement KPI's

To designate an objective for the analysis, Brand Engagement KPI's need to be defined and a sensible way to measure them needs to be determined. Brand engagement KPI's aim to measure what kind of relationship users have with the brand. To do this, there is a need to classify user opinion and sentiment in several areas related to the brand's operations, without focusing too

[1] Omnichannel - The unification of all the communication channels and platforms through which the customers interact with the company in order to create a seamless customer experience.

much on one thing, but on the overall brand. In this sense, four major KPI's were selected: Services/products quality, customer service quality, social page popularity, and website quality.

NOTE: The chosen KPI's are of general use, and serve as a standard template; Each retailer is different, with distinct needs, and might choose different KPI's depending on its business strategy.

Services/products quality is a KPI that aims to evaluate what is the customer opinion about the value that the brand provides. It can be obviously drilled down into services quality and products quality. Services quality reefers to the value provided by the brand in services; in this case it was identified that the main service the company provided and that the customers mentioned was deliveries. Products quality regards the overall quality of the brand's items in terms of price, style, durability and utility.

The **Customer service quality** KPI seeks to measure the users' opinion about the customer service they are being provided overall (inside or outside the stores). A good customer service is a powerful way of building brand engagement amongst the customers, making them feel valued and appreciated, while providing them with all the help they need to profit from the brand's services or products.

Social page popularity is the KPI that evaluates how well the brand's social pages are doing in terms of popularity. To do this, user activity is measured frequently along with general sentiment to identify its growth or decline along time and devise counter-measures if needed. This KPI also seeks to track how much the contents shared by the page are having success amongst users and how to improve that rate.

Finally, the **Website quality** KPI is based on the users' opinion about the brand's website. Navigability, simplicity of use, and clarity are characteristics that users value particularly when visiting a web page. This KPI is important because the brand's web page is the main representation of the company on the web and as such should induce a positive response on its visitors.

4.8.2 Comment Selection

This section will cover how to select specific comments to calculate the chosen KPI's regarding brand engagement. Firstly, since the focus is on the brand in general, it would be useful to associate to this analysis the general sentiment trend that was observed in the preliminary analysis, just to have a notion of how the public sentiment in the social pages is shifting over time.

Service/products quality conveys all comments that mention experiences with products or services. The search however, may not be that straightforward as the NLP algorithm used needs to be able to correctly and consistently classify entities as products when it sees them. Services such as deliveries or store card functionalities also need to be correctly identified to select the right comments. It may be difficult to define specific keywords for products because of their variety and because customers usually don't mention the word "product" to refer to it, instead they use the product's name such as "dress" or "shoes". The algorithm should be then trained to identify products by their names, maybe by feeding it with the company's product's database. As for the services, the keywords can be chosen around the services the company provides, which in this case are mainly deliveries. As such, the best keywords are usually "order", "delivery", "parcel",

"shipping", "shipment", "package", "box", "packaging", etc. The following real comments are an example of customers commenting on products and services.

- *"Absolutely baffled that my order ended up in bloody Ipswich rather than Newcastle! It's not even in close proximity!"*
- *"I only bought these recently, first wear today and the heel has broken off.... not impressed "*
- *"Very disappointed only swim-wear you have is a plain black costume."*

The user comments to be utilized in the customer service KPI are usually easy to find since when users talk about the subject they usually use the exact words of "customer service". However there are some instances harder to identify where users talk about customer service without using that nomenclature; as such, the search should be conducted using other related words such as "assistance", "help", or "support". It is worth mentioning again that all kinds of customer service is being targeted here, be it on the store, phone, website, or on the social media (since there are several visible instances where customers receive assistance through the brand's social page). Below are some collected comments of people talking about customer service.

- *"The same to you. Really appreciate your assistance."*
- *"Been holding to speak to your customer service department for over 20 minutes so far. Trying to find out where the order is that I placed on 07th December. Not happy."*
- *"absolute joke of a service. Disgraceful customer service. Unable to get through on phone and I have to wait 5 DAYS to receive a response via twitter. I predict with that service you may not last past 2018"*

For measuring the social page popularity KPI, there aren't really a great deal of comments that can be used since people don't usually comment about the contents of the page, but simply react to them accordingly to the how they are stimulated to do so. This KPI will mainly be measured through statistical analysis of the data received such as number of comments and general sentiment.

As for the web page quality KPI, the search should focus on all comments that mention the brand's web page. Since users usually have no motive to talk about any other website on the brand's page, any comment with the words "web page", "web site", or "site" will most likely be referring to the brand's web page. The following comments were collected while searching with the keywords mentioning above.

- *"Your website is still wrong guys..."*
- *"Hi is there a problem with the website? I can't get past the delivery address page and don't want to lose my order."*
- *"Are you having a problem with your website? Been trying to access it for an hour on my phone but am unable to do so."*

4.8.3 Analysis and Outputs

Just as stated in the analysis for Store Performance, the comments should be correctly weighed according to their quantity and time of posting. The same goes for using a natural language processing algorithm to process text and more effectively find and interpret comments for the KPI's chosen, allowing for a more accurate drill-down. Identically, detecting trends on the data for each KPI drill-down will help understand underlying factors that are influencing customer sentiment.

For this analysis, a major part of the preliminary analysis done previously is going to be used because most of its data fits in the brand engagement topic. General customer sentiment is probably the most important of those, because it allows to track the overall sentiment the public has about the brand on a progressive time basis; This will be one of the outputs for this KPI. Any kind of demographical statistics are also pertinent to be shown along with the rest of the data for this KPI as it provides a sense of the type of users interacting with the brand.

The analysis of user distribution across platforms will be of value for the social page management KPI, to help the brand allocate the right resources in each social media appropriately, addressing the number of customers in each platform. In the same sense, the monthly comment count stat can help understand the amount of user activity in the brand's social page; Pairing that up with a like or share count will help calculate popularity. Something that would be interesting to measure is how appealing are the posts that the brand is making on their page; In other words: how relevant for the user are the contents being posted by the brand. Though this is not easy to measure, if it would be possible to obtain the data on how many people view a certain post vs the number of likes/shares/comments, that could provide an idea of how many people actively react to the contents, showing interest, and how many just pass by them, showing indifference.

Excluding the social page management KPI drill-down, that has its own method of measurement, for the rest of the drill-downs, with the collected comments about each of them, the outputs should be generated comprising trends and totals of sentiment and keywords, weighed by quantity and time. The same analysis should be done for the whole data collected for Brand Engagement, to present a global overview of the use-case. Once again it can be suggested that each KPI drill-down has a threshold, defined by the retailer, that could make a warning go off whenever reached, making it easier for the retailer to identify abnormal events.

Chapter 5

Conclusions and Future Improvements

The present chapter serves to present the main conclusions drawn from this study and to discuss what future work and improvements are to be done to finish the development of the project and improve its features.

5.1 Conclusions

The Customer Xperience project has the objective of extracting data from social media websites, analyze that data from a retailer point of view, and deliver actionable insights through a dashboard integrated with the Oracle Retail suite.

With the focus of this particular thesis being the analysis and insight generation of the social media data, all study made on the existing state of the art indicated that this is a feasible concept, though it will most likely require a large development time and the use of complex tools.

What makes this project possible is the existence of relatively developed tools of web crawling and sentiment analysis; These are the two main technologies responsible for the project's success. Web crawling tools are what makes the extraction of data from the social media possible, with relative ease. Sentiment analysis, on the other hand, is a vital element for the analysis of the data; The ability of classifying sentiment, identifying categories and keywords is essential to automate the analysis process in an accurate way. Unfortunately this is where the project was weaker, since the sentiment analysis performed by the DDP was very limited and even inaccurate; Any improvements for this project would be dependent of a better sentiment analysis algorithm. Despite these shortcomings, the current state of the art indicates that there are highly developed sentiment analysis algorithms that convey the functionalities needed to improve the project. The SotA also shows the existence of projects similar to Customer Xperience, such as IBM's Analytics for Social Media, which further prove the viability and potential of the assignment.

Though the objective of Wipro when proposing this thesis was to reach an implementation of the project by the end of it, it became clear early on that such was an unrealistic goal. As such, the aim became to deliver a more theoretical approach of the project, devising a "blueprint" that can be understood by a future developer (or developers) that picks up on the project. The time setbacks

that were forced on the project, such as the uncertainty of the initial plan or the huge delay on achieving tool access, also meant that not all of its aspects could be explored. Nevertheless, the work accomplished here was a very important benchmark in the overall project, as it proved the validity of the concept, explored the most important problems and pitfalls, and helped devise a work plan for the future states of development.

5.2 Future Work and Improvements

The Customer Xperience project is a big and complex assignment with several parts and details to have into account. As such this thesis was only able to cover a rather small portion of the project, at a theoretical point of view; There is still a considerable amount of work to be done in order to make it operational. This section will cover all the work that needs to be done and can be done to further continue and finish the project, serving as a guideline for anyone that wishes to continue the work developed until here.

5.2.1 Improved Sentiment Analysis

To allow the project to progress more fluently, refining and working the incoming data is of big importance. The cleaner and the clearer the data, the easier it is to use it on a complex analysis. Right now sentiment analysis is a big part of that analysis and, unfortunately, it is also a problematic part. The sentiment analysis that is being done by default by the Data Discovery Platform is not good enough for the objectives of the project; It lacks consistency, reliability, and clarity. Since the possibility of directly altering the DDP's algorithm is not on the table, the suggestion would be to perform an independent sentiment analysis on the incoming comments, one that is more reliable and that conveys more depth and detail. Doing this will greatly improve the results of the analysis and open new possibilities for more detailed analysis.

Another aspect that a better sentiment analysis algorithm could cover is the language range. Right now the analysis can only be done for English language. Though that might be enough when collecting data for a non-international English company, that won't be the case for most of them. Obviously, being able to analyze Portuguese text would a major asset for the project right now, since this solution is meant for Wipro's Portugal branch. Ideally, it should be able to analyze a range of selected languages, covering at least a handful of the most spoken languages around the world.

5.2.2 Remaining Use Cases

One of the most obvious work to be done to finish the project is to approach the remaining use-cases of the project. This thesis only covered the use-cases of Store Performance Drivers, Demand Forecast, Demand Driven Replenishment, and Brand Engagement; there still remain Category Adjacency, Store Staff Optimization, Stock Out & Lost Sales, Size Pack Optimization, Outlet

Locations, Price Optimization, Markdown Optimization, Product Improvement, Promotion Effectiveness, Demand Driven Localized Assortment, Social Advocacy, and Media Mix Optimization. Each of these use-cases would require a theoretical study just like the one done on this thesis before being implemented.

5.2.3 Implementation and Integration with the Oracle Retail Suite

The logical path after approaching all the use cases theoretically is to actually work on their implementation. Implementing isn't always as simple as translating theory to practice, there are always snags along the way and sometimes even restructuring is needed. When implementing a project there are always hidden problems and things for which the developers couldn't be prepared; in this case, for a project of this size and complexity, it won't certainly be easy to implement its features.

As was the main objective of this project, this implementation should be done in a integrated way with the Oracle Retail Suite, which is the main software suite used by Wipro in the majority of its projects. This would mean that Customer Xperience would then be part of the Oracle Retail solutions used by Wipro Portugal and could be employed in virtually any project.

5.2.4 Crossing Social Media Data with Historical Company Data

Once the use cases are implemented with oracle retail the project is ready to take the next step: crossing both social media data and historical company data together. This combination of data sets will dramatically improve results, as analysis made by both sets can confirm their accuracy with each other and create a feedback loop. The integration of this new data set will incontrovertibly call for big changes in the analysis structure for each use case; Now, there will be a need to account for completely new data that is related to the data collected so far. The relationships between the two data sets will need to be correctly identified and the analysis processes will need to be adjusted accordingly; this might take some trial and error.

5.2.5 Use of Machine Learning

To maximize the potential of all the analysis done, it would be highly beneficial to implement a machine learning algorithm. Having this (and having combination with the historical company data) would allow to completely automate the process; the algorithm would be able to compare its analysis predictions to the real results and progressively adapt them, learn relationships between the data sets and improve its processes with minimal human intervention.

A machine learning algorithm would also be an important tool to implement an automatic keyword picker, which accordingly to [42] would be an important step in developing filters for social media data.

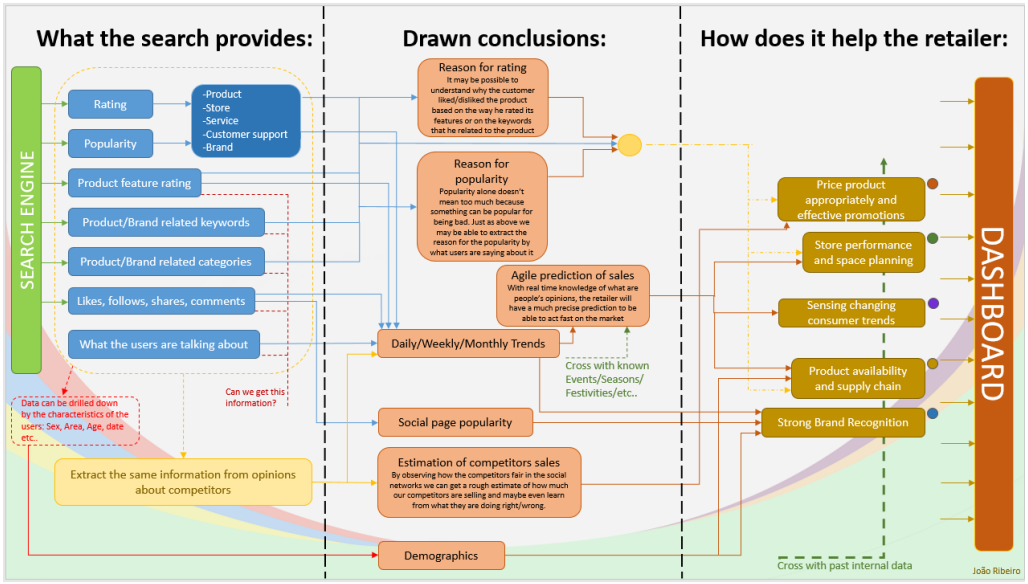
Though all of this is easier said than done, it would certainly be the icing on the cake.

5.2.6 Dashboard

The final phase of the project is the construction of a dashboard to display all the analyzed data; That will be the subject of another thesis. The dashboard has the function of displaying, in a clear and understandable way, all of the relevant data for the retailer to be able make decisions; It's the front-end of the tool and it's supposed to be simple to use and clear. This dashboard should be done in a integrated way with the Oracle Retail Suite, and should finalize the development of the project.

Appendix A

Appendix 1 - Customer Xperience Final Project Outline



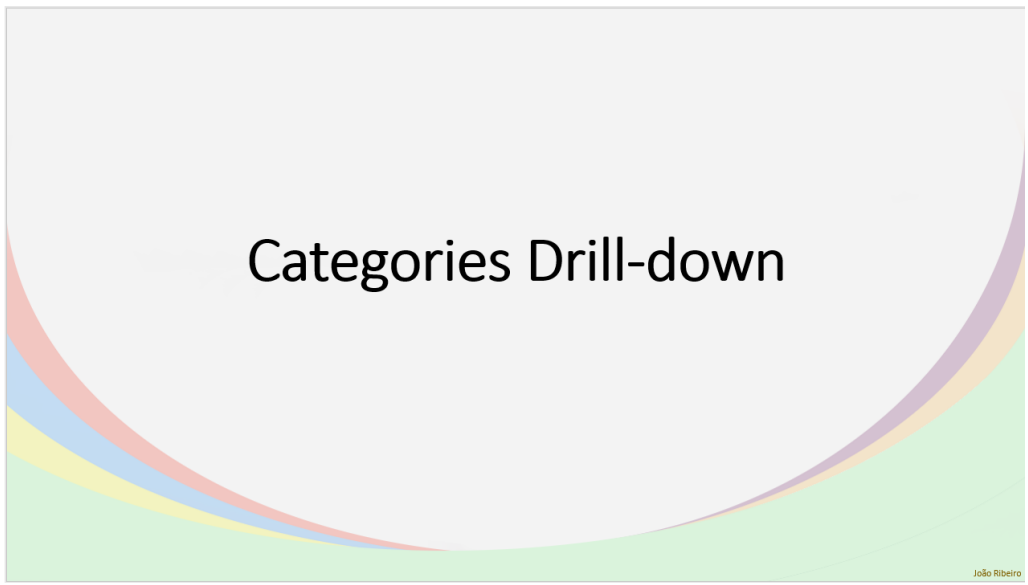
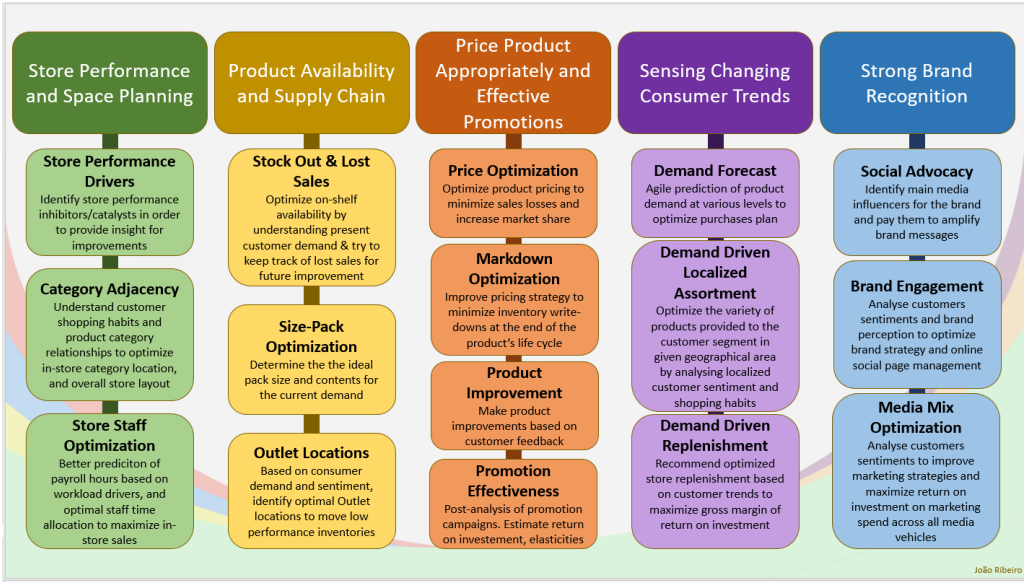


Figure A.1: Customer Xperience Final Project Outline - Pages 1,2 and 3



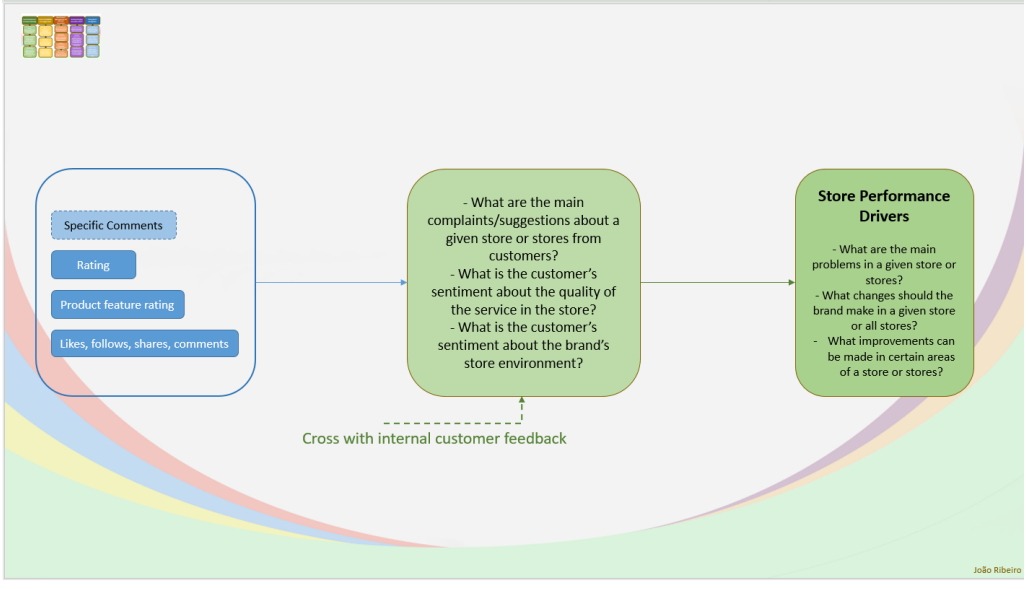
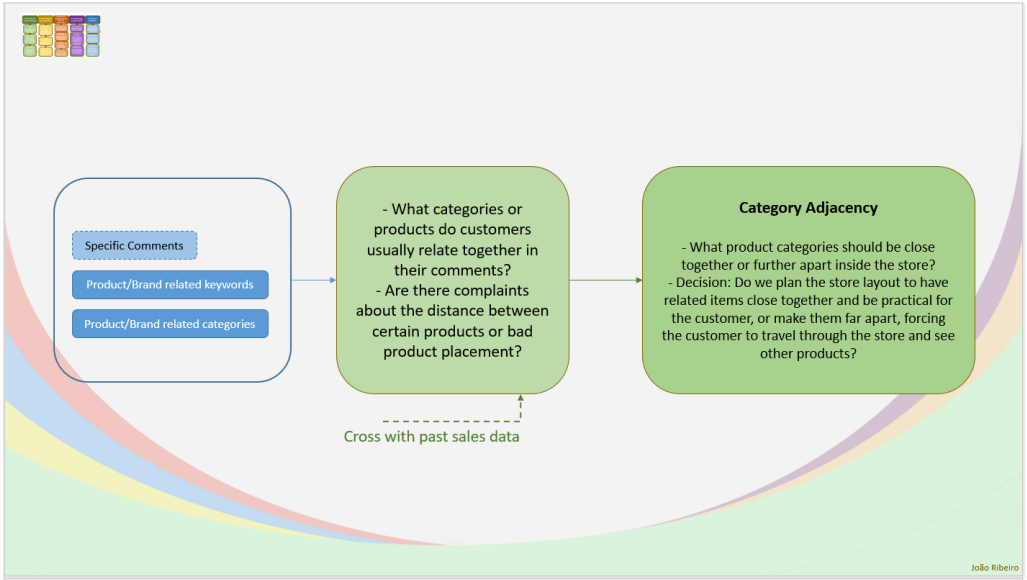


Figure A.2: Customer Xperience Final Project Outline - Pages 4,5 and 6



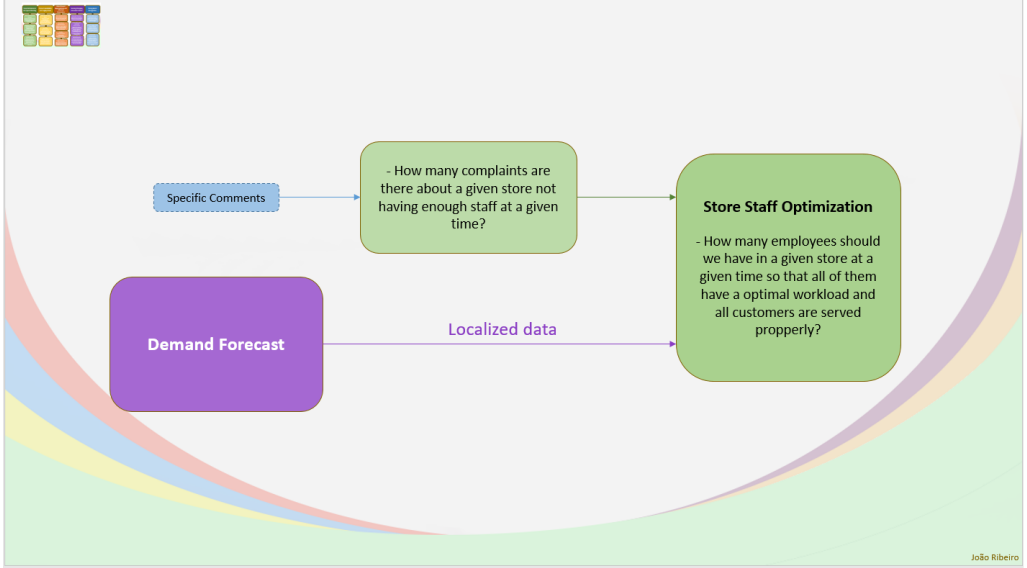


Figure A.3: Customer Xperience Final Project Outline - Pages 7,8 and 9

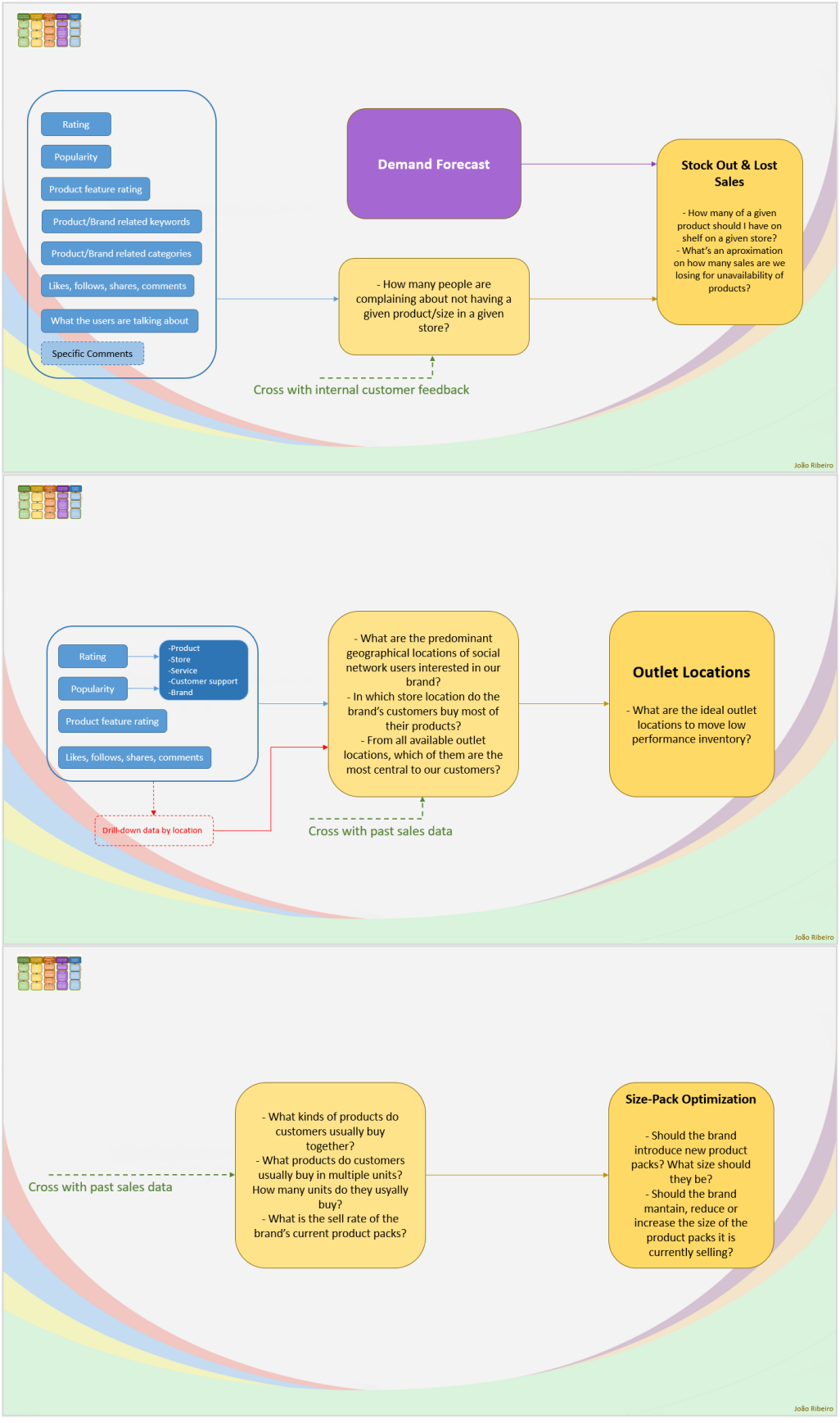


Figure A.4: Customer Xperience Final Project Outline - Pages 10,11 and 12

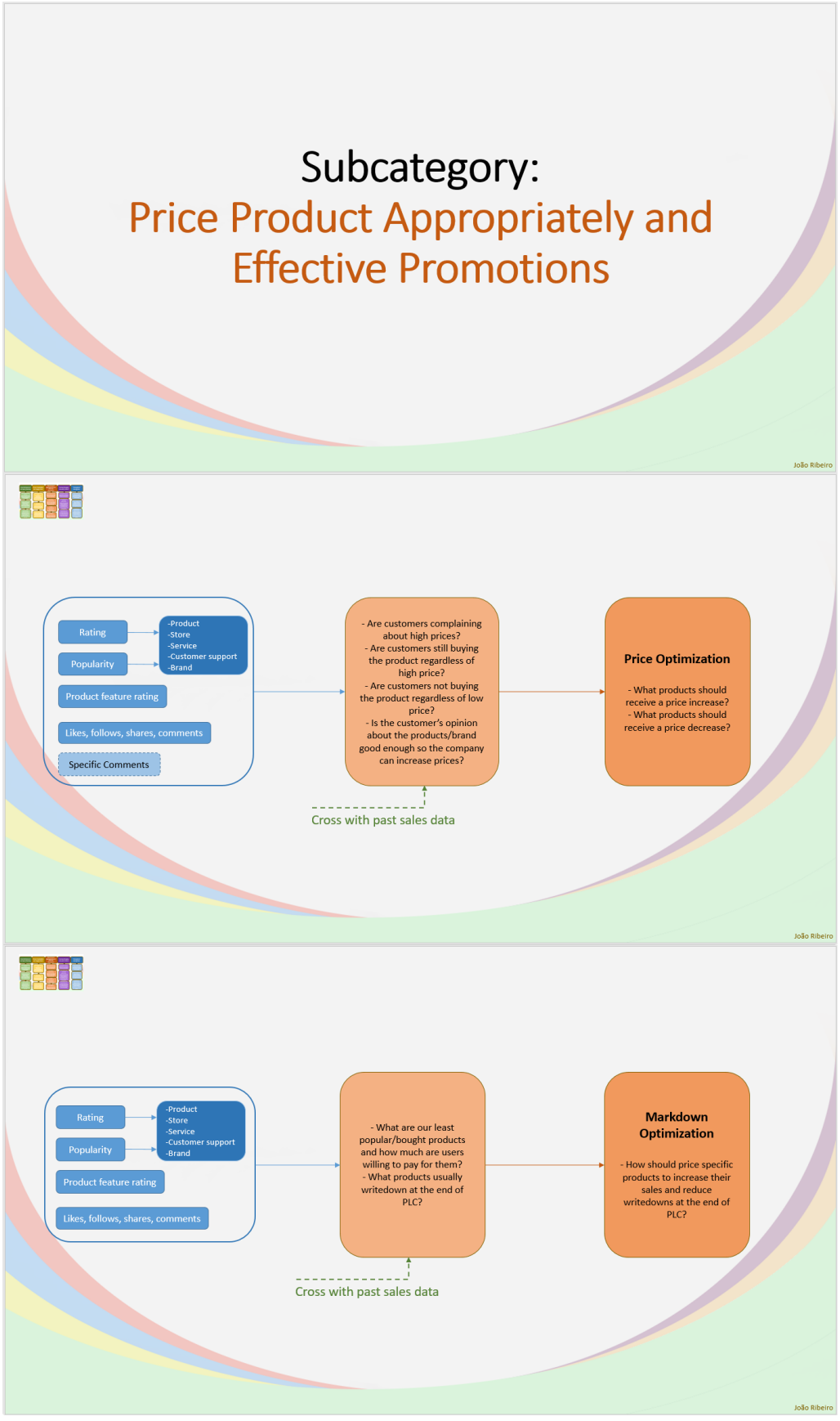
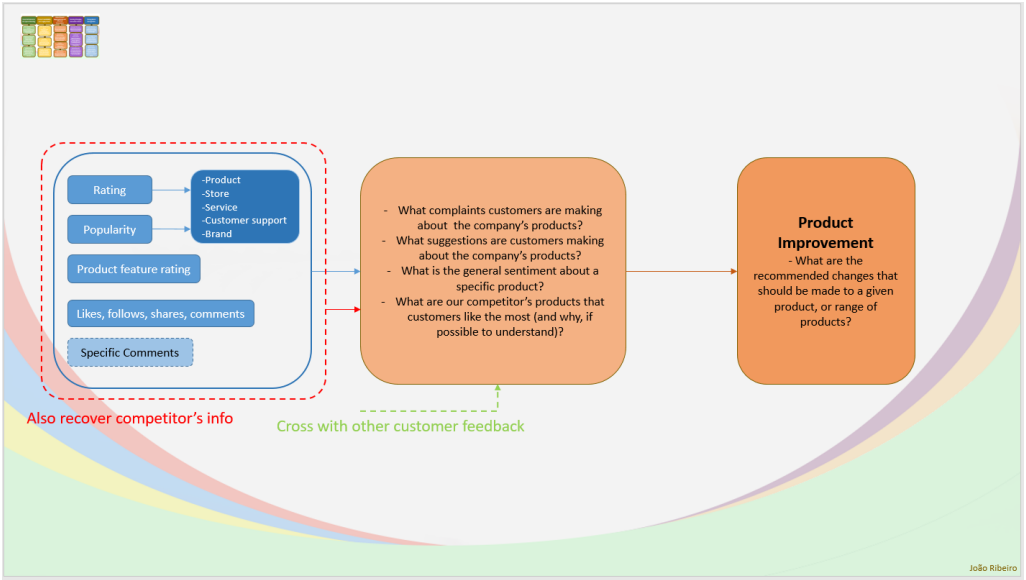
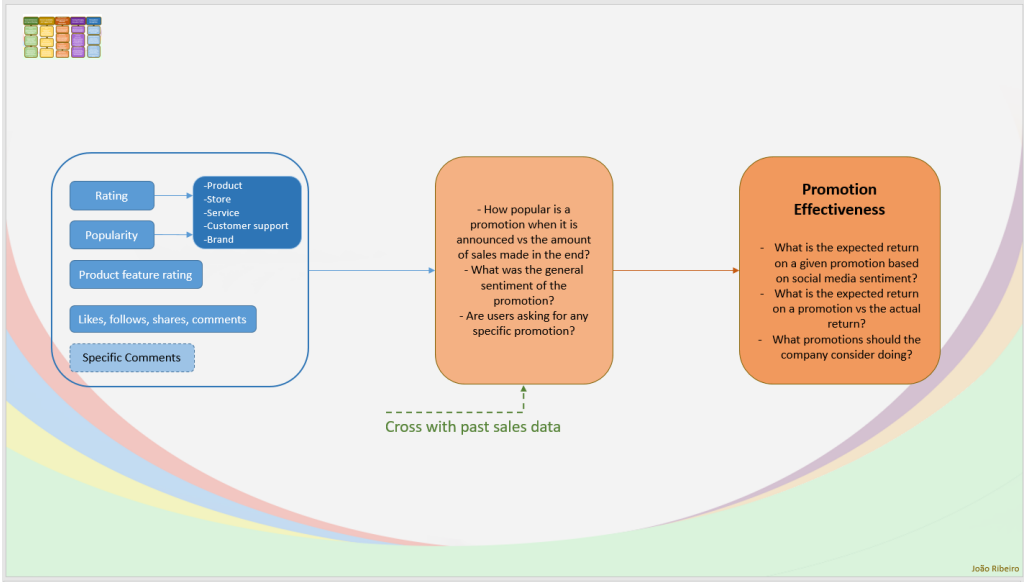


Figure A.5: Customer Xperience Final Project Outline - Pages 13,14 and 15





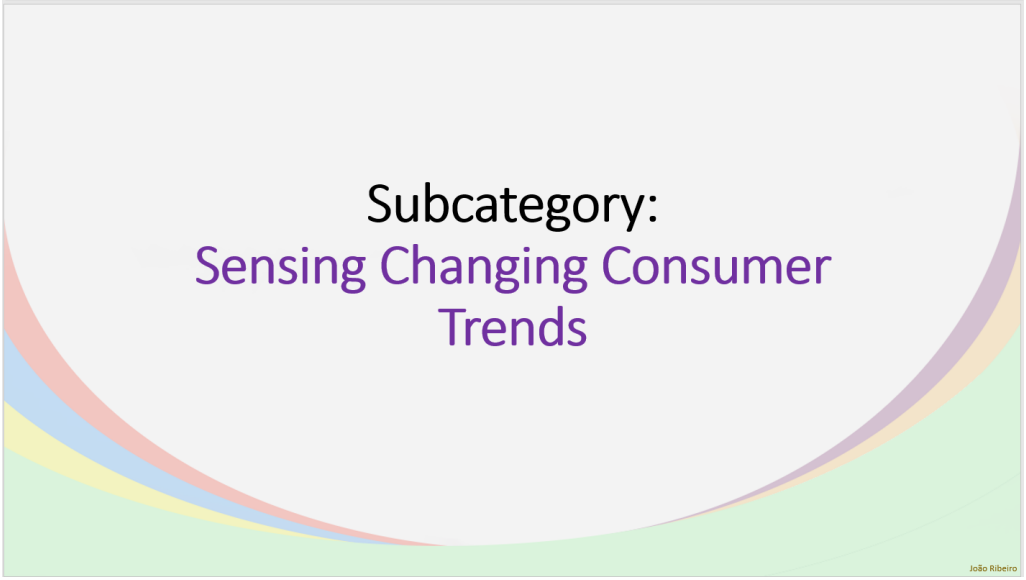
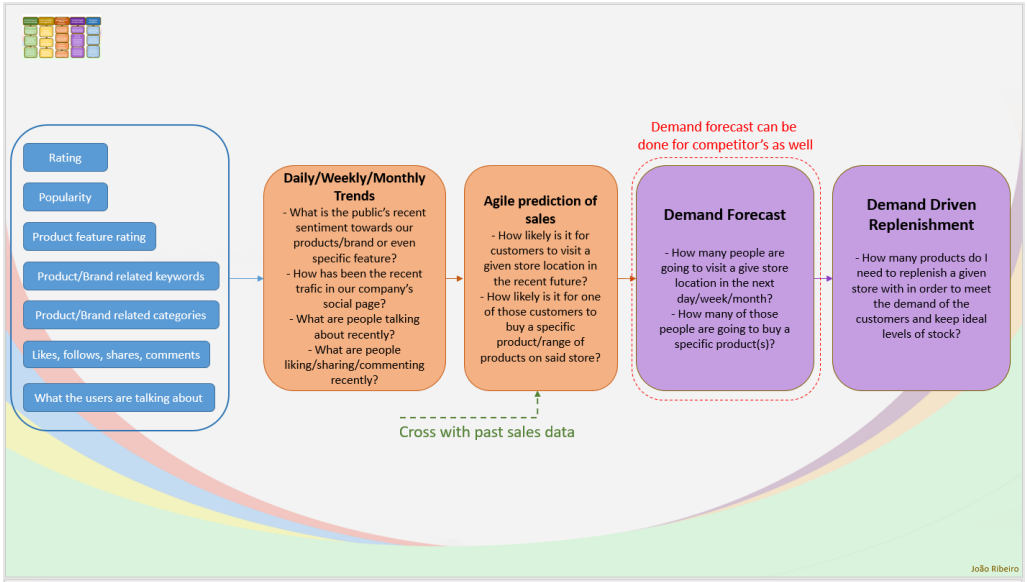
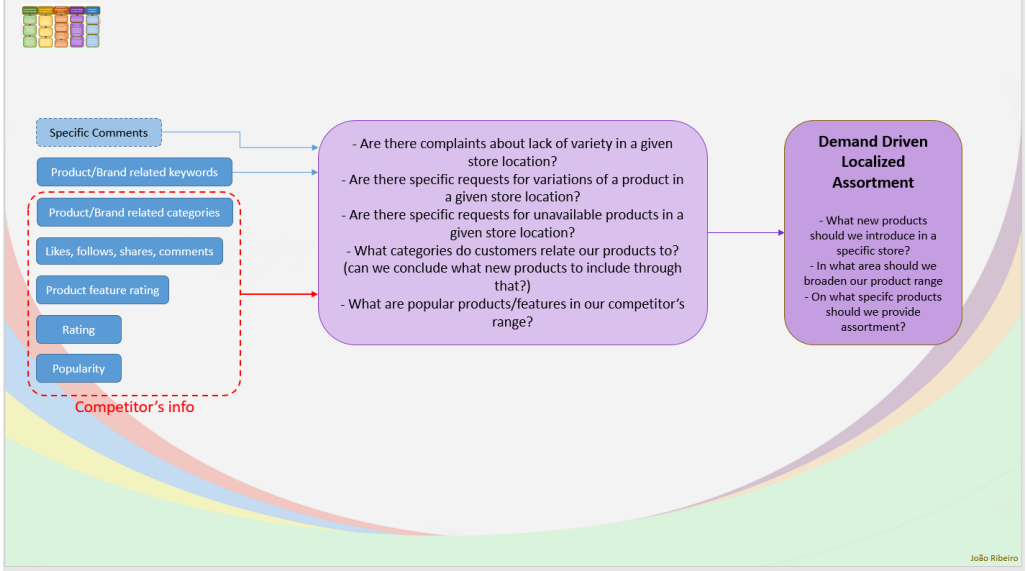


Figure A.6: Customer Xperience Final Project Outline - Pages 16,17 and 18





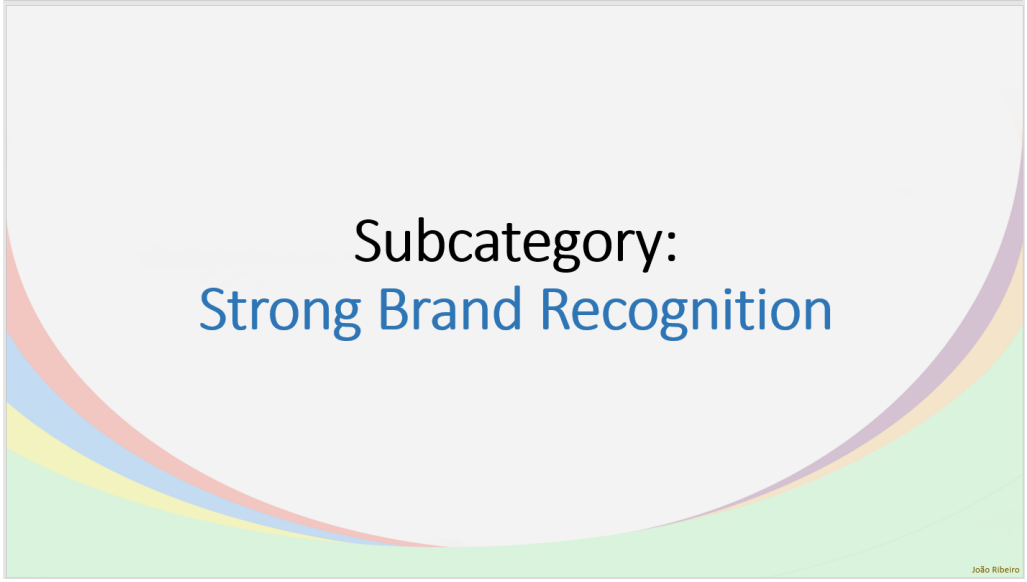


Figure A.7: Customer Xperience Final Project Outline - Pages 19,20 and 21

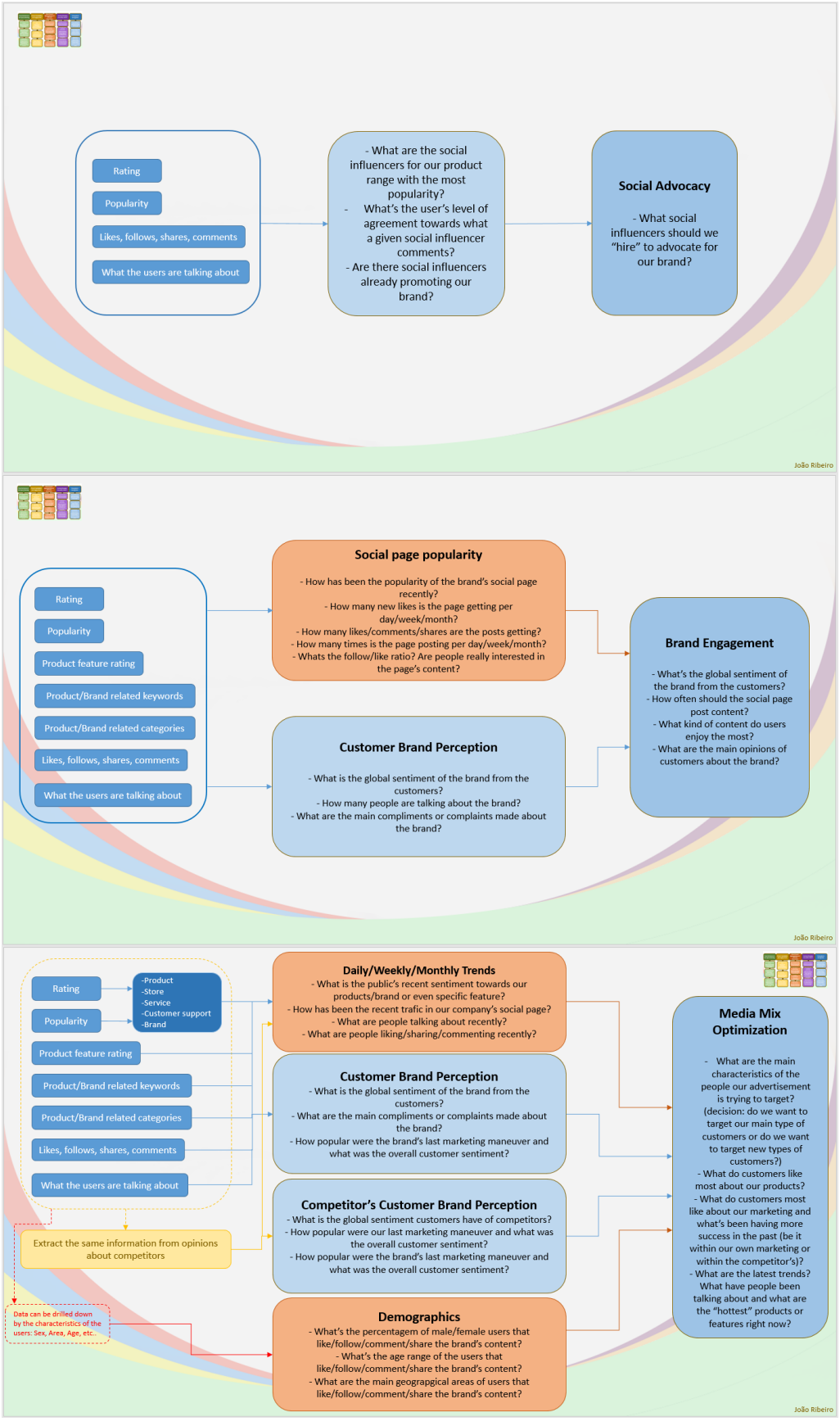


Figure A.8: Customer Xperience Final Project Outline - Pages 22,23 and 24

Appendix B

Appendix 2 - Data Discovery Platform - Data Sample

Dimensior	Category	Media	Sentimen	Date	Social ID	Subscriber Nam	State	Latitude	Longitud	Comment	Spell Corrected Text (Twitte	Bank Name	Message Sourc	Friends(Twitter	Listed (Twitte	Like Count	Share Count	No of Comments (Faceboo	Gender	Keywords	Handle	Date_Created
Account	Account	Twitte	Neutral	30/04/2018 17:29								DEBENHAM	Tweets	490	4	0	0		Male	account		01/05/2018 00:19
Account	Account	Twitte	Negative	07/05/2018 13:36								DEBENHAM	Tweets	408	0	0	0		Female	account		07/05/2018 23:26
Account	Account	Twitte	Negative	28/03/2018 07:37								DEBENHAM	Tweets	247	0	0	0			account		29/03/2018 02:37
Account	Account	Twitte	Negative	27/03/2018 13:02								DEBENHAM	Tweets	843	9	0	0		Female	account		28/03/2018 00:35
Account	Account	Twitte	Negative	16/03/2018 11:56								DEBENHAM	Tweets	710	2	0	0		Male	account		19/03/2018 07:10
Account	Account	Twitte	Negative	25/02/2018 00:28								DEBENHAM	Tweets	621	1	0	0		Male	account		26/02/2018 01:29
Account	Account	Twitte	Negative	25/03/2018 22:51								DEBENHAM	Tweets	31	11	0	0		Male	account		26/03/2018 00:12
Account	Account	Twitte	Neutral	11/03/2018 20:06								DEBENHAM	Tweets	1017	4	0	0			account		11/03/2018 22:09
Account	Account	Twitte	Negative	06/02/2018 13:55								DEBENHAM	Tweets	1486	123	0	0			account		10/02/2018 07:49
Account	Account	Twitte	Negative	28/03/2018 21:33								DEBENHAM	Tweets	156	10	0	0		Female	account		29/03/2018 02:37
Account	Account	Twitte	Negative	02/03/2018 00:50								DEBENHAM	Tweets	3	0	0	0		Female	account		02/03/2018 22:36
Account	Account	Twitte	Negative	23/12/2017 00:02								DEBENHAM	Tweets	177	1	0	0		Male	account		10/02/2018 07:49
Account	Account	Twitte	Negative	05/03/2018 20:42								DEBENHAM	Tweets	487	26	0	0			account		05/03/2018 22:13
Account	Account	Twitte	Negative	26/12/2017 16:18								DEBENHAM	Tweets	34	0	0	0		Male	account		10/02/2018 07:49
Account	Account	Twitte	Negative	01/04/2018 17:09								DEBENHAM	Tweets	630	0	0	0			account		01/04/2018 23:37
Account	Account	Twitte	Negative	09/01/2018 16:40								DEBENHAM	Tweets	172	0	0	0			account		10/02/2018 07:49
Account	Account	Twitte	Negative	03/05/2018 13:37								DEBENHAM	Tweets	70	1	0	0			account		04/05/2018 02:42
Account	Account	Twitte	Negative	15/04/2018 10:11								DEBENHAM	Tweets	76	0	0	0		Female	account		16/04/2018 12:49
Account	Account	Twitte	Neutral	06/01/2018 17:27								DEBENHAM	Tweets	474	12	0	0			account		10/02/2018 07:49
Account	Account	Twitte	Negative	09/04/2018 11:32								DEBENHAM	Tweets	186	3	0	0			account		09/04/2018 12:33
Account	Account	Twitte	Negative	17/12/2017 14:50								DEBENHAM	Tweets	180	0	0	0		Male	account		10/02/2018 07:49
Account	Account	Twitte	Negative	28/12/2017 13:41								DEBENHAM	Tweets	231	5	0	0			account		10/02/2018 07:49
Account	Account	Twitte	Negative	05/01/2018 14:43								DEBENHAM	Tweets	122	1	0	0			account		10/02/2018 07:49
Account	Account	Twitte	Negative	18/12/2017 23:05								DEBENHAM	Tweets	307	7	0	0		Male	account		10/02/2018 07:49
Account	Account	Twitte	Negative	03/05/2018 17:20								DEBENHAM	Tweets	358	14	0	0			account		04/05/2018 02:42
Account	Account	Twitte	Negative	12/12/2017 08:20								DEBENHAM	Tweets	374	15	0	0		Male	account		10/02/2018 07:49
Account	Account	Twitte	Negative	16/12/2017 00:20								DEBENHAM	Tweets	221	3	0	0		Female	account		10/02/2018 07:49
Account	Account	Twitte	Neutral	21/12/2017 19:13								DEBENHAM	Tweets	807	26	0	0			account		10/02/2018 07:49
Account	Account	Twitte	Negative	29/03/2018 09:23								DEBENHAM	Tweets	526	11	0	0		Male	account		29/03/2018 23:20
Account	Account	Twitte	Neutral	02/05/2018 20:38								DEBENHAM	Tweets	70	1	0	0			account		03/05/2018 03:19
Account	Account	Twitte	Negative	27/03/2018 10:34								DEBENHAM	Tweets	246	0	0	0			account		28/03/2018 00:35
Account	Account	Twitte	Neutral	28/02/2018 18:43								DEBENHAM	Tweets	434	0	0	0		Female	account		28/02/2018 23:54
Account	Account	Twitte	Negative	15/04/2018 10:08								DEBENHAM	Tweets	76	0	0	0		Female	account		16/04/2018 12:49
Account	Account	Twitte	Negative	18/12/2017 16:30								DEBENHAM	Tweets	177	1	0	0		Male	account		10/02/2018 07:49
Account	Account	Twitte	Negative	20/12/2017 13:22								DEBENHAM	Tweets	262	4	0	0		Male	account		10/02/2018 07:49
Account	Account	Twitte	Negative	28/04/2018 10:21								DEBENHAM	Tweets	317	0	0	0			account		28/04/2018 23:57
Account	Account	Twitte	Neutral	23/02/2018 10:43								DEBENHAM	Tweets	270	11	0	0			account		23/02/2018 23:34
Account	Account	Twitte	Negative	23/12/2017 09:23								DEBENHAM	Tweets	231	4	0	0			account		10/02/2018 07:49
Account	Account	Twitte	Negative	28/03/2018 18:38								DEBENHAM	Tweets	904	4	0	0			account		29/03/2018 02:37
Account	Account	Facebo	Negative	28/12/2017 08:19								DEBENHAM	status			0	0	0		account		10/02/2018 07:49
Account	Account	Facebo	Negative	03/01/2018 15:41								DEBENHAM	Comment			0	0	0		account		10/02/2018 07:49
Account	Account	Facebo	Negative	22/12/2017 09:48								DEBENHAM	status			0	0	0		account		10/02/2018 07:49
Account	Account	Facebo	Negative	17/01/2018 23:07								DEBENHAM	status			0	0	0	Female	account		10/02/2018 07:49
Account	Account	Twitte	Negative	13/12/2017 14:01								DEBENHAM	Tweets	116	8	0	0		Male	card		10/02/2018 07:49

Figure B.1: Data Discovery Platform - Data Sample

Appendix C

Appendix 3 - Questionnaire 1

Questionnaire 1

Be advised that these questions are regarding data from Facebook and Twitter only.

Please give your answers in red colouring and always in English.

1.Participant Background – Understanding the participant closeness to the social networks.

1.1 How often do you utilize social networks? (please respond with an X)

- Facebook: Never ( ) Rarely ( ) Sometimes ( ) Regularly ( ) Every Day ( )
- Twitter: Never ( ) Rarely ( ) Sometimes ( ) Regularly ( ) Every Day ( )

Even if you don't currently use any social network, please continue the questionnaire based on what transpires to you about the subject in your day-to-day life.

2. Social Network Distinction – Evaluating the differences between Facebook and Twitter.

2.1 How similar do you consider user behaviour in facebook and twitter? (please respond with an X)

Not at all similar ( ) Not very similar ( ) Somewhat similar ( ) Quite similar ( ) Extremely similar ( )

2.2 What characteristics distinguish the two social networks? (select as many as you want with X)

Areas of interest ( )

Contents ( )

Nature of the comments ( )

User segment ( )

User interactions ( )

Others ( ) What others? _____

2.3 If you think there are differences, how much do those differences influence the forecasting process? (please respond with an X)

No influence 1 ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) Strong influence

Figure C.1: Questionnaire 1 - Page 1

3. Key Words for Forecasting – Identifying keywords that imply intent to purchase.

3.1 What key words on a comment do you consider most indicative of a potential future purchase from the customer? Mention at least 5 please.

_____

_____

_____

_____

_____

_____

. . . . . (add more lines if needed)

3.2 Order the 5 most important key words you chose, from most relevant to less relevant.

1-_____

2-_____

3-_____

4-_____

5-_____

4.Mutual User Influence – Understanding how user biasing affects purchasing intent.

4.1 How do you weigh the influence that users exert on each other's opinions? (please respond with an X)

No influence 1 ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) Strong influence

4.2 Assuming a user makes a comment about a product that can induce purchase intent on viewers, how do you weigh the number of friends that user has as a forecasting factor for that product's demand? (please respond with an X)

No influence 1 ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) Strong influence

4.3 How do you weigh the amount of likes on a comment/tweet as an influence on other people's intent to purchase a product mentioned in that comment/tweet? (please respond with an X)

No influence 1 ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) Strong influence

4.4 How do you weigh the amount of shares/retweets of a post/tweet as an influence on other people's intent to purchase a product mentioned in that post/tweet? (please respond with an X)

No influence 1 ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) Strong influence

Figure C.2: Questionnaire 1 - Page 2

Appendix D

Appendix 4 - Questionnaire 1 Answers

Questionnaire 1 - ALL ANSWERS GIVEN

In the checkbox questions, the red numbers represent the number of people that chose that answer. In the keywords questions all of the answers are stated

1.Participant Background – Understanding the participant closeness to the social networks.

1.1 How often do you utilize social networks?

- Facebook: Never (1) Rarely (1) Sometimes (2) Regularly (2) Every Day (3)

- Twitter: Never (7) Rarely (1) Sometimes (1) Regularly ( ) Every Day ( )

2. Social Network Distinction – Evaluating the differences between Facebook and Twitter.

2.1 How similar do you consider user behaviour in facebook and twitter?

Not at all similar (2) Not very similar (3) Somewhat similar (4) Quite similar ( ) Extremely similar ( )

2.2 What characteristics distinguish the two social networks?

Areas of interest (3)

Contents (6)

Nature of the comments (1)

User segment (4)

User interactions (7)

Others (1) What others: Timely matters versus timelessness;

2.3 If you think there are differences, how much do those differences influence the forecasting process?

No influence 1 ( ) 2 ( ) 3 (4) 4 (3) 5 (2) Strong influence

3. Key Words for Forecasting – Identifying keywords that imply intent to purchase.

3.1 What key words on a comment do you consider most indicative of a potential future purchase from the customer? Mention at least 5 please.

WANT(2) LOOKING BUY(4) NEED(3) WISH LIKE(4) BOUGHT INTERESTING USEFUL(2)

GOOD DEAL CHEAP(2) MUST HAVE(3) TRENDY GOOD LOOKING OPPORTUNITY PROMOTION

GOOD PRODUCT PRICE COST HOW MUCH RECOMMEND LOVE EXCELLENT SAVING BARGAIN

FAST DELIVERY DISCOUNT COUPON FREE SHIPPING TOP WEAR NICE AMAZING HANDY ~

GREAT EXPERIENCE RELIABLE AS ADVERTISED EASY CHECKOUT GREAT CUSTOMER SERVICE

Figure D.1: Questionnaire 1 Answers - Page 1

3.2 Order the 5 most important key words you chose, from most relevant to less relevant.

1-BUY

2- MUST HAVE

3- LIKE

4- NEED

5- CHEAP

(This is an approximation of the combination of all of your answers)

4.Mutual User Influence – Understanding how user biasing affects purchasing intent.

4.1 How do you weigh the influence that users exert on each other's opinions? No influence 1 ( ) 2 ( ) 3 (2) 4 (5) 5 (2) Strong influence

4.2 Assuming a user makes a comment about a product that can induce purchase intent on viewers, how do you weigh the number of friends that user has as a forecasting factor for that product's demand?

No influence 1 ( ) 2 ( ) 3 (3) 4 (3) 5 (2) Strong influence

4.3 How do you weigh the amount of likes on a comment/tweet as an influence on other people's intent to purchase a product mentioned in that comment/tweet?

No influence 1 ( ) 2 ( ) 3 (1) 4 (7) 5 (1) Strong influence

4.4 How do you weigh the amount of shares/retweets of a post/tweet as an influence on other people's intent to purchase a product mentioned in that post/tweet?

No influence 1 ( ) 2 ( ) 3 (2) 4 (2) 5 (5) Strong influence

Figure D.2: Questionnaire 1 Answers - Page 2

Appendix E

Appendix 5 - Questionnaire 2

Questionnaire 2

Be advised that these questions are regarding data from Facebook and Twitter only.

Please give your answers in red colouring and always in English.

1.1 Based on the answers given by everyone in the previous questionnaire, would you like to change the five most important keywords that indicate intent for future purchase? If not, leave empty.

_____

_____

_____

_____

_____

2.1 Consider the following real comments, selected based on the keywords you've chosen in the previous questionnaire.

1- "I'm trying to buy my husband some clothes. Won't let me add 2 gift cards?"

2- "Amazing dresses I've seen in store and online. Will have to buy soon."

3- "Can I just have a refund please I'll buy one in store instead thanks"

4- "I want to buy the KVD Saint/Sinner palette but my local store doesn't have the brand, + I can't use Love2Shop vouchers on your website. Is there any way of going into store and paying for it in my local store and having it delivered to the store or my house?"

5- "Wasted journey thanks. Oxford does not have Little Mistress; you must have confused the city of Oxford with Oxford St. London"

6- "Disappointed today. Found a bag I liked on a rack marked 20% off marked price. However, the one I wanted was not 20% off...only designer bags were. It was not on the wrong rack as there were about 4 others. Not good."

7- "I'm sure they were cheaper yesterday online?"

8- "Hi! I just want to know if you have any Cat Von D make up products in your store in Canterbury? Thank you!"

Classify these comments on how much they indicate an intent to purchase. (please respond with an X)

1 - Not at all( ) Not much( ) Somewhat( ) Very much( ) Extremely( )

2 - Not at all( ) Not much( ) Somewhat( ) Very much( ) Extremely( )

3 - Not at all( ) Not much( ) Somewhat( ) Very much( ) Extremely( )

4 - Not at all( ) Not much( ) Somewhat( ) Very much( ) Extremely( )

5 - Not at all( ) Not much( ) Somewhat( ) Very much( ) Extremely( )

Figure E.1: Questionnaire 2 - Page 1

6 - Not at all( ) Not much( ) Somewhat( ) Very much( ) Extremely( )

7 - Not at all( ) Not much( ) Somewhat( ) Very much( ) Extremely( )

8 - Not at all( ) Not much( ) Somewhat( ) Very much( ) Extremely( )

2.2 Consider the following comments regarding the "Jade Winds" dress.

Comment	Likes	Shares/Retweets	User friends
"Hello, do you have the size M for the Jade Winds dress?"	3	0	400
"OMG the Jade Winds dress is AMAZING!! Will be sure to buy ;;;)))"	5	1	500
"HUGH, this jade dress is disgusting, the color doesn't fit any scenario....."	4	0	200
"Hi, I just bought the jade winds dress and I noticed the waist seams are very fragile. A friend of mine also bought it and says the same thing.. Will you be refunding our money?"	5	1	300
"The jade dress is very pretty.. Will pass in the store later to see it"	1	0	600
"Like the Jade winds dress but the price is not for my pocket! Will it be in discount soon?"	2	0	200

Based on the reasoning used until now, make a qualitative forecast for the dress. (please respond with an X)

No purchases 1( ) 2( ) 3( ) 4( ) 5( ) 6( ) 7( ) 8( ) 9( ) 10( ) Large amount of purchases

NOTE: Consider that this forecast is based only on the people that are related to these comments. In that sense, "Large amount of purchases" means that almost all the people that wrote or viewed these comments is going to buy the dress, and "No purchases" means that almost none will.

2.3 If you could, please roughly estimate a quantifiable number for the sales of this dress, based only on these comments. (i.e. 35 units sold)

_____

Figure E.2: Questionnaire 2 - Page 2

Appendix F

Appendix 6 - Questionnaire 2 Answers

Questionnaire 2

In the checkbox questions, the red numbers represent the number of people that chose that answer. In the keywords questions all of the answers are stated

1.1 Based on the answers given by everyone in the previous questionnaire, would you like to change the five most important keywords that indicate intent for future purchase? If not, leave empty.

BUY_____**NEED**_____

MUST HAVE_____**WANT**_____

NEED_____**BUY**_____

LIKE_____**MUST HAVE**_____

CHEAP_____**LIKE**_____

2.1 Consider the following real comments, selected based on the keywords you've chosen in the previous questionnaire.

1- "I'm trying to buy my husband some clothes. Won't let me add 2 gift cards?"

2- "Amazing dresses I've seen in store and online. Will have to buy soon."

3- "Can I just have a refund please I'll buy one in store instead thanks"

4- "I want to buy the KVD Saint/Sinner palette but my local store doesn't have the brand, + I can't use Love2Shop vouchers on your website. Is there any way of going into store and paying for it in my local store and having it delivered to the store or my house?"

5- "Wasted journey thanks. Oxford does not have Little Mistress; you must have confused the city of Oxford with Oxford St. London"

6- "Disappointed today. Found a bag I liked on a rack marked 20% off marked price. However, the one I wanted was not 20% off...only designer bags were. It was not on the wrong rack as there were about 4 others. Not good."

7- "I'm sure they were cheaper yesterday online?"

8- "Hi! I just want to know if you have any Cat Von D make up products in your store in Canterbury? Thank you!"

Classify these comments on how much they indicate an intent to purchase. (please respond with an X)

1 - Not at all( ) Not much( ) Somewhat( ) Very much(**4**) Extremely(**3**) **4,42**

2 - Not at all( ) Not much( ) Somewhat(**3**) Very much(**3**) Extremely(**1**) **3,71**

3 - Not at all( ) Not much(**2**) Somewhat(**5**) Very much( ) Extremely( ) **2,71**

4 - Not at all( ) Not much( ) Somewhat(**4**) Very much(**1**) Extremely(**3**) **4,42**

5 - Not at all(**2**) Not much(**1**) Somewhat(**2**) Very much(**1**) Extremely(**1**) **2,71**

Figure F.1: Questionnaire 2 Answers - Page 1

6 - Not at all( ) Not much(2) Somewhat(3) Very much(1) Extremely(1) **3,14**

7 - Not at all(1) Not much( ) Somewhat(6) Very much( ) Extremely( ) **2,71**

8 - Not at all( ) Not much(1) Somewhat(3) Very much(3) Extremely( ) **3,28**

2.2 Consider the following comments regarding the "Jade Winds" dress.

Comment	Likes	Shares/Retweets	User friends
"Hello, do you have the size M for the Jade Winds dress?"	3	0	400
"OMG the Jade Winds dress is AMAZING!! Will be sure to buy ;;;))))"	5	1	500
"HUGH, this jade dress is disgusting, the color doesn't fit any scenario....."	4	0	200
"Hi, I just bought the jade winds dress and I noticed the waist seams are very fragile. A friend of mine also bought it and says the same thing.. Will you be refunding our money?"	5	1	300
"The jade dress is very pretty.. Will pass in the store later to see it"	1	0	600
"Like the Jade winds dress but the price is not for my pocket! Will it be in discount soon?"	2	0	200

Based on the reasoning used until now, make a qualitative forecast for the dress. (please respond with an X)

No purchases 1(1) 2(3) 3(1) 4( ) 5(1) 6( ) 7(1) 8( ) 9( ) 10( ) Large amount of purchases

NOTE: Consider that this forecast is based only on the people that are related to these comments. In that sense, "Large amount of purchases" means that almost all the people that wrote or viewed these comments is going to buy the dress, and "No purchases" means that almost none will.

2.3 If you could, please roughly estimate a quantifiable number for the sales of this dress, based only on these comments. (i.e. 35 units sold)

8_5__10__15__8__36_____

Figure F.2: Questionnaire 2 Answers - Page 2

References

[1] Wipro Logo. Last access: 24/04/2018. URL: https://upload.wikimedia.org/wikipedia/en/f/ff/Wipro_Logo_1.png.

[2] Enabler Logo. Last access: 24/04/2018. URL: http://natura.di.uminho.pt/join2004/imagens/enabler.png.

[3] Amazon Comprehend – Continuously Trained Natural Language Processing | AWS News Blog. Last access: 23/04/2018. URL: https://aws.amazon.com/blogs/aws/amazon-comprehend-continuously-trained-natural-language-processing/.

[4] Natural Language Understanding Demo. Last access: 24/04/2018. URL: https://natural-language-understanding-demo.ng.bluemix.net/?cm_mc_uid=18095932157115198324270&cm_mc_sid_50200000=13351831520617886516&cm_mc_sid_52640000=23041121520617886523.

[5] Cloud Natural Language | Google Cloud. Last access: 30/04/2018. URL: https://cloud.google.com/natural-language/.

[6] Text Analytics API | Microsoft Azure. Last access: 03/05/2018. URL: https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/.

[7] AWS | Retail | Unified Commerce - Amazon Web Services. Last access: 16/05/2018. URL: https://aws.amazon.com/retail/unified-commerce/.

[8] IBM Watson Analytics for Social Media - Overview - Singapore. Last access: 16/05/2018. URL: https://www.ibm.com/sg-en/marketplace/social-media-data-analysis?

[9] F. Robert. Jacobs, Richard B. Chase, and F. Robert. Jacobs. *Operations and supply chain management*. McGraw-Hill Irwin, 2010.

[10] About Us - Wipro. Last access: 20/04/2018. URL: https://www.wipro.com/about-us/.

[11] Data Mining – How is it Important for Business? - iNurture. Last access: 20/04/2018. URL: http://www.inurture.co.in/data-mining-how-is-it-important-for-business/.

[12] New tweets per second, and how!, 2013. Last access: 21/05/2018. URL: https://blog.twitter.com/engineering/en_us/a/2013/new-tweets-per-second-record-and-how.html.

[13] Web Crawlers — Everything You Need to Know – Cabot Technology Solution – Medium. Last access: 16/05/2018. URL: https://medium.com/@cabot_solutions/web-crawlers-everything-you-need-to-know-6dce26ee8ad8.

[14] Web Crawling: Data Scraping vs. Data Crawling | | PromptCloud. Last access: 17/05/2018. URL: https://www.promptcloud.com/data-scraping-vs-data-crawling/.

[15] Chi-In Wong, Kin-Yeung Wong, Kuong-Wai Ng, Wei Fan, Kai-Hau Yeung, and R de Luis Gonzaga Gomes. Design of a Crawler for Online Social Networks Analysis. URL: http://www.wseas.org/multimedia/journals/communications/2014/a165704-469.pdf.

[16] StormCrawler. Last access: 17/05/2018. URL: http://stormcrawler.net/.

[17] Meesala Shobharani. Related content A Framework for Sentiment Analysis Implementation of Indonesian Language Tweet on Twitter Asniar and B R Aditya - A Peculiar Sentiment Analysis Advancement in Big Data Manisha Valera and Yash Patel - Latent Dirichlet Allocation (LDA) for Sentiment Analysis Toward Tourism Aspect level sentiment analysis using machine learning. 2017. doi:10.1088/1757-899X/263/4/042009.

[18] What Is Amazon Comprehend? - Amazon Comprehend. Last access: 23/04/2018. URL: https://docs.aws.amazon.com/comprehend/latest/dg/what-is.html.

[19] Amazon Comprehend - Natural Language Processing (NLP) and Machine Learning (ML). Last access: 23/04/2018. URL: https://aws.amazon.com/comprehend/.

[20] AboutNLU. Last access: 24/04/2018. URL: https://console.bluemix.net/docs/services/natural-language-understanding/index.html#about.

[21] Customizing. Last access: 24/04/2018. URL: https://console.bluemix.net/docs/services/natural-language-understanding/customizing.html#customizing.

[22] Language Support | Cloud Natural Language API | Google Cloud. Last access: 30/04/2018. URL: https://cloud.google.com/natural-language/docs/languages.

[23] Content Categories | Cloud Natural Language API | Google Cloud. Last access: 30/04/2018. URL: https://cloud.google.com/natural-language/docs/categories.

[24] Text Analytics API overview (Microsoft Cognitive Services on Azure) | Microsoft Docs. Last access: 03/05/2018. URL: https://docs.microsoft.com/en-us/azure/cognitive-services/text-analytics/overview.

[25] What is Retail Analytics? | Sisense. Last access: 03/05/2018. URL: https://www.sisense.com/glossary/retail-analytics/.

[26] Omnichannel Analytics | SAS. Last access: 16/05/2018. URL: https://www.sas.com/en_us/industry/retail.html#omnichannel-analytics/.

[27] Merchandising Intelligence | SAS. Last access: 16/05/2018. URL: https://www.sas.com/en_us/industry/retail.html#merchandising-intelligence/.

[28] Customer Experience Insight | SAS. Last access: 16/05/2018. URL: https://www.sas.com/en_us/industry/retail.html#customer-experience-insight/.

[29] Supply Demand Planning | SAS. Last access: 16/05/2018. URL: https://www.sas.com/en_us/industry/retail.html#supply-demand-planning/.

[30] IoT for Retail | SAS. Last access: 16/05/2018. URL: https://www.sas.com/en_us/industry/retail.html#iot-for-retail/.

[31] More Solutions | SAS. Last access: 16/05/2018. URL: https://www.sas.com/en_us/industry/retail.html#more-solutions/.

[32] AWS | Retail - Amazon Web Services. Last access: 16/05/2018. URL: https://aws.amazon.com/retail/.

[33] AWS | Retail | Customer Engagement - Amazon Web Services. Last access: 16/05/2018. URL: https://aws.amazon.com/retail/customer-engagement/.

[34] AWS | Retail | Data Analytics - Amazon Web Services. Last access: 16/05/2018. URL: https://aws.amazon.com/retail/data-analytics/.

[35] What Is Demand Forecasting & Estimation? | Chron.com. Last access: 04/05/2018. URL: http://smallbusiness.chron.com/demand-forecasting-estimation-32783.html.

[36] How to perform a Multiple Regression Analysis in SPSS Statistics | Laerd Statistics. Last access: 14/05/2018. URL: https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php.

[37] How the USGS uses twitter data to track earthquakes, 2015. Last access: 21/05/2018. URL: https://blog.twitter.com/official/en_us/a/2015/usgs-twitter-data-earthquake-detection.html.

[38] Naren Ramakrishnan, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Khandpur, Parang Saraf, Wei Wang, Jose Cadena, Anil Vullikanti, Gizem Korkmaz, Chris Kuhlman, Achla Marathe, Liang Zhao, Ting Hua, Feng Chen, Chang-Tien Lu, Bert Huang, Aravind Srinivasan, Khoa Trinh, Lise Getoor, Graham Katz, Andy Doyle, Chris Ackermann, Ilya Zavorin, Jim Ford, Kristen Summers, Youssef Fayed, Jaime Arredondo, Dipak Gupta, and David Mares. 'Beating the news' with EMBERS: Forecasting Civil Unrest using Open Source Indicators. 2 2014. URL: http://arxiv.org/abs/1402.7035.

[39] Lawrence Phillips, Chase Dowling, Kyle Shaffer, Nathan Hodas, and Svitlana Volkova. Using Social Media To Predict the Future: A Systematic Literature Review. 2017. URL: https://arxiv.org/pdf/1706.06134.pdf.

[40] Yan Chen, Jichang Zhao, Xia Hu, Xiaoming Zhang, Zhoujun Li, and Tat-Seng Chua. From Interest to Function: Location Estimation in Social Media. URL: https://pdfs.semanticscholar.org/cda1/757d7134885dc9d070f605157e1608e9c204.pdf.

[41] Aron Culotta, Nirmal Kumar Ravi, and Jennifer Cutler. Predicting the Demographics of Twitter Users from Website Traffic Data.

[42] Evangelos Kalampokis, Efthimios Tambouris, and Konstantinos Tarabanis. Understanding the predictive power of social media. *Internet Research*, 23(5):544–559, 10 2013. URL: http://www.emeraldinsight.com/doi/10.1108/IntR-06-2012-0114, doi:10.1108/IntR-06-2012-0114.

[43] The Negativity Bias in User Experience. Last access: 29/05/2018. URL: https://www.nngroup.com/articles/negativity-bias-ux/.

[44] Paul Rozin and Edward B Royzman. Negativity Bias, Negativity Dominance, and Contagion. URL: https://pdfs.semanticscholar.org/b074/3ad11614cf1f4861dc2eca0e4a264b528990.pdf?_ga=2.222560671.586736382.1527591828-739469435.1526908075.

[45] RA Thompson Empathy development and undefined emotional understanding: The early development of empathy. 1987. Empathy and emotional understanding: The early development of empathy. *books.google.com*. URL: https://books.google.co.uk/books?hl=pt-PT&lr=&id=PVQ4AAAAIAAJ&oi=fnd&pg=PA119&dq=Thompson,+R.+(1987).+Empathy+and+emotional+understanding:+The+early+development+of+empathy&ots=Km_R6xmktr&sig=v8RpHFTGfp0aLBiBLeSFBpJjZzY.