FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Customer Xperience - Using Social Media Data to Drive Actionable Insights for Retail

Pedro Abrunhosa Martins

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

Supervisor: Pedro Alexandre Rodrigues João

Second Supervisor: José Nuno Ferreira

February 22, 2019

Resumo

Com a profileração de plataformas de redes sociais e a sua adopção em massa por todo o mundo, o espaço virtual social representa um ecosistema rico em matéria de estudo. O ritmo de produção de novos dados por utilizadores destas plataformas é considerável. Tanto investigadores como interesses comerciais já se aperceberam do valor latente da produção de contéudo em massa presente nas plataformas sociais, surgindo assim a necessidade de ferramentas diversas que operem eficazmente neste espaço tanto no sentido de obter informação, construir conhecimento e descobrir novos padrões e relações, como no sentido de apoiar a criação de valor e a construção de vantagens competitivas.

No papel de apoio às actividades de retalho as redes sociais desempenham dois papeis: Por um lado são o ponto de contacto directo com o cliente, o canal através do qual é estabelecida comunicação birdirecional e por onde são executadas campanhas de marketing e ao mesmo tempo ações de apoio ao cliente; Por outro são uma fonte inesgotável de informação sobre o consumidor. Este projecto surge com foco no segundo aspecto da relação do retalho com as redes sociais. Foi identificada a necessidade de recolher, processar, interpretar e analisar em massa a opnião do consumidor por forma a gerar informação operacional que auxilie decições executivas e operacionais nas actividades de retalho. Para tal foi criado um produto desenvolvido através da integração três sistemas distintos, por via das suas respectivas APIs, recorrendo a um programa de automação e controlo baseado em Python. Depois da definição dos parâmetros iniciais de pesquisa, o programa faz uma seriação e recolha de submissões de utilizadores da plataforma Twitter, garante a execução de processos de NLP para extração de sentido e polaridade de sentimento expressa no texto e passa a pacote final de texto e processamento a uma base de dados não relacional capaz de produzir visualizações via funções *MapReduce*.

Abstract

With the proliferation of social media platforms, as well as a global mass adoption, the social virtual space represents an ecosystem that's rich in research matter. The rate of data production by users is considerable. So much so that both researchers and investors have realized the inherent value of social media data and a need has arisen for proper tools that can effectively operate in that space. For obtaining information, build knowledge and uncover new patterns and relations, but also tools that create value and build a competitive advantage.

In the role of aiding retail activities, social media plays two roles: The first role is that it lays out a bi-directional communication channel for direct contact with the customer, through which both marketing moves are broadcast, but at the same time customer support is offered; The second role is social media is an inexhaustible source of data on customers. This project focuses fundamentally on the second role. A need was identified for solution able to discover, extract, process, interpret and analyse massive volumes of opinionated user submissions on social media so that actionable insight is produced for aiding key retail activities. In that sense, a product was developed that consists of the integration, control and automation of three distinct systems through their respective APIs, using Python. After defining initial search parameters, the program discovers and extracts user submissions to the Twitter platform, guarantees execution of NLP tasks for element extraction and sentiment polarity analysis and passes the final package of text plus analysis to a non relational database where from data can be extracted or visualised using MapReduce functions.

Agradecimentos

Antes de mais quero agradecer à minha Mãe. Pelo seu esforço a criar um filho estarlhe-ei para sempre grato. De seguida quero agradecer à minha namorada, por quem tenho um amor tão natural e inquestionável como o ar que respiro. Agradeço também aos meus dois orientadores, José Ferreira e Pedro João, que me guiaram por esta viagem sem me deixar perder firmeza. Finalmente agradeço aos Professores Maria Antónia Caravilha e José Oliveira pelo esforço e dedicação aos alunos da especialização em Gestão Industrial no MIEEC

Pedro Abrunhosa

"I have a lot of beliefs and I live by none of them."

Louis C.K.

Contents

List of Figures

Abreviaturas e Símbolos

API	Application Programming Interface
JSON	Java Script Object Notation
ML	Machine Learning
NLP	Natural Language Processing
NLU	Natural Language Understanding
REST	Representational State Transfer
SA	Sentiment Analysis
SMA	Social media analytics

Chapter 1

Introduction

The first chapter aims to convey the context and motivation inherent to the project being developed within the Masters Course in Electrical and Computers Engineering at Facudade de Engenharia da Universidade do Porto as well as a general overview on the company hosting the project, *Wipro Portugal*.

1.1 Hosting Company

Wipro Limited is an Indian multinational focused information technology, consulting and business process services. With over 160 000 employees in 62 countries, *Wipro* has a vast portfolio of services available to clients, made possible by leveraging technologies and expertise such cognitive computing, hyper-automation, robotics, cloud and analytics to build modern solutions for modern problems. *Wipro* promotes a culture of corporate citizenship, a commitment to sustainability and adherence to ethical standards. The companies core values and guiding principles are summarized by the *Spirit of Wipro*:

- **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."

1.1.1 Wipro Portugal

Established in 1997 under the name *Enabler* - at the time, a spin-off from *Sonae*, a Portuguese multinational mainly focused on Retail - the company was acquired by *Wipro Limited* in 2006 and renamed as *Wipro Portugal S.A.*. The company specializes in developing IT solutions for retailers based on the *Oracle Retail Suite*, having been distinguished as an *Oracle Retail Center of Excellence* based on the expertise and proficiency of implemented solutions.

Services provided by *Wipro Portugal* include:

- Business Transformation: Substituting legacy systems for *Oracle Retail* software products, adapted to customer needs.
- Implementation: Implementation of *Oracle Retail* packages, integration and migration of other applications and tools.
- Application management and support: Teams distributed throughout Portugal, India and Brazil, guaranteeing management and support 24/7.
- Consulting: Industry specialists capable of evaluating and advising on different areas of business.

1.2 Context

Through the continuous growth of social media platforms, user generated data is at an all time high with every indication of steady growth. The advent of social media brought about the opportunity to gain new insight into how customers act before and after purchases, how brand relations play out in the new medium and what new forces drive consumption or the lack thereof, all by virtue of processes able analyse natural language.

Development of machine-learning algorithms with sets of big data has shown tremendous results for a variety of scenarios, with numerous case studies readily available. Such is the fate of Natural Language Processing (NLP), a research area that at it's genesis relied on laborious manual labeling of Parts of Speech (PoS) among other NLO tasks, and that is now growing heaps and bounds due to the development of machine-learning algorithms.

Understanding consumer behaviour has long been a priority for goods and service providers. As such, companies push for more sophisticated methods of understanding the patterns ruling consumer behaviour. The advent of social media brought about the opportunity to gain new insight into how customers act before and after purchases, how brand relations play out in the new medium and what new forces drive consumption or the lack thereof.

1.3 Project Description and Objectives

The project at hand aims at conceiving a reliable way to discover, extract, process, store and analyse data stemming from social media platforms in an effort known as opinion mining. The ever growing amount of data produced by users on social media seems to be a promising source of insights into customer behaviours, brand engagement, customer experience, trend determination and understanding of influencing agents' reach and impact on public perception and intent of purchase.

This is achieved by gauging consumer opinion, sentiment and emotion on identifiable topics and entities and applying opinion summarization techniques so that a clearer picture is formed and insights can be extracted regarding the entity under scrutiny. The final goal is enabling the creation of driving feeds for retail activities such as Pricing, Store Layout, Customer Experience, Targeted Campaigns, Allocation of stock, and Planning. As such, the proposed solution should be an automated process with a data source and discovery parameters limiting breadth and scope of the search as inputs and an output feed consisting of fully processed and analysed data for the purpose of driving said activities.

For unstructured, user created text data, from which entities will be identified and opinion and sentiment mined, NLP techniques will be employed. Structured and numerical data and metadata will be used to clarify context surrounding user text through traditional means of analysis.

Data storage should allow for distributed access and versatile querying. Given the unstructured nature of the bulk of collected data, a NoSQL approach is preferred.

The ultimate goal of this project is to leverage social media data so as to produce unique actionable insights for retail through a solution that can scale, expand in scope and that's easy to deploy to areas not related to retail but that may benefit from social media data gathering and opinion mining.

1.4 Adopted Methodology

Choosing a methodology is contingent on the task at hand. Tasks are defined so as to further the project towards the defined objectives. An overview of project tasks and corresponding methodologies goes as follows. To define realistic expectations of what type of data is to be found when mining social media, as well as what sorts of insight such data may provide, thorough research on the subject was conducted. With a clearer picture on what to expect, research was conducted for finding the right tools for the job of extracting and processing data.

In the meantime, it was to be expected that gathering data from unstructured text wouldn't amount to a trivial task. Techniques making up a Natural Language Processing toolkit where researched, looking for practicality, functionality and effectiveness.

In terms of volume, it was expected gathered data would amount to tens of thousand of text units, each unit consisting of each user's individual text submission, regardless of size. The expected results of processing each unit for analysis gave rise to the need of a robust way of dealing

with such amounts of data. Given that most of the data gathered was to be unstructured by nature, database chosen had to be able to deal with object oriented documentation. As before, research was to be conducted, with the aim of confirming how this is being done and what methods suit the project best. Development of solutions followed a reasoning of first and foremost gathering the low hanging fruit and accomplishing a proof of concept capable of being improved upon if stipulated time for project completion allowed it.

1.5 Document Structure

The following document is composed of five more chapter, aside from the introduction, namely: Literature Review, Current Situation Analysis, Proposed Solution, Project Development and Results, and Conclusion and Further Improvements.

In chapter 2 the full theoretical framework this project is based in will be explained and relevant concepts to further understanding of the project as a whole will be provided.

In Chapter 3 an overview of the field of retail is given, specifically concerning activities governed by the Oracle Retail Suite, as well as an overview of current techniques and objectives of Natural Language Processing and Sentiment Analysis applied to retail. Chapter 4 delves into a proposed solution where it is discussed which tools where used and to what end, as well as what each technique brings to the table and furthers the project. In chapter 5 a detailed explanation of the progress the project has seen and each increment was added since it's birth and how the finished product was arrived at. Results will be presented for discussion. Chapter 6 is where conclusion are discussed future improvements laid out.

Chapter 2

Literature Review

This chapter delves into the various scientific and technological domains relevant to the project at hand, mainly relating to data, data acquisition, data analysis and data storage, several retail areas and new technologies applied to the project

2.1 Social Media Data

Social media is understood to be the collection of platforms where users create, share and exchange content, following the technological and ideological foundations of Web 2.0 [5]. It gives users an unprecedented way to communicate and collaborate on content creation. Social media platforms are growing in size and number. As such, vast quantities of data are produced on a daily basis . Considering aspects of volume, velocity and variety (3Vs), as defined by [6], social media data can be labeled as "big data".

- **Volume** - Evaluating volume of data depends on context. Data sets equal in size but varying in type may need very different processing power for analysis, making one high volume and the other easily manageable. Social media generated data is neigh inexhaustible, such is the speed it's being produced at [1].

- **Variety** - Variety in data occurs when several types of data are extracted simultaneously, namely structured data, consisting of data that can be tabulated or fed into a relational database and unstructured data, that has no predefined structure or data model, and semi-structured data, i.e. data that carries no separation between data and schema. A social media submission has several metrics associated - view counts, "likes", "favourites", that can be classified as structured data. The text, video or image shared by the user makes up unstructured data. Meta data accompanying each post makes up most of the semi-structured data on social media.

- **Velocity** - Velocity is the speed at which data is generated. Mass availability and mass adoption of smart devices result in rising amounts of interactions on social media platforms

each minute. Figure 2.1 gives a by-the-numbers overview of the rate at which interactions occur.

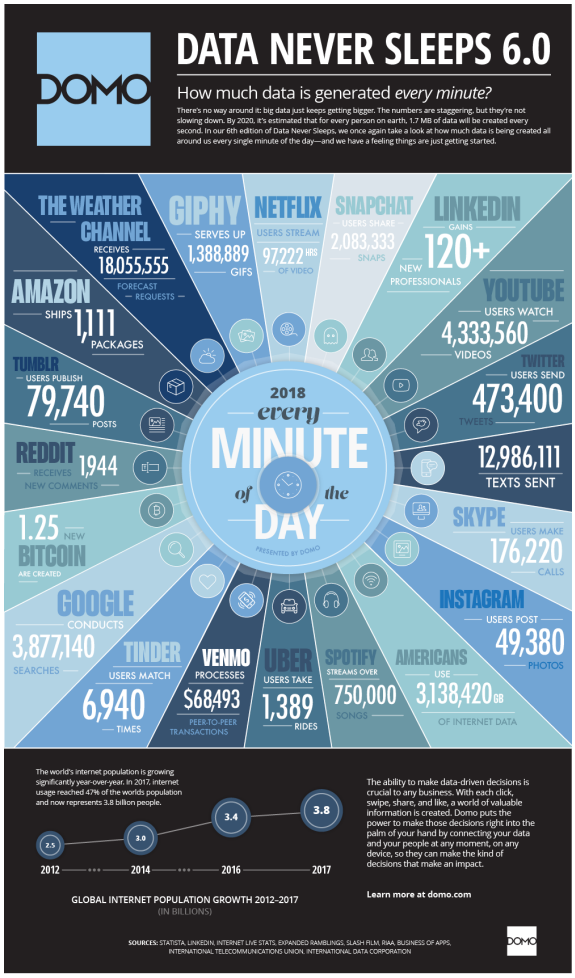

Figure 2.1: Data generated every minute on a variety of platforms for 2018 [1]

Data generated on social media platforms is considered Community Data. Regarding it's composition, already showed to be varied enough to be considered Big Data, unstructured data is to be found mostly in the form of text, image and video and structured data in the form of metrics that illustrate appreciation, influence, reach and in some cases, condensed, fully fledged opinions. The first constitute the actual content produced or at least share by the user, the second are manifested through "like" buttons, share links, favourites, etc. Shortcuts for opinion are often found on review sites and Customer Support section, in the form of an "I found this useful" button [7].

The 3V model for describing big data isn't fixed or established, meaning some researchers actually suggest other categories. It seems sensible to discuss a fourth V, Value, when regarding sentiment analysis on a large scale, since what's being gauged is the overall sentiment by the

customer toward a given entity. High volume and high velocity make it a necessity to have and efficient storage system so data can be amassed and value accrued. It's data aggregation that enables analysis that ultimately answers business questions and paint a clearer picture of the market. [8].

2.1.1 Sources

Data generated on a social media platform is highly dependant on the functionalities of the platform. Online social networks (e.g. Facebook, LinkedIn) allow users to connect and form communities. Interaction happens through status updates, comments, sharing of media messaging services; Blogs generally allow for broadcasting of long form text formats and media. Usually they are maintained by a single user or group (e.g. Blogger and WordPress); microblogs (e.g. Twitter and Tumblr) are similar to blogs, but users broadcast short text or media links and usually implement a subscription based feed that presents the latest or most relevant status update from subscribed users; media sharing platforms (e.g.YouTube and Instagram) are mainly used to share multimedia content; Opinion, review and ratings platforms (e.g. Yelp and TripAdvisor) allow for users to share their subjective views on businesses, services or products [5].

Of the platforms mentioned above, undoubtedly Facebook and Twitter are the preferred platforms for bi-directional communication between business and customers [9]. However, the content of such communication varies substantially between platforms, by way of the environment the communication takes place in. On Facebook, a brand has a commercial page where they share status updates and media and user feedback comes mostly in the form of reactions and comments. Comments are rich in unstructured data with value for analysis, but they are likely bounded by the context surrounding the status. Twitter, on the other hand, allows for more dynamic and free communication between a brands and consumers because no user page exists and no moderation or context to bound discourse. Brands can Tweet at all user subscribed to their feed, users can Tweet at brands by tagging them in the Tweet and brands can respond by tagging the user in the response Tweet.

Extracting social media data from a platform is contingent on availability of data, availability of tools, popularity of the platform and ease of analysis [10] making these core factors when choosing a data source. Communications between customers and businesses through these platforms is largely publicly available. Tools will be discussed in 2.1.2. By the numbers, Facebook is more popular than Twitter, with over 2271 million active daily users [11] to Twitter's 326 million monthly active users [12], but both have enough users to extract meaningful data.

2.1.2 Data Acquisition

There are essentially two methods for extracting data from social media: using an Application Programming Interface (API) or, alternatively, web crawling and scraping indexed pages. Proprietary APIs allow users to interface two software systems and essentially make requests remotely. Both Facebook and Twitter have proprietary Representational State Transfer (REST) APIs made available to developers. These APIs are capable of fulfilling a vast number of varied requests and

are used as tools for anything from targeted advertising to managing accounts. Keeping with the scope of this dissertation, endpoints of interest are those that allow for data discovery and extraction from the platform. Twitter makes their timeline (i.e. every tweet since the first, posted in 2006) available through the Twitter Search API. Facebook used to make their public data available through the Graph API but has since changed that policy [13].

A web crawler is a bot that traverses and indexes every page on a website, following hyperlinks it finds along the way. They are widely employed for many different purposes, most notably page indexing for web bases search engines. Basic data like page URLs and hyperlinks are gathered and stored and often Hypertext Markup Language (HTML) checked for syntax. Paired with a web scraping tool, this allows for a very powerful data extraction tool. Web scraping tools parse HTML structures and Extended Markup Language structures that make up a web page and extract data contained in those structures. It essentially copies an entire page into local storage, making it available for analysis. For such platforms that present data as continuous feed with endless scrolling built in, data discovery through web crawling becomes inefficient. In this case, automated scrolling of the feed can be implemented through software.

APIs are the preferred method for data extraction when available, however that isn't always the case, be it because the platform is unwilling (i.e. keeping competitive advantages, privacy policy, etc.) or unable (i.e. legal requirements, technological deficits, etc.) to provide such a service.

Revisiting the above example, Facebook altered policy surrounding their public API and now doesn't allow for mass extraction of user data from public pages. In such cases data can and should be extracted via crawling or scrolling through a feed and scraping, though software should conform to both legal and polite practices when doing so. Server-side protection may be put in place to prevent bots from operating on the network. These checks work by detecting patterns that don't appear to be human interaction, which can be simulated, and by detecting high rates of requests to the server, in which case the control is put in place to prevent attacks on the server that may cause down-time for regular users. Furthermore inclusion of "/robots.txt" files by web site owners so as to instruct robots on what parts of the server are off-limits for crawling [14] is a viable way to enforce polite web crawling and scraping. Bad agents are able to simply ignore this file and carry on, however.

2.2 Social Data Analytics

The characteristics of social media platforms make analysis of data generated by its users a matter of identifying trends that reveal risk or opportunity buried in the communication's and derive insightful actions in accordance. Social media can and should be the primary channel for communication with a target audience, specially for broadcasting messages and immediate availability of the brand for the customer. Communications flowing both ways can be analysed for a variety of attribute's like entities and sentiment regarding them.

Information spreads through the network at very high speeds in social media, making flexibility and reactivity in response to insights discovered an important factor in competitiveness [15].

As mentioned, Social Media Analytics (SMA) entail three major steps. A rephrasing suggested by [15] is "capture", "understand" and "present". The framework for SMA presented in [16] goes as follows:

- **Tracking -** Decisions on the data source, approach, method and output;
- **Preparation -** noise removal and data packaging for next step;
- **Analysis -** Any and all methods to be applied, from network analysis to sentiment analysis.

The analysis process must be further broken down in order to get a clearer picture at how insight is generated. Suffice it to say that for the most part, meaning will be extracted from unstructured data and paired with context created from structured data, upon which, through aggregation of all posts around the same topic, a general feel of the public opinion for said topic will surface.

2.2.1 Sentiment Analysis

2.2.1.1 Occurrence of "Brand talk" in social media discourse

Research conducted in 2009 shows that from a sample of 150,000 Tweets, 19% of microblogs contain mentions of a brand. Of these 20% express an opinion or sentiment [17]. More recent data show that 75% of B2B and 65% B2C brands market on twitter, meaning they have since broadned their outreach and availability, inviting many more opinionated submissions

"Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes" - [18].

Opinion influences behaviour, making it the crux of social media data analysis. Opinionated user submissions have helped revamp businesses, sway public sentiment and overall influence the public sphere [18]. But Opinion also serves as a predictor for emerging trends because of it's potential for studying human behaviour [10]. Results of sentiment analysis are usually by way of polarity - negative, positive or neutral - and/or by way of a scoring system, e.g. -1 to 1, also indicative of polarity.

There are essentially three levels of analysis for sentiment:

- Document level - Analysis results in overall sentiment polarity of the document. For any text containing more than one entity, no valuable information can be extracted aside from a general appraisal.
- Sentence level - A sentence is nothing more than a small document and as such, the same reasoning applies
- Aspect level - Isolating an entity and its different aspects and being able to extract sentiment polarity allows for the most complete interpretation of an opinionated text.

To further clarify Aspect level SA, [18] poises that opinion, in the context of analysis should be though of as a quintuplet, $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$, where e_i denotes entity i, a_{ij} is the aspect j of the entity the opinion is about, h_k is the holder k of the opinion, t_l is the instant in time l the opinion is held in and s_{ijkl} is the sentiment on aspect a_{ij}. Most every opinion can be dissected to fit this quintuplet, and certainly every opinion expressed by a review. [18]. By the structure alone, it becomes self evident that each opinion has a holder and is only valid in a moment in time, and only about a given aspect of an entity, making it very context dependent and thus subject to change through definition of new business process that transform the customer experience.

To extract value from opinion, a high volume on the same entity must be collected. Insights are discovered by how the public leans in opinion regarding some predefined subject. Regarding the same aspect, no two differing opinions are more or less right, rather, both must be taken into account to arrive at the general opinion that most resembles the truth as perceived by the customer base.

Methodologies employed for sentiment analysis are as follows: Sentiment analysis can be performed mainly in two ways: The first is a lexicon-based approach. Be it by comparing to a dictionary where each word has a sentiment value attributed dependent on the context of the sentence or be it corpus based, where context can be analysed and compared and a statistical or semantic score is given to each possibility of meaning, with the highest score prevailing. The second approach is through machine learning. Using supervised or unsupervised learning methods where each word in each context of a sentence or paragraph has a certain weight that adds up to a final sentiment polarity score [2]. Figure 2.2, page 11 shows the different methodologies for sentiment analysis.

One set of challenges facing sentiment analysis stem mostly from the human element in text production. Be it misspellings, sarcasm or pure deception, result validity decreases as a function of lack of clarity in communication, or communication through context clues the algorithms are not able to pick up on, at present. [19]

It's of note that in terms of opinion extraction, brevity of text might prove an advantage. In social media platforms focused on microblogging - such as Twitter, when opinion is presented, it has to be compact and explicit due to text size constraints. Research suggests it's easier to analyse sentiment in short form microblogs than long form blogs, regarding classification accuracy. [20]

2.2.2 Natural Language Processing

Natural Language Processing (NLP) is the means by which computers process and analyse natural language. SA can be considered a sub-field of NLP, but for the project at hand, the NLP toolkit serves the purpose of preparing text for sentiment analysis. As with Sentiment Analysis, NLP tasks became far less laborious and far quicker in the advent of machine learning algorithms. NLP has three main task domains of action: Syntax, Semantics and Discourse. It's by performing tasks in these domains NLP creates understanding of natural language by computers. Syntactic analysis, or parsing, breaks down phrases into their components and determines the relation between components.

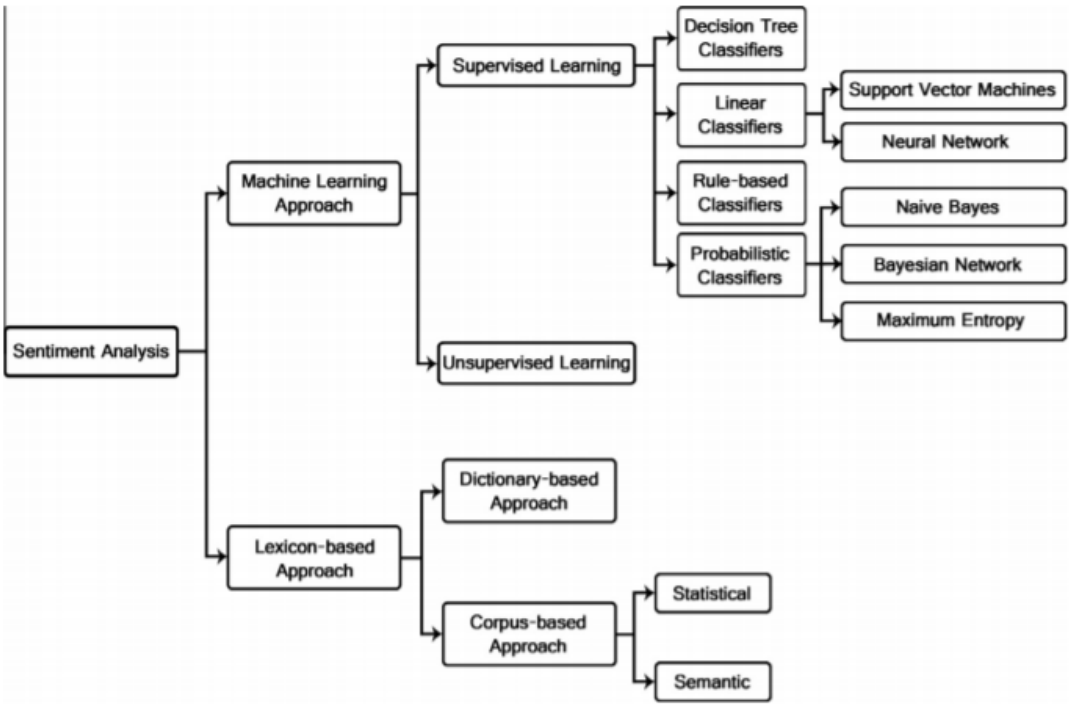

Figure 2.2: Sentiment classification techniques [2]

Examples of NLP syntax tasks are part of speech tagging, stemming (i.e. determining the derived or inflected word's stem), and terminology extraction (i.e. retrieving relevant terms from text). Semantics tasks retrieve intelligible meaning from the words, something humans do unconsciously but is the biggest source of challenge in NLP. Interpretation of signs an attribution of meaning is at the base of automated chat box and computer generated text and speech as well as speech understanding, just to give some examples. In the domain of discourse, NLP tasks are, for example, coreference resolution, i.e. finding all attributes of the same entity mentioned in a text document, paragraph or phrase. What enables this is first and foremost the technology of categorization, or rather, the ability to group distinct entities into the same categories (e.g. rice, bread, steak are meronym of food). Categorization is most easily done through machine learning, achieved by first running a test data set, where each item is assigned manually, and feeding it into a machine learning algorithm, which in turn, all going well, is able to apply what it has learned from the patterns detected in the manual categorisation and apply it across new data that is fed. Another method to categorize is through a rule based approach. Checking Figure 2.2 again, the different ways of going about these tasks still apply, because, as was said, SA is merely a sub-task of NLP. The order of appearance was stitched to emphasise importance of SA to the final solution of the project. A rule based approach starts out as a set of categories manually defined by specialists, wherein each categories is governed by a set of rules that determine if a word belongs to that category or not. To be clear, both approaches result in the ability to correctly identify and extract entities, therefore aspects of said entities, and therefore make and informed guess on the polarity

of opinion or class of emotion with sensitivity to context.

Furthermore NLP allows to connect entities and uncover relations between them. Relations between entities are just as rich in valuable information as opinion on entities, and suitable targets for sentiment analysis.

Chapter 3

Current Situation Analysis

In this chapter Retail Supply Chains' biggest challenges and current implemented tools ans solutions will be analyzed and discussed, thereby identifying needs and shortcomings in need of being tackled by this project's proposed solution.

3.1 State of the Retail Supply Chain

In [3], a report produced in 2017 laying out the State of the Retail Supply Chain, three groups of challenges were outlined as the biggest and most in need of support for retail activities. The report was composed after interviewing 80 retailers with €100m p.a. sales, from over seven European countries. Presented statistics are sourced from this document, unless explicitly said otherwise. The highlighted challenges are:

- *Promotions*: Growing competition - choice and availability, in the eye of the customer - has consumers asking "when" and "where" to shop. Promotions are an attempt of answering these questions with a "here" and "now". 40% of companies interviewed mention supply chain planning of promotions as their single biggest challenge and 82% are dissatisfied with their current promotion and new line planning system.
- *Forecasting*: Effective forecasting has long been an important research subject for sales industries and is still on the list of retailers biggest challenges. 76% of interviewed retailers reported discontent with the current system of analysis and reporting on forecasting.
- *Use of external data and space data*: Reportedly, few retailers integrate system for using external data, with some doing so manually. Many use no external data at all when forecasting, and thus lose out on valuable data for more reliable forecasting models.

Another important insight the report in establishes is a list [3] of analysis and reporting retailers wish they had. This can be looked up in Figure 3.1, page 14. Particularly relevant are the 70% of retailers wishing for more accurate forecasting of future out-of-stocks, 54% wanting near-time replenishment and forecast calculation and 47% wanting analysis on allocating scarce stock based on forecast margin contribution.

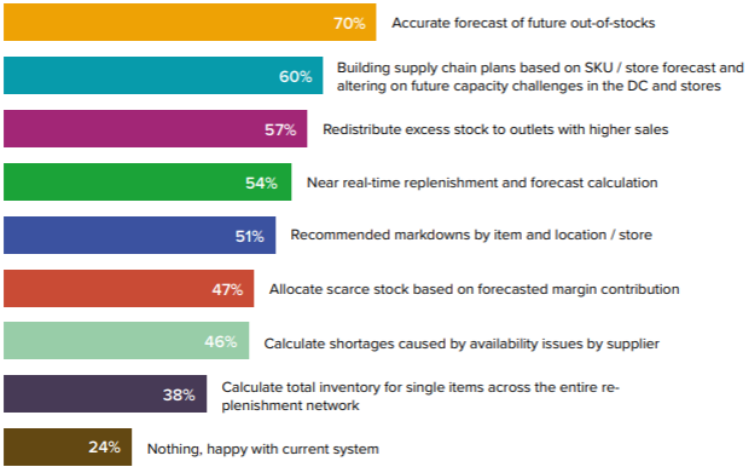

Figure 3.1: "Top Three Issues Regarding Supply Chain Planning and Execution" [3]

3.2 Oracle Retail

The following retail activities are simultaneously relevant to the project and handled by Oracle through one or more products entailing the Oracle Retail Suite.

3.2.1 Inventory Planning

Oracle describes inventory as the retailer's largest financial investment and one of the most complex operations to manage efficiently. In agreement with the aforementioned report, balancing cost of inventory with cost of out-of-stock is a continuous task. Growth of availability is making cost of out-of-stock grow considerably.

Oracle Retail Inventory Planning creates optimized inventory targets by location to meet demand and satisfy business and financial objectives. [21]

3.2.2 Allocation

Also in tune with the report, Oracle claims allocating resources to be vitally important. The right item at the right time is an excellent precedent any sale and grows customer satisfaction. In the age of omni-channel retail agile and efficient allocation are a must-haves in order to stay competitive.

Oracle Retail Allocation translates merchandise plans into location level allocation in order to fulfill customer order demands while being aware of impacts on warehouse holdback quantities. Currently, Oracle allows for allocation planning both in advance of an orders arrival as well as last minute in order to leverage real-time sales and inventory information. [22]

3.2.3 Customer Engagement

Recognizing the importance of brand awareness and brand loyalty in a landscape where the customer is mainly on the lookout of most bang for their buck, Oracle created and integrated cloud service designed to drive customer loyalty, increase average spend and drive repeat purchases. The software suite cleans, consolidates an organizes data stemming from across the organization so as to be able to connect to the customer with one voice, wherever and however the customer chooses to engage. [23]

3.2.4 Category Management Planning

To meet the need of keeping up with flexibility standards of customers who want to engage with retailers on their own terms, from buying online to visiting a brick-and-mortar store, Oracle built a product that implements consolidation of internal and external data sources into an easily intelligible format to provide retailers with insights regarding all levels of activity, from the national level to the store specific level.

This leverages two modular activities. The first, Category Planning - combining datapoints from various sources to recommend formal category based on consumer insight and product performance that can be fed downstream into assortment planning, pricing, promotions, inventory and space processes. The second, assortment planning, allows to chose between a catalog of approaches based on internal and third-party data to create optimized customer-centric and targeted assortments, unique to each retail store. In combination, the main goal is to maximize customer satisfaction and overall category profitability. [24]

3.2.5 Offer Optimization

Oracle estimates 50% of customers respond to personalized offers and 65% think personalized offers are most important in their shopping experience and compels them to buy. Personalizing offers through omni-channels is a common practice, specially through customer accounts on online retail platforms. This sort of customization can also be adapted to promotions targeted offers and markdowns in order to maximize results [4]. The diagram in Figure 3.2, page 16, showcases how a personalized approach can look like.

3.3 Social Media as a Business Execution Platform

Social media savvy companies already push for a strong presence online, acknowledging that being more connected to their costumers brings huge advantages and increased competitiveness. Even the use of simple readily available metrics such as any platform's equivalent of "likes", shares and tags to gauge content engagement, to then tailor content for increased engagement. Predictive analytics are already employed in order to tailor content for generating the highest buzz metrics (user engagement with the post or submission).

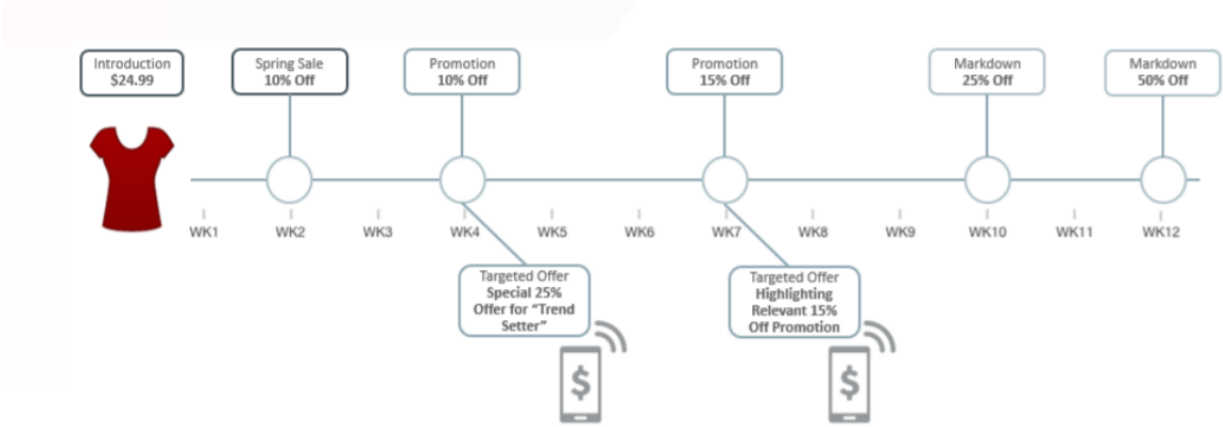

Figure 3.2: Offer Optimization Complete Lifecycle [4]

More sophisticated approaches utilize analytics packages such as Google Analytics™ to enrich a social media data set with a search data set, thus being able to create tailor made targeted ad campaigns, for example.

Even more sophisticated companies rely on predictive analytics and machine learning to produce content that most assuredly will be seen online through interest and sharing, instead of relying on instinct when putting something out. [25]

Social media analytics and artificial intelligence is being tapped into as a source of information and a platform that aids in developing product design, innovation, and managing relations with consumers through customer support and lighthearted, anthropomorphized interactions with the purpose marketing and campaigning, effectively funneling marketing and support through one channel.

Chapter 4

Proposed Solution

The solution outlined in this chapter takes into account the analysis of the current situation regarding retailers, described in chapter 3, specifically targeting their greatest challenges and looks to do so by harnessing the potential of social media data, thereby creating a product capable of tapping into it, preferably in a way that allows for flexibility, portability and scaling.

4.1 Choice of tools

Choosing the right tools for the right job facilitates work all the while yielding greater results. Therefore, this section is dedicated to the tools chosen and the justification behind the choice

4.1.1 Twitter Search API

For the project at hand, the choice of tools is of utmost importance. Regarding sources of data, differences between platforms in regards to quality and subject of discourse and style as well as form of expression, are important aspect to consider when choosing a platform out of which to build a workable, valuable dataset.

For that reason Twitter was chosen as the preferred platform to serve as the source of all user submitted content to be analysed for sentiment, and the reasons are as follows:

User submissions (i.e. Status Updates) are mostly public, making the Twitter timeline (i.e. every status update not deleted in chronological order) a wealth of accessible information without fear of breaching privacy; Twitter's development kit is composed of a multitude of API services that facilitate integrating the platform in applications and business processes seamlessly and with near perfect reliability, thereby rendering useless crawling and scrapping the internet, avoiding all complication arising with this approach, as where previously described.

The Standard Twitter Search API delivers all tweets matching a given search query, out of a considerable sample from the last 10 days. Paid versions allow for full access to the entire timeline, however, for the purpose of establishing proof concept, the standard version was enough for a test trial.

API responses come as a list of Status Objects, displayed in a JSON file, such as the one presented in Appendix A.

In order to interact with the API, the programming language Python was chosen. Python, besides being easy to use for it's high level abstraction, is object oriented and as such easy to integrate with an API. It's one of the most popular programming languages and integrates perfectly with a variety of web applications. Scalability, portability and the potential for developing a web service out of the solution are added benefits of using python for the framework of the proposed solution.

4.1.2 IBM Watson Natural Language Understanding

For Sentiment Analysis and Natural Language processing the choice was made to use cloud computing. Given the predicted volume of data, arrived at by witnessing the rate of incoming status updates with reference to a certain Retailer X on twitter, it seemed sensible to outsource computing power to a full fledged AI based web service capable of delivering an in-depth analysis based on Natural Language Processing with integrated Sentiment analysis. The choice fell on IBM Watson's Natural Language Understanding for reuniting the exact tools that were planned for and being unrestricted in access, rather limiting number of uses. From a certain limit on, each use would accrue a cost.

Watson's NLU is equipped to analyse each extracted tweet and mine it for categories, entities and associated sentiment and emotion, relations, keywords and associated sentiment and emotion, semantic roles, concepts, and lastly, document level sentiment and emotion.

Watson NLU displays sentiment on a scale from -1 (Negative polarity) to 1 (positive polarity), 0 being neutral. Emotion is shown through a combination of five main emotions and a corresponding relevancy score, namely sadness, joy, fear, disgust and anger. The relevance score allows for understanding what if any emotion or combination of emotions is actually being expressed in the text. An example on an analysis, as well as a demonstration of the format the analysis comes in, which is also an object in a JSON file, can be found in Appendix B.

4.1.3 Storage and views

The third API integrated into the system is CouchDB, a NoSQL, non-relational database especially equipped reliably deal with big-data. CouchDB uses the JSON format to receive data and as such is perfectly suited to receive both the Status Object JSON and the NLU Object JSON and allow for viewing and look-ups. CouchDB deals with data processing and visualization through MapReduce functions, or Views in CouchDB. This method filters and sorts documents with a similar structure to JSON through a Map procedure, for example, showing every document with the key "Name" or the value "John", if the document was composed by a list of people. Then it executes a reduce method by which documents are grouped on a common property, such as the number of people named "John" in said list. These methods' advantage is in providing for redundancy and fault tolerance.

4.2 Adopted Approach

The solution proposed consist in utilizing the three described systems in a automated and chained manner. Periodically, the Twitter API is called, the search query run, matched statuses are sent through the NLU API for analysis and the result stored in CouchDB where data analysis is carried out through data visualization methods such as CouchDB's View, List and Show functions. Other analytic tools can be applied, specially to structured data. Influencing the outcome is possible at two separate steps: Changing the search query to be broader or narrower, thereby changing the dataset and alternatively, uncovering the right MapReduce functions in order to present the Views that actually promote actionable insight.

By way of example, a view showing summarized opinion, achieved by the average of sentiment scores, about a relevant category uncovered in a considerable number of tweets. In more practical terms, if a big enough number of tweets identify the entities "employee" and the sentiment is positive the majority of the time, this is an indication of positive customer experience. The opposite could also be true. Generalizing, any entity that is identified with a certain frequency, including derivations on a word or words with the same meaning is prone to be evaluated based on customer opinion and changed in accordance if deemed fit and backed up by the other information systems in place.

Chapter 5

Project Development and Results

5.1 Defining the approach

Originally, the present project was to be a continuation of work based on a foundation laid out by two previous projects developed in the same context and with similar goals. This was concluded to be impractical and unfruitful for two reasons - the tools made available for developing a solution where insufficient both in efficacy and scalability, since they were dependent on external agents for data collection and analysis. As such, it was decided a fresh start was necessary, aiming to achieve the goals originally set out through different means. The proposed solution was thus prepared with no dependencies to past work. This was decided after pursuing all avenues that could facilitate coherence between the different iterations on the Customer Xperience project, including opening communication with the Data Discovery Platform, the service utilized before the current project for data collection and analysis. Communication was established for finding ways of improving the service to allow for achieving the laid out objectives. From this contact, we where able to amass a large volume of data from a number of Retailers, originating mainly from Facebook comments and Twitter status updates. That content, however, proved to be useless in the scope of the project, since text was truncated and entity extraction and sentiment analysis proved to be too inconsistent. Such concerns had been raised when the previous iteration of the the project was underway, but not wanting to rule any possibilities out, another attempt was made in hopes of shining new light on old data. Efforts to pursue this line of development further were soon extinguished.

5.2 Performing data extraction

Starting over presented it's own set of challenges, mainly regarding decisions on how to tackle the problem at hand - extracting useful, actionable insight through social media data analysis. The first decision to be made was what platform data should be gathered from. This was no trivial matter. As mentioned before, every social media platform has an unique social ecosystem based on it's purpose and the communities that formed around it. The checklist defined for choosing the networks that would serve as data source was as follows, in no particular order: a sufficient

amount of users for guaranteed variety of people; public channels of communication between brands and customers; public discourse about brands between customers; a high volume of user generated text (be it by customer or brands); primarily discourse must happen in English. The reasoning behind the first point of the checklist was to avoid as much bias as possible in the data, except for the natural bias of only being able to gather opinions from people that actually engage with the brand or with discourse on the brand. The reasoning behind the second point was to guarantee opinionated data on the brand existed to begin with. The reasoning for the third point was augmenting the quality of opinions broadcast by social media users. Finally, English discourse was a necessity because of barriers of understanding and limitation of analysis tools. China has massive social networking platforms, some bigger than western models with essentially the same functionalities, but language would make those impossible to analyze. The choice came down to Facebook and Twitter, for crossing all points on the checklist.

5.2.1 Settling on Twitter Search API

The tie breaker came down to mapping out the tools most suited to extract data from each source. As was mentioned in Subsection 2.1.2, page 7, Facebook has significantly cut down on access for data aggregation from their platform, alternatives being automating a browser to scroll through brand page feeds and scrape data from the screen, filtering, and cleaning such data afterwards for analysis. If the added work meant more meaningful data, this would have been a fine method to execute. It was deemed, however, that the way user interaction on brand pages works, is not conducive to truly free discourse, as it ends up being bounded by the published content the user is commenting under. Added to this, Twitter, as also mentioned in Subsection 2.1.2, has an API that grants much more freedom for data collection, even in its free version, which for proof of concept was deemed enough. So it was settled. Twitter was chosen to be the source for social media data and the Twitter Search *RESTful* API the medium through which data was to be gathered.

5.2.2 Web page Interface for Tweet retrieval

The Representational State Transfer (REST) architectural style for software facilitates computers on the Internet to operate with each other. Twitter's Search API is HTTP based and as such accepts requests via GET and POST methods for retrieving and sending data to and from the service being accessed. For this reason, a first approach at collecting Tweets was achieved by building a *PHP*-based web page with a very simple interface, see Fig. 5.1 page 23, allowing the user to input every search parameter they wanted to control for and receive a response in the form of a JSON object, as seen in Appendix A.

This proved to be satisfactory proof of concept for the extraction part of the project, however some issues arose that needed to be addressed.

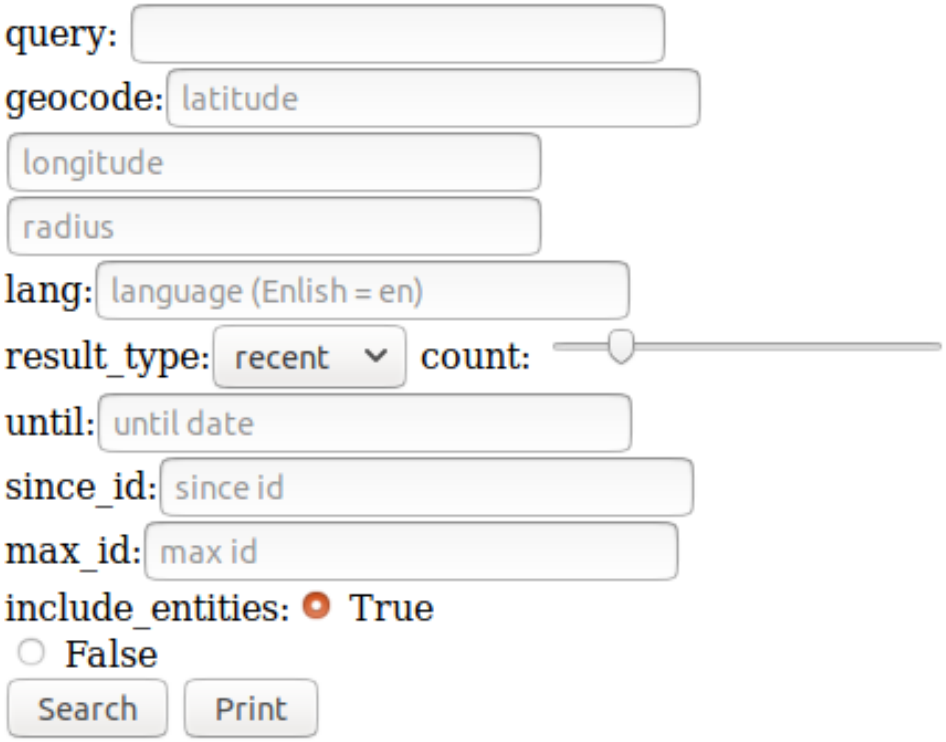

Figure 5.1: Web page Interface for Tweet Search

5.2.3 Move to Python Script Interface

As a programming language *PHP* has many drawbacks, and those soon became apparent when trying to further data gathering through this process. The final goal being automation of the process, a *PHP* based solution would prove unfit and unnecessarily cumbersome to make improvements on. Python presented itself as an alternative that would solve some if not all of these programs. As a scripting language, writing a script for searching Twitter for status updates on a regular basis was trivial, especially considering the existence of a Python wrapper capable of requesting the Twitter API for services (a wrapper being package that translates python code into HTML requests and back with the aim of working with the API using Python exclusively). One of the limitations of the free version of the Twitter Search API, as mentioned, is that searches can only be made against seven days worth of archived data, meaning the script had to be run at least once every seven days. Afterwards, the script was set up to automatically search Tweets based on a predefined query, grouping one hundred at a time - the limit of retrievable Tweets per search - and saving them into files with time stamps for names, in order to create unique ID's for unequivocal access should the need arise. Each Tweet is stored as JSON object instance and henceforth will be called a Tweet Object.

5.2.4 Catch-all Search Query

It was decided that the query chosen to search Tweets should be as broad as possible, since it was clear value was accrued through high volume rather than narrow specificity. Thus every time the search was carried out, it would be looking for every status update (i.e. Tweet) with the words corresponding to a large retailers brand name, namely Primark, be it in text or as a tag identifying the account. Being one of the biggest retailers in the world, Primark's Twitter account has approximately 234 thousand followers and preliminary searches showed hundreds of daily Tweets that where either exchanges between the retailer and users or users mentioning the retailer in a status update, by name or using a hashtag.

5.3 Feeding Watson NLU

For reasons mentioned in Subsection 4.1.2, page 18, IBM's Watson NLU was the tool of choice for analysing each Tweet collected through the process explained above. As with Twitter, Watson NLU services can be requested through an API with much the same mode of operation, and such as before, a wrapper was used to translate Python code into HTML requests and back. Constraints on number of uses and features per use, imposed by utilizing the lite version of the service (i.e. free of charge) meant a decision had to be made on what features where deemed useful for analysis. It was decided text would be processed for extraction of the following elements:

- Keywords - Returns important keywords in the content;
- Entities - Identifies people, cities, organizations, and other other entities in the content;
- Concepts - Returns high-level concepts in the content;
- Categories - Returns a five level taxonomy of the content;
- Emotion - Detects anger, disgust, fear, joy, or sadness that is conveyed in the content: Emotion can also be analyzed for detected entities with entities and for keywords;
- Sentiment - Analyzes the general sentiment of your content: Sentiment can also be analyzed for detected entities and for detected keywords.

An example of a full analysis can be found in Section B.2, Appendix B. The JSON object containing this analysis is called an NLU object, henceforth.

5.3.1 Automating analysis

As before, through the Python script interface, a Tweet's text, i.e. the actual user-written text, was sent to NLU through the API. The received JSON response was then concatenated with the JSON object representing the retrieved status and together they formed a single document containing all information regarding a single Tweet. By this point, the script automatically searches, extracts, processes and stores each Tweet and its analysis automatically and this process is run routinely.

5.4 Storing Analysis Documents in CouchDB and Insight Extraction

When it became clear that there where to many single documents to process locally through scripting code to run through folders, and rather a database was needed for indexing and retrieving documents was to be employed, the issue arose of how such varied types of data and could be stored in a relational database. This realization coming rather late in the project life-cycle, familiarization with the tools was limited, specially considering the tool chosen utilized a language that was not in the project developers curriculum. That being said, the potential of the tool was very clear and will be laid out as best as can be managed.

5.4.1 Feeding CouchDB

The CouchDB API also functions like the previous two, receiving http requests and outputting http strings that can be parsed into a JSON file. And just as before, a Python wrapper takes care of constructing HTTP requests based starting from Python, and translating responses back into Python. For each document destined to be stored an unique ID is attributed that was decided to be the original status id from the Tweet JSON object. A check is performed so no repetition of documents occurred. One by one documents are loaded into the database. Each document consists of the Tweet Object and the corresponding NLU object. Due to how CouchDB works, this change will be spread out to other nodes using the same database, and even if there's changes being made elsewhere on the dame data base, CouchDB's "Eventual Consistency" guarantees, sooner or later, every user on the system is looking at the same set of documents.

5.4.2 Data Handling on CouchDB

Most work involving data documents of CouchDB is done through views. Tasks involving views include but are not limited to: filtering documents relevant to a process, extracting data from a document and presenting in desired order, build indexes to find documents by value or structure and many calculations using data contained in the documents viewed. This is first achieved by designing map functions. When a view is queried, CouchDB runs the source code contained in the map unction on every document in the database the view was defined in. Based on your code, a document may or may not trigger a built in emit() function that takes two arguments, a "key" and a "value". CouchDB takes whatever you pass into the emit() function and puts it into a list, with each row containing a *key* and a *value*, with the list being sorted by *key*. *Row* happens to be the name to be given to each paired result. View results are stored in B-trees (Fig. 5.2, each B-tree being stores as their own file for high performance usage.

The next possible step is writing a reduce function. Because of the B-tree structure, similar keys can be reduced to one, with mapped values being processed to give a single result.

A very simple search function would look like this:

```
function(keys, values, rereduce) {
    return sum(values);
```

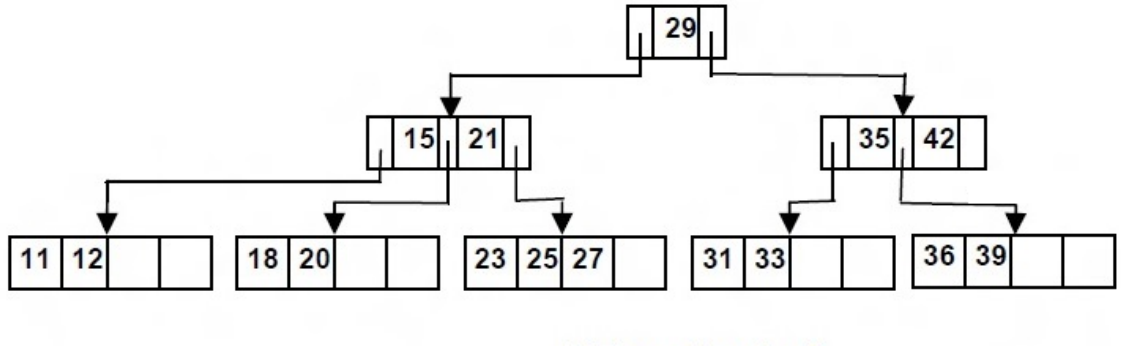

Figure 5.2: B-tree data structure

}

For a Map result (key, value) pair like:

(a , 1)
(a , 1)
(a , 1)
(b , 1)
(b , 3)

The resulting reduce would look like:

(a , 3)
(b , 4)

In CouchDB, map functions and reduce functions can be run sequentially in a view, therefore being named *MapReduce* functions. It's by applying map reduce functions to the documents created through the aforementioned, for each user status, that insight is extracted. These functions are written in JavaScript, since that's the native language of the CouchDB viewing server. Hence, all work from this point on was actually performed on Futon, a Web based graphic user interface for managing CouchDB. After building the appropriate views, the Python script can retrieve them through the API and visualisation tools may be applied for making sense of analysis.

5.5 Results

In this section, an example of a Tweet and its corresponding stored document, containing the Tweet object and the NLU object, is shown, as well as and example showing average sentiment regarding a chosen subset of identified categories. Choice sec This example serves the purpose of showing how the finished process can be made to target and study customers by themes of conversation directed to the brand. In figure 5.3, page 27, some of the attributes of the Tweet Object are displayed, highlighting the tweet text, quoted bellow, date of creation, location of the user and a description the user wrote about themselves. The tweet goes as follows:

"Almost 2019 and @Primark still haven't got an online store"

This text contains a clear and unequivocal opinion, expressed implicitly through the way the sentence is arranged. Date and location are relevant for pinning down the moment in time and space an opinion is being expressed, so a localized analysis for any given store can be carried out, within a given timeline. In figure 5.4, page 27 results of document-level sentiment analysis carried

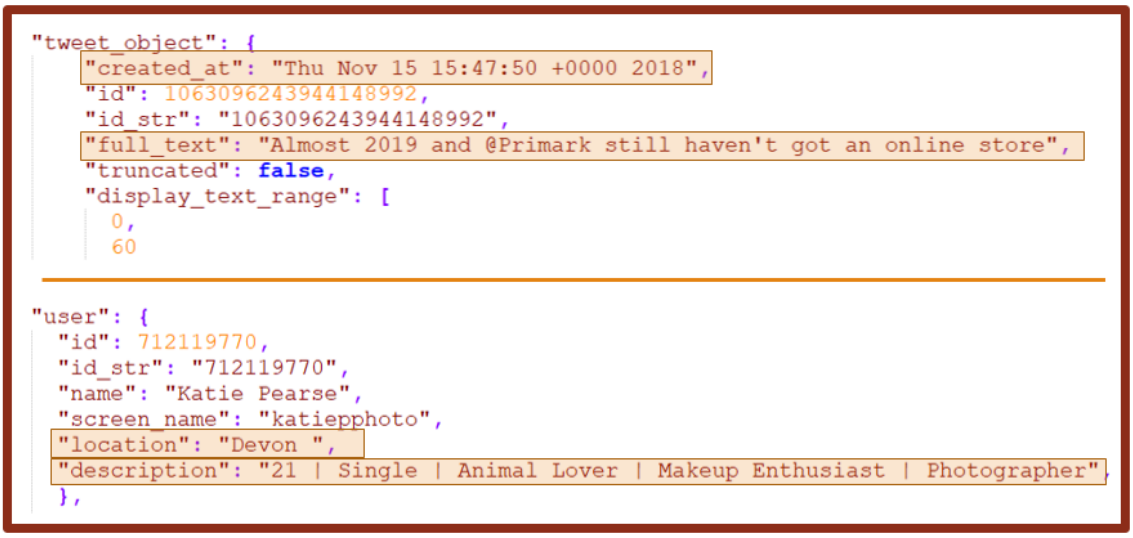

Figure 5.3: Example of a Tweet Object, highlighting text, date, location and user description

out on the given Tweet, using NLU is presented. As can be seen, sentiment is correctly identified as negative, despite the text expressing the opinion implicitly, without use of any negatively charged words, meaning, extracting analysis purely from context. In figure 5.5, page 28 results of keyword

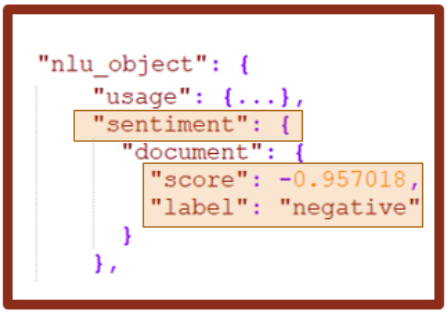

Figure 5.4: Results of document-level sentiment analysis

extraction, namely "Primark" and "online store", are presented, as well as an aspect-level sentiment analysis result and aspect-level emotion analysis, again correctly identifying negative sentiment and mainly sad emotion for each keyword and the context they appear in.

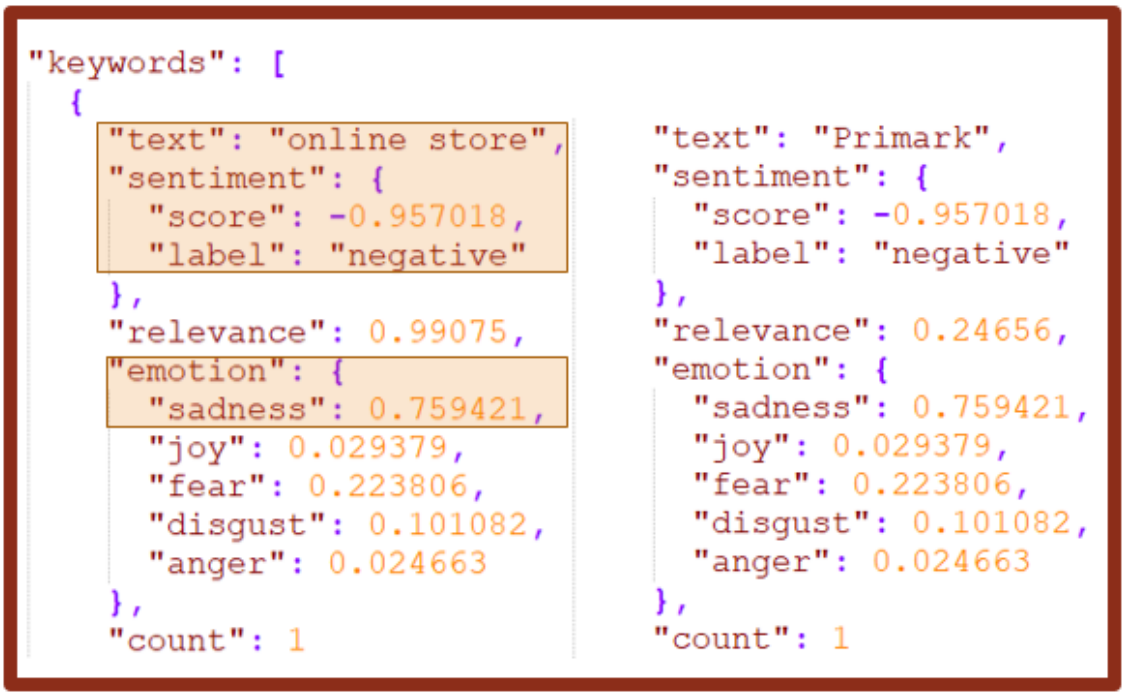

Figure 5.5: Results of keyword extraction and aspect-level sentiment analysis and emotion detection

Finally, in figure 5.6, page 28, the category of the tweet is extracted, identifying it taxonomically within 3 levels of depth. At the first level, the text is within the category of "shopping", at the second level, "retail" and lastly, "online stores".

Figure 5.6: Results of category classification

In the Table 5.1, page 29 and example of a CouchDB View is presented, indexing the average document-level sentiment regarding each category taxonomy extracted from text for the first level category *"/shopping"*.

5.5.1 Discussion of Results

The depth achievable in insight extraction using this combination of tools is incredible. And lamentably, lack of time didn't allow for deeper exploration of flashy results. However, some

Table 5.1: Average Sentiment by Category

Category	Sentiment [-1,1]
/shopping/toys/action figures	0.801
/shopping/toys/dolls	0.23212975
/shopping/retail/outlet stores	-0.014775613041984144
/shopping/retail/online stores	-0.46258987500000004
/shopping/retail/department stores	0.2133975182342175

thought was put into how these insights can be arrived at. One obvious step is isolating every entity discovered by the NLU and analysing sentiment and emotion regarding those that appear more frequently. Summarizing sentiment scores reveals how user or customers lean regarding that entity. Drilling down on this notion, users can be grouped by demographics of one owns choosing based on parameters and metrics retrieved by the Twitter Search API regarding each status's original poster and research sentiment and emotion regarding each entity for a determined demographic. Comparing differences in opinion between demographics concerning the same entity may be cause for further exploration so one arrives at the root-cause of such phenomena. Also, getting back to the opinion quintuplet presented in Section 2.2.1, page 9, explorations of sentiment in time can be made, including drill down criteria for demographics, and thus observing change of opinion over time, be it opinion of the brand overall, opinion on a certain entity, by a certain group of people, etc. The case for exploring social media data in search of patterns and unknown information that might be of use is self evident and this tool may well prove useful in that search.

Chapter 6

Conclusion and future Improvement

It concludes that, as stated before, Social Media is a vast pool of data waiting to be harnessed and put to use. That being stated, the Retail service area faces challenges that can be severely mitigated through the use of ever more sophisticated and complete systems of transforming raw data into useful information and actionable insight. Having identified the lack of a tool to harness such information from such a rich source, this project attempted to do just that.

The laid out data discovery and extraction methods, using the Twitter Search API proved a success, even wile utilizing the standard version of the service. The process of collecting statuses was seamless and the JSON format allowed for easy handling through the rest of the process.

The IBM Watson NLU API performed as expect of self learning algorithms. Some content was mislabeled or misunderstood, but overall the quality of the analysis was deemed good enough to be useful. Features regarding creating user-made models of analysis where available, not used, but interesting to explore in further iterations.

Data gathered in CouchDB was safe from corruption and accessible from any terminal and the platform behaved beautifully processing hundreds of thousands of documents for indexing and as such proved to be the right tool for the matter.

More effort could have and should have been exerted in actually creating the MapReduce functions to tell a convincing story of its value, but the possibilities laid out are feasible and of intrinsic value for a number of applications.

6.1 Future Improvements

As far as future improvements, two seem the most urgent. The first is assessing the influence of the breadth of search when querying social media. The second is the creation of well thought out MapReduce functions that can integrate retail processes efficiently, presenting data need when it's needed, thereby being actual actionable insight.

Another improvement is the inclusion of other platforms, starting with Facebook. Enough data on the platform is public, and difficulty of access does not mean impossibility. It would be

interesting how discourse changes between platforms and how through data gathered from each, the same entity or aspect is evaluated. If such differences exist, the question pends: Why?

The architecture of the project allows for distribution as a web service. When proven value can be extracted from the tools devised, deploying the project as a service could prove beneficial for users and profitable for owners of the service.

Appendix A

Twitter Status Object JSON


```
{
"created_at": "Thu Nov 15 14:54:32 +0000 2018",
"id": 1063082831000358912,
"id_str": "1063082831000358912",
"full_text": "@lol_itskatie Hi, sorry to hear this. If the item is faulty yo
"truncated": false,
"display_text_range": [
14,
119
],
"entities": {
"hashtags": [],
"symbols": [],
"user_mentions": [
 {
"screen_name": "lol_itskatie",
"name": "Resident Fatass\u270c",
"id": 1436311554,
"id_str": "1436311554",
"indices": [
0,
13
]
 }
],
"urls": []
},
"metadata": {
```

```
"iso_language_code": "en",
"result_type": "recent"
},
"source": "<a href=\"http://www.zendesk.com\" rel=\"nofollow\">Zendesk</a>",
"in_reply_to_status_id": 1063056936998965248,
"in_reply_to_status_id_str": "1063056936998965248",
"in_reply_to_user_id": 1436311554,
"in_reply_to_user_id_str": "1436311554",
"in_reply_to_screen_name": "lol_itskatie",
"user": {
"id": 1630182978,
"id_str": "1630182978",
"name": "Primark",
"screen_name": "Primark",
"location": "",
"description": "\ud83c\udf08 Wear.Share.Inspire. Customer Service enquiries: htt
"url": "https://t.co/vMwt3SW1fd",
"entities": {
 "url": {
"urls": [
{
 "url": "https://t.co/vMwt3SW1fd",
 "expanded_url": "http://www.primark.com",
 "display_url": "primark.com",
 "indices": [
0,
23
 ]
}
]
 },
 "description": {
"urls": [
{
 "url": "https://t.co/uaeJa6yQfO",
 "expanded_url": "https://www.help.primark.com/",
 "display_url": "help.primark.com",
 "indices": [
50,
73
```

```
 ]
}
]
 }
},
"protected": false,
"followers_count": 232287,
"friends_count": 3185,
"listed_count": 616,
"created_at": "Mon Jul 29 12:04:15 +0000 2013",
"favourites_count": 27748,
"utc_offset": null,
"time_zone": null,
"geo_enabled": true,
"verified": true,
"statuses_count": 67922,
"lang": "en-gb",
"contributors_enabled": false,
"is_translator": false,
"is_translation_enabled": true,
"profile_background_color": "FECC3F",
"profile_background_image_url": "http://abs.twimg.com/images/themes/theme12/
"profile_background_image_url_https": "https://abs.twimg.com/images/themes/t
"profile_background_tile": false,
"profile_image_url": "http://pbs.twimg.com/profile_images/998127801600430082
"profile_image_url_https": "https://pbs.twimg.com/profile_images/9981278016O
"profile_banner_url": "https://pbs.twimg.com/profile_banners/1630182978/1541
"profile_link_color": "00AFDB",
"profile_sidebar_border_color": "000000",
"profile_sidebar_fill_color": "DDEEF6",
"profile_text_color": "333333",
"profile_use_background_image": true,
"has_extended_profile": false,
"default_profile": false,
"default_profile_image": false,
"following": false,
"follow_request_sent": false,
"notifications": false,
"translator_type": "none"
},
```

```
"geo": null,
"coordinates": null,
"place": null,
"contributors": null,
"is_quote_status": false,
"retweet_count": 0,
"favorite_count": 0,
"favorited": false,
"retweeted": false,
"lang": "en"
 }
```

Appendix B

NLU Object JSON

B.1 Example of an NLU Object


```
{
"usage": {
"text_units": 1,
"text_characters": 209,
"features": 2
},
"sentiment": {
"document": {
"score": 0.636049,
"label": "positive"
}
},
"language": "en",
"emotion": {
"document": {
"emotion": {
"sadness": 0.15373,
"joy": 0.611206,
"fear": 0.03094,
"disgust": 0.054218,
"anger": 0.170435
}
}
},
"analysed_text": "This lovely, Primark Checkered Shoes is up for grabs today
 \n\nWant to make a bid??\n\nTick Tock. Tick Tock
```

}

B.2 Full NLU analysis


```
"nlu_object": {
"usage": {
"text_units": 1,
"text_characters": 266,
"features": 6
},
"sentiment": {
"document": {
"score": 0.957615,
"label": "positive"
}
},
"language": "en",
"keywords": [
{
"text": "part of this exciting project",
"sentiment": {
"score": 0.974402,
"label": "positive"
},
"relevance": 0.725813,
"emotion": {
"sadness": 0.039368,
"joy": 0.92289,
"fear": 0.024058,
"disgust": 0.013492,
"anger": 0.002848
},
"count": 1
},
{
"text": "\ud83e\udd17 https://t.co/kklVn15u9a",
"sentiment": {
"score": 0.907082,
"label": "positive"
```

```
},
"relevance": 0.630667,
"emotion": {
"sadness": 0.365452,
"joy": 0.411231,
"fear": 0.189583,
"disgust": 0.073751,
"anger": 0.020486
},
"count": 1
},
{
"text": "kid",
"sentiment": {
"score": 0.974402,
"label": "positive"
},
"relevance": 0.580907,
"emotion": {
"sadness": 0.039368,
"joy": 0.92289,
"fear": 0.024058,
"disgust": 0.013492,
"anger": 0.002848
},
"count": 1
},
{
"text": "ambassador",
"sentiment": {
"score": 0.98342,
"label": "positive"
},
"relevance": 0.537694,
"emotion": {
"sadness": 0.036009,
"joy": 0.929479,
"fear": 0.006834,
"disgust": 0.009396,
"anger": 0.010697
```

```
},
"count": 1
},
{
"text": "mum",
"sentiment": {
"score": 0.974402,
"label": "positive"
},
"relevance": 0.434106,
"emotion": {
"sadness": 0.039368,
"joy": 0.92289,
"fear": 0.024058,
"disgust": 0.013492,
"anger": 0.002848
},
"count": 1
},
{
"text": "Primark",
"sentiment": {
"score": 0.98342,
"label": "positive"
},
"relevance": 0.397244,
"emotion": {
"sadness": 0.036009,
"joy": 0.929479,
"fear": 0.006834,
"disgust": 0.009396,
"anger": 0.010697
},
"count": 1
}
],
"entities": [
{
"type": "TwitterHandle",
"text": "@Primark",
```

```
"sentiment": {
"score": 0.0,
"label": "neutral"
},
"relevance": 0.01,
"count": 1
}
],
"emotion": {
"document": {
"emotion": {
"sadness": 0.049787,
"joy": 0.930548,
"fear": 0.014246,
"disgust": 0.007398,
"anger": 0.001537
}
}
},
"concepts": [
{
"text": "Rhythm and blues ballads",
"relevance": 0.884595,
"dbpedia_resource": "http://dbpedia.org/resource/Rhythm_and_blues_ballads"
}
],
"categories": [
{
"score": 0.963118,
"label": "/shopping"
},
{
"score": 0.88551,
"label": "/shopping/resources"
}
],
"analyzed_text": "I\u2019m SO happy to finally announce that I am @Primark f
}
```

References

[1] Data Never Sleeps 6 | Domo. URL: https://www.domo.com/learn/data-never-sleeps-6.

[2] Medhat Walaa, Hassan Ahmed, and Korashy Hoda. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 2013. URL: https://ac.els-cdn.com/S2090447914000550/1-s2.0-S2090447914000550-main.pdf?_tid=f7297a09-84b2-4bbe-b994-470d65312989&acdnat=1542631531_4ab4ea6ff4adea276852bb821b92b3b7.

[3] COMMISSIONED BY RELEX SOLUTIONS State of the Retail Supply Chain. Technical report, 2017. URL: https://cdn2.hubspot.net/hubfs/317879/Martec%20Report%202017/State_of_the_retail_supply_chain_2017_web.pdf?__hssc=197871792.1.1512560418832&__hstc=197871792.cefdd9b43943d87feb2e68379ccd3e12.1512552769149.1512552769149.1512560418832.2&__hsfp=3867347377&hsCtaTracking=cae6c95c-3721-47df-a5b0-a34c4a13901b%7C2a5ce5e5-c23f-4054-9cad-78e3a14afc9b&t=1513857192611.

[4] Oracle. Offer Optimization Brief. URL: http://www.oracle.com/us/industries/retail/retail-offer-optimization-brief-4861980.pdf.

[5] Pritam Gundecha and Liu Huan. Mining Social Media: A Brief Introduction. 2014. URL: http://pubsonline.informs.orghttp://www.informs.org, doi:10.1287/educ.1120.0105.

[6] Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144, 4 2015. URL: https://www.sciencedirect.com/science/article/pii/S0268401214001066, doi:10.1016/J.IJINFOMGT.2014.10.007.

[7] Gerard George, Martine R Haas, and Alex Pentland. Big Data and Management: From the Editors. *Academy of Management Journal*, 57(2):321–326, 2014. URL: http://ink.library.smu.edu.sg/lkcsb_researchhttp://ink.library.smu.edu.sg/lkcsb_research/4621, doi:10.5465/amj.2014.4002.

[8] Klaus Lyko, Marcus Nitzschke, and Axel-cyrille Ngonga Ngomo. Big Data Acquisition. pages 39–61, 2016. doi:10.1007/978-3-319-21569-3.

[9] New Tools for Managing Communication on Your Page | Facebook for Business. URL: https://www.facebook.com/business/news/new-tools-for-managing-communication-on-your-page?__mref=message_bubble.

[10] Zeynep Tufekci. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. Technical report, 2014. URL: www.aaai.org.

[11] investor.fb.com. Facebook Q3 2018 results. Technical report. URL: https://s21.q4cdn.com/399680738/files/doc_financials/2018/Q3/Q3-2018-Earnings-Presentation.pdf.

[12] Investor Fact Sheet MAU (Monthly Active Users) DAU (Daily Active Users) Year-Over-Year Growth. Technical report. URL: https://s22.q4cdn.com/826641620/files/doc_financials/2018/q3/TWTR-Q3_18_InvestorFactSheet.pdf.

[13] Cracking Down on Platform Abuse | Facebook Newsroom. URL: https://newsroom.fb.com/news/2018/03/cracking-down-on-platform-abuse/.

[14] The Web Robots Pages. URL: http://www.robotstxt.org/robotstxt.html.

[15] Stefan Stieglitz, Milad Mirbabaie, Björn Ross, and Christoph Neuberger. Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, 39:156–168, 4 2018. URL: https://www.sciencedirect.com/science/article/pii/S0268401217308526, doi:10.1016/j.ijinfomgt.2017.12.002.

[16] Stefan Stieglitz, Linh Dang-Xuan, Axel Bruns, and Christoph Neuberger. Social Media Analytics. *WIRTSCHAFTSINFORMATIK*, 56(2):101–109, 4 2014. URL: http://link.springer.com/10.1007/s11576-014-0407-5, doi:10.1007/s11576-014-0407-5.

[17] Bernard J Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter Power: Tweets as Electronic Word of Mouth. 2009. URL: www.interscience.wiley.com, doi:10.1002/asi.21149.

[18] Bing Liu. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012. URL: https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdfinternal-pdf://0744994148/SentimentAnalysisandOpinionMining.pdfhttp://www.morganclaypool.com/doi/abs/10.2200/S00416ED1V01Y201204HLT016, doi:10.2200/S00416ED1V01Y201204HLT016.

[19] Tomáš Ptáček, Ivan Habernal, and Jun Hong. Sarcasm Detection on Czech and English Twitter. Technical report. URL: http://www.mturk.com.

[20] Adam Bermingham and Alan Smeaton. *Classifying Sentiment in Microblogs: Is Brevity an Advantage?* 2010. URL: http://www.computing.dcu.ie/~abermingham/data/.

[21] Oracle. Oracle Retail Advanced Inventory Planning. URL: http://www.oracle.com/us/products/applications/047071.pdf.

[22] Oracle. Oracle Retail Allocation. URL: http://www.oracle.com/us/products/applications/062048.pdf.

[23] Oracle. Oracle Retail Customer Engagement Cloud Services. URL: http://www.oracle.com/us/industries/retail/customer-engage-cloud-service-ds-2503875.pdf.

[24] Oracle. Oracle Retail Category Management Planning & Optimization. URL: http://www.oracle.com/us/industries/retail/retail-category-management-ds-2616834.pdf.

[25] The Power of Predictive Analytics and Social Media Data - Christopher S. Penn Marketing Blog. URL: https://www.christopherspenn.com/2017/08/the-power-of-predictive-analytics-and-social-media-data/.