LISBON
SCHOOL OF
ECONOMICS &
MANAGEMENT
UNIVERSIDADE DE LISBOA

MASTER

INFORMATION SYSTEMS MANAGEMENT

2019/2020

MASTER'S FINAL PROJECT

PROJECT

ASSESSING PUBLIC FIGURES' REPUTATION THROUGH SENTIMENT ANALYSIS ON TWITTER USING MACHINE LEARNING: CREATION OF A SYSTEM

CATARINA CORREIA VIEGAS

NOVEMBER - 2020

LISBON
SCHOOL OF
ECONOMICS &
MANAGEMENT
UNIVERSIDADE DE LISBOA

MASTER

INFORMATION SYSTEMS MANAGEMENT

2019/2020

MASTER'S FINAL PROJECT

PROJECT

ASSESSING PUBLIC FIGURES' REPUTATION THROUGH SENTIMENT ANALYSIS ON TWITTER USING MACHINE LEARNING: CREATION OF A SYSTEM

CATARINA CORREIA VIEGAS

ORIENTATION:

PROFESSOR DOUTOR JESUALDO CERQUEIRA FERNANDES

NOVEMBER – 2020

Acknowledgments

Throughout the completion of this project I have received an immense deal of support from different people. Therefore, I would like to acknowledge my deepest appreciation…

…to my parents, who made me who I am. I will never be able to thank them enough.

…to my sister, who is the better half of me. She has always been by my side and I would not have managed to finish this chapter of my life, nor any other one, without her.

…to my grandparents, who never doubted me and have shown nothing but love and support to whatever choices I decide to make.

…to my aunt and uncle, as well as my cousins, that have always been there for me.

…to my friends, my soul sisters that mean so much to me and give me unconditional support on everything.

…to the colleagues that became friends. This process would not have been the same without them, and I will never forget the encouragement they gave me. A very special thank you to two of them, not only because they revised all my materials and offered valuable insights, but mostly because they cheered and uplifted me throughout the entire project completion. I am eternally glad to have been able to meet them and share this journey with them.

…and because they say that a dog is a woman's best friend, and I could not possibly agree more, to my dog, who was always by my side while I worked on this project, as well as on everything I did for the past 9 years.

Finally, I would also like to express my utmost gratitude for my supervisor, who believed in my project, even when I did not. He welcomed me with enthusiasm when I reached out and accepted this challenge without hesitation. I am deeply thankful for his guidance and support.

List of Acronyms

API – Application Programming Interface

BD – Big Data

CRISP-DM – Cross-Industry Standard Process for Data Mining

IDE - Integrated Development Environment

JSON – JavaScript Object Notation

KDD – Knowledge Discovery in Databases

ME – Maximum Entropy

NB – Naïve Bayes

NER – Named Entity Recognition

NLP – Natural Language Processing

POS – Part-of-speech

SA – Sentiment Analysis

SEMMA – Sample, Explore, Modify, Model, Assess

SVM – Support Vector Machine

TF-IDF – Term Frequency-Inverse Document Frequency

URL – Uniform Resource Locator

Abstract

Never has so much data been generated and at such an astounding rate as nowadays. This is undoubtedly an era of Big Data and this term does not go unnoticed, bearing within innumerous challenges, but also a multitude of opportunities. Of the generated data, roughly 80% comes unstructured, which makes its analysis more challenging. Within this type of data there is a special focus on the text format, a format that is frequent and carries great potential. There are several applications, techniques and tools connected to the analysis of textual documents and this area is strongly linked to Natural Language Processing, another topic of extreme importance in this field. One of the greatest challenges of both is related to Sentiment Analysis, an area that has attracted both academics and professionals given its many applications. This analysis also has a particular impact on social networks. From the multiplicity of topics that could be studied, it was interesting to combine trends and address issues such as online reputation and social image. Thus, this project focused on creating a system capable of identifying the sentiment associated with public figures, demonstrated through publications on Twitter. For this purpose, the first step was to carry out a literature review capable of exploring the topics and recent developments associated with the chosen subject. Regarding the system, a Machine Learning approach using supervised learning methods was adopted. To this end, a manually annotated dataset that intended to be as inclusive as possible was created. Afterwards, and succeeding the text transformation, three of the most used classifiers (Naïve Bayes, Support Vector Machines and Maximum Entropy) were trained in an attempt to gauge which one would demonstrate better results. After an initial training phase, the individual impact of some pre-processing techniques was assessed. The obtained results were not as good as initially desired, nonetheless the best model was chosen to be incorporated into the system. This project contributes to increase the knowledge base of the areas in which it is comprised, and also provides a manually annotated dataset that can be used in further research.

Keywords: Big Data, Machine Learning, Online Reputation, Natural Language Processing, Sentiment Analysis, Twitter

Resumo

Nunca se geraram tantos dados e a um ritmo tão alucinante como atualmente. Vive-se indubitavelmente numa era de *Big Data* e este termo não passa despercebido, trazendo consigo inúmeros desafios, mas também múltiplas oportunidades. Dos dados gerados, cerca de 80% encontra-se de forma desestruturada, o que torna a sua análise um pouco mais desafiante. Dentro deste tipo de dados há um foco especial para o formato de texto, formato esse que para além de comum, agrega um grande potencial. Existem várias aplicações, técnicas e ferramentas associadas à análise de documentos textuais. Esta área surge fortemente ligada ao Processamento de Linguagem Natural, um tópico de extrema importância neste domínio. Um dos grandes desafios de ambos encontra-se relacionado com Análise de Sentimentos, uma área que tem atraído tanto académicos, como profissionais, dada as suas inúmeras aplicações. Esta análise tem ainda uma particular incidência no âmbito das redes sociais. Da multiplicidade de tópicos que poderiam ser estudados, é interessante aliar tendências e abordar a questão da reputação online e a imagem social. Dessa forma, o presente projeto focou-se na criação de um sistema capaz de identificar o sentimento associado a figuras públicas demonstrado através de publicações no Twitter. Com essa finalidade, o primeiro passo consistiu em levar a cabo uma revisão de literatura capaz de explicitar os tópicos e tendências associadas à temática escolhida. Relativamente ao sistema, optou-se por uma abordagem de *Machine Learning* com recurso a métodos supervisionados de aprendizagem. Para tal, criou-se um *dataset* manualmente anotado, que tentou ser o mais inclusivo possível, e procedeu-se ao treino de três classificadores (*Naïve Bayes*, *Support Vector Machines* e Entropia Máxima) numa tentativa de aferir qual demonstraria os melhores resultados. Após uma fase inicial de treinos, investigou-se ainda o impacto individual que alguns procedimentos e técnicas teriam na *perfomance* dos classificadores escolhidos. Os resultados obtidos não foram tão bons como inicialmente esperado, mas, no final, escolheu-se o melhor modelo para ser incorporado no sistema. Este projeto contribuiu para aumentar a base de conhecimento das áreas em que se insere, e fornece ainda um *dataset* manualmente anotado que poderá ser utilizado em investigações futuras.

Palavras-chave: *Big Data*, *Machine Learning*, Processamento Natural de Linguagem, Análise de Sentimentos, *Twitter*, Reputação Online

Table of contents

Table of contents – Figures

Table of contents – Tables

Table of contents – Equations

1. Introduction

1.1. Contextualization

Data has never been created at this rate. In fact, statistics suggest that most of the world's data has been created in the last two years alone. About 1.7MB is created per second by every person and 2.5 quintillion bytes of data a day are produced by humans (Bulao, 2020). Data is generated in multiple formats at an astonishing rate and it comes from a multitude of sources, thus one can say that this is undoubtedly a Big Data (BD) era (Sivarajah et al., 2017). Although its genesis is uncertain and several definitions have arisen, almost everyone is familiar with the term BD nowadays. This thematic has been subject of much attention from various entities, from researchers to corporate leaders, but has also raised some fear, as its gigantic potential comes alongside with numerous challenges (Gandomi & Haider, 2015). Decisions based on data instead of intuition are deemed as better decisions (McAfee & Brynjolfsson, 2012), therefore, unlocking BD's potential in order to support the decision-making process should be prerogative (Gandomi & Haider, 2015). Creativity to achieve that and deal with the forthcoming challenges is crucial (John Walker, 2014).

The data generated can be deemed as structured, unstructured or semi-structured, the latter being the category that falls in between the others (Gandomi & Haider, 2015). Whilst structured data is the easiest to manipulate, only 5% fall into that category (Cukier, 2010) apud (Gandomi & Haider 2015, p.138). On the other hand, unstructured data, which is more difficult to work with and often overlooked by companies due to the hardships it presents (Rogers, 2019), represent 80% of the data generated (Grimes, 2008) and is created at a faster rate (Lee, 2017). Most of it appear in text format (Grimes 2008) and analysing this type of data has the potential to bring more business-based useful information (Tan, 1999; Chen, Mao & Liu, 2014). Text Analytics, or Text Mining, the *process of extracting interesting and non-trivial patterns or knowledge from text documents*" (Tan, 1990; p.65) has emerged as a field to deal with this data format. There are several techniques and tools within this dimension (Gandomi & Haider, 2015; Fan et al., 2006), however, while machines are prepared to process structured databases automatically, text was meant to be read by people (Hearst, 2003). Natural Language Processing (NLP) has then appeared to produce technologies able to teach natural language to computers and allow them to comprehend, analyze and even produce it (Fan et al., 2006).

Within the Text Mining realm, one of the most challenging NLP research topics is Sentiment Analysis (SA) (Liu, 2012). SA, also referred to as Opinion Mining, is the "*computational study of people's opinions, attitudes and emotions toward an entity*" (Medhat, Hassan & Korashy, 2014; p.1903), an entity being a topic, an event, an individual, a brand/organization, a product and so forth (Medhat, Hassan & Korashy, 2014). Due to its high potential, SA is being broadly studied and applied to almost every domain (Liu, 2012). To perform it, there are two main group of approaches, one being lexicon-based and the other relying on machine learning methods (Medhat, Hassan & Korashy, 2014). One of the richest sources of opinionated data sources are the social media platforms (Liu, 2012), and Twitter has attracted the SA community, who has taken a special interest in studying the platform (Martínez-Cámara et al., 2012).

Nowadays, it is undeniable that social media has a huge impact on both organizations and people. Anyone can share their opinions, whether it is their outtake on an experience with a product or service, or simply their regards towards an organization or person. Aula (2010) states that social media has the capability of presenting a collective truth. The interactive aspect that is so defining of social networks presents several implications to organizations and they must adapt and follow proper strategies. If they do not, their reputation can be put at risk (Aula, 2010). Reputation, which has become a strategic asset, when lost can have astounding costs (Floreddu, Cabiddu & Evaristo, 2014). Moreover, this is an important topic not only for corporations, but also for humans, as reputation plays a major role in people's life and is a central aspect of social identity (Jazaeiri et al., 2018). One can say that a person is their own brand, and this is clearer in the case of public figures, whose reputations can define their success or downfall.

Taking this into account, the present project intends to explore the Big Data - Sentiment Analysis trend, combining it with another important and emerging topic such as reputation and how it can be defined by public opinion. For that, a system capable of assessing public figures' reputation through Twitter will be created. The first step in order to do that is to conduct a literature review, where all concepts will be found logically chained, starting with a broad view on general topics and narrowing it down to the specifics of the problem on which the project relies. After the review, the focus of the present document will be the methodology regarding the elaboration of the system, followed by a discussion and final remarks.

1.2. Relevance and Main Goals of the Project

Unstructured data in text format is the most common type of generated data (Grimes, 2008) and social media is a rich source of it, therefore, it would be interesting to showcase an approach at one of BD's problems. Reputation and image are also major issues in our society; hence it makes sense for this project to specify on the people's domain. Nowadays, it is impossible to manually gather, label and classify all the data that is generated (Chikersal et al., 2015), thus having a system to tackle this could be a great advantage for companies/Public Relationships management. In summary, this project aims to create a system capable of understanding the overall sentiment towards public figures to assess their reputation. Twitter, which has been widely studied, will be used as the source of data. Due to its potential to generate good results, a supervised machine learning approach will be used to deal with the data. In order to achieve this goal, it is necessary to:

- Create a dataset for the specific chosen domain (people).
- Train and test different classifiers to identify which produces better results.

At the end, with the best performing model, there will be a system able to analyse new data and predict the overall sentiment regarding a specific public figure, thus assessing his reputation.

2. Literature Review

In order to start a project, the first step is to understand the state-of-the-art of the concepts it relies on. This is of utmost importance, and therefore the present document provides a Literature Review that aims to briefly introduce the key ideas and topics. It begins with the broader topics, such as BD and succinctly explains the concept, its evolution and the techniques and tools available to unveil its potential. Then it narrows down to the specifics of textual data analysis, exposing not only its benefits, but also the challenges it brings. A chapter about the field of NLP can also be found, followed by the crucial mention of SA. That includes the explanation of the topic and a generalized view of the approaches and techniques commonly used. Its connection and applications with social media are also explored. Afterwards, and since the project focuses on that particular domain, it is finalized with a short chapter about Public Image and Reputation and the role they play both on people's lives and on corporations' performance.

2.1. Big Data

2.1.1. Concept and Evolution

Almost everyone has heard the term BD, but, although widely known, its genesis is uncertain and it has been awarded several explanations, which ultimately generated confusion (Gandomi & Haider, 2015). The concept itself can be deemed as abstract, and even nowadays there is no consensus on its true definition (Chen, Mao & Liu, 2014). In their paper, Gandomi and Haider (2015) quote some of the most used interpretations, including Tech's American Foundation (2012) apud (Gandomi & Haider 2015, p.138) that states that BD is a term used to describe large volumes of complex and variable data that is generated at high speed and require advanced techniques and technologies to allow its collection, storage, distribution, management and analysis. One could also say that "*BD is the artefact of human individual as well as collective intelligence generated and shared mainly through the technological environment*" (Sivarajah et al.,2017, p.264). Wamba et al (2015) tried to compile and display some published definitions in their work.

When it comes to BD, the sources and features inherent to its concept have not always been the same. Lee (2017) highlights three main stages: **Big Data 1.0 (1994-2004)** – the beginning of this phase goes back to the outset of e-commerce. Businesses focused on establishing online presence, being able to maintain transactions with clients and improving their efficiency regarding operations (Provost & Fawcet, 2013; Lee, 2017). Therefore, the companies that operated online were precisely those which contributed mostly to web content. User generated content was not yet significant due to the technical limitations of the applications (Lee, 2017); **Big Data 2.0 (2005-2014)** – As the name suggests, it is related to Web 2.0. Technologies and applications became more advanced, and social media rose and started being prominently used. End-users were now able to engage and interact with websites, as well as contribute with their own content (Kaplan & Haenlein, 2010; Lee, 2017). Social media was a major driver to the growth of user generated content, and this also led businesses to improve their techniques to extract data that could potentially be useful for them (Bjurstrom & Plachkinova, 2015); **Big Data 3.0 (2015 –)** – adding to the interactions previously mentioned in the previous phases, it also includes Internet of Things devices and applications, which can generate data in text, audio and/or video format without human intervention (Lee, 2017). Although being quite recent and not so expressive nowadays, it is believed that this source of data and the

amount of information it produces will be substantial by 2030 (Chen, Mao & Liu, 2014; Yaqoob et al., 2016).

2.1.2. Characteristics

BD has certain inherent attributes that are widely agreed upon and commonly denominated by Vs. Initially, Laney (2001) proposed **Volume**, **Velocity** and **Variety** and those have become the most acknowledged and used to describe BD. Volume is related to the amount of data generated and/or collected, either by an individual or an organization (Lee, 2017). Currently, data magnitudes are appearing as numerous exabytes and petabytes (Gandomi & Haider, 2015), however, as everything evolves, what is considered BD today will probably not be the same in the future (Gandomi & Haider, 2015; Lee, 2017). Recently, Reinsel, Gantz and Rydning (2018), in a sponsored IDC white paper, stated that IDC has predicted that the global datasphere will go from 33 zettabytes[1] in 2018 to 175 zettabytes by 2025. When it comes to Velocity, it is referring to the speed at which data is generated and how fast it can be handled (Gandomi & Haider, 2015; Lee, 2017). This is decisive for companies, as being able to obtain and work with real time information may enable them to be sharper than their contenders (McAfee & Brynjolfsson, 2012). Desjardins (2019), posting for the World Economic Forum, reckons that by 2025, around 463 exabytes of data will be produced daily. Lastly, Variety points to the diversity in data types and structures generated. Data can be considered structured, when found in relational databases or spreadsheets (e.g., Excel spreadsheet), unstructured, when machines normally are not prepared to deal with them at a first instance and without prior intervention/pre-processing (e.g., text and audio), or semi-structured, this category falling in between the previously mentioned (e.g., Extensible Markup Language) (Gandomi & Haider, 2015; Lee, 2017). Whereas structured data is the easiest to manipulate, only 5% of all data fits that category (Cukier, 2010) apud (Gandomi & Haider 2015, p.138). As of unstructured data, it is generated at a faster rate when compared with the other types (Lee, 2017) and expanding at around 55-65% early (Marr, 2019). Although more difficult to deal with, and thus often overlooked by companies because of the hardships it presents, developing strategies to use the information it hides should be prerogative (Rogers, 2019). New technologies and techniques are emerging

[1] For reference, a zettabyte is equivalent to a trillion gigabytes.

and becoming available, so dealing with unstructured data will gradually require less effort (Lee, 2017).

Besides the previous well-known Vs, there have been other suggestions for dimensions that could be intrinsic to the BD concept. For instance, SAS advocated for both **Variability** and **Complexity** (note that some this last term is also sometimes referred to Veracity), while IBM proposed **Veracity** (different than SAS' definition) and Oracle **Value** (Gandomi & Haider, 2015; Lee, 2017). Variability appears related to the erratic, and somehow difficult, way to predict data flow rates at which data is generated, whilst Complexity comes linked to the heterogeneity of data sources that poses as a challenge when trying to connect, match, clean data across multiple systems (SAS, nd; Gandomi & Haider, 2015; Lee, 2017). George, Haas & Pentland (2014) enumerate 5 pivotal sources of high-volume data: **(1) public data**, often held by government institutions, **(2) private data**, that is associated with private firms and other organizations, **(3) data exhaust**, referring to passively created data that when combined can reveal important information, **(4) community data**, which is predominantly non-structured data in text format generated in social networks, and **(5) self-quantification data**, that point out to data created by devices that monitors actions and behaviours. IBM's suggestion to include Veracity refers to the uncertainty of data. (IBM, nd; Lee, 2017). Data can be incomplete, inaccurate, inconsistent, defective, subjective and so on, and that is a challenge when deciding what information to trust and use (Lee, 2017). At last, it is possible to state that data has an undiscovered intrinsic value (Oracle, nd). When it comes to Value, organizations must fathom the importance BD can represent to their decisions (Lee, 2017). Although originally considered low-level (the value/volume ratio is small) (Gandomi & Haider, 2015; Lee, 2017), data value can be transformed to high-level when the right tools and techniques are applied (Lee, 2017). Higher value can also be attained by analysing greater data volumes (Gandomi & Haider, 2015). Although the previous concepts are the most established and acknowledged on literature after the 3 initial Vs Laney (2001) proposed, on Sivarajah et al (2016) there is evidence that **Visualization** could also represent a major topic connected to BD, as it is a challenge to represent efficiently the information gathered so it can be consulted easily and effectively.

2.1.3. Value Extraction

"*Big Data used to be a technical problem, now it's a business opportunity*" (Russom, 2011, p.3). Its potential has been recognized and the number of publications

around this theme have risen. Multiple domains and sectors have been exploring the opportunities within this realm and benefited from developments around the subject. (Rodríguez-Mazahua et al., 2015; Sivarajah et al., 2017). For instance, Sagiroglu and Sinanc (2013) point to a Mckinsey Report that identifies opportunities in healthcare, retail, public sector, manufacturing and personal location data.

Decisions based on data instead of on intuition or hunches are considered better decisions (McAfee & Brynjolfsson, 2012). BD can be thought of as a gold mine, but it holds no significant value if its potential is not properly addressed and used towards better decisions (Gandomi & Haider, 2015). In order to do that there are two main processes: **Data Management** and **Data Analytics**, the first comprising the phases of acquiring and preparing the data for the analysis, and the latter being directly related to the process of acquiring insightful information (Labrinidis & Jagadish, 2012) apud (Gandomi & Haider 2015, p.140). The subject of BD Analysis is still developing, but, in a broader sense, there are three types of analytical methods: **Descriptive**, the most straightforward that summarizes and describes datasets, **Predictive**, which focuses on trying to predict future events or possibilities and **Prescriptive**, a type of analysis used to study cause-effect relationships (Sivarajah et al., 2017). There are a set of techniques and tools available. Among the several, Yaqoob et al (2016) enumerate **(1) Data Mining**, that enables the discovery of patterns/relationships between variables, **(2) Web Mining**, which allows the identification of patterns of online use **(3) Visualization Methods**, related to information presentation in form of graphs and dashboards, **(4) Machine Learning**, referring to the computational behaviours that use data as a basis **(5) Optimization Methods**, which are used to deal with quantitative questions and, finally, **(6) Social Network Analysis**, used to study social relationships network wise. In their article, Gandomi and Haider (2015) also refer **Text Analytics** (also known as Text Mining) and **Video Analytics**, both focusing on the respective data format the name suggests, while Philip Chen and Zhang (2014) add **Statistics** to the list. Statistics aim to explore causal relationships and correlations among different objectives. When it comes to tools, Rodríguez-Mazahua et al (2015) organize them into groups according to the type of analysis that is being made. They could either be tools for **batch processing** (data is gathered, stocked and only then analysed), for **stream processing** (when data needs to be analysed promptly) or **interactive analysis** (where data is processed and allows users to perform their own analysis).

However, aside from its clear potential and the tools and techniques that are emerging, extracting knowledge from BD bears considerable challenges. Even nowadays, it is not an easy task to uncover the potential of BD related to unstructured data (Misra et al., 2014). Although Agarwal & Dhar (2014) bespeak about the progresses made, deciding how to collect the most decisive data as it is created and make it reach the right person in perfect timing is still arduous (Misra et al., 2014). Sivarajah et al (2017) divide the existing challenges into different categories, one related to data itself, another regarding processes and finally one linked with management. In a broader sense, it is common to consider issues such as data privacy and security, data quality, investment justification and lack of qualified personnel (Lee, 2017).

2.2. Text Mining

About 80% of the world's data comes in an unstructured format (Schneider, 2016). Tan (1999) had previously referred a study that showed that 80% of an organization's information was enclosed in text documents. Coincidently, years later, Grimes (2008) also stated that 80% of the business relevant data is not structured and presents itself mostly in text, so it appears to be a trend that remains throughout time. Thus, and considering text is the most prevailing type of stored information, analyzing it presents more business-based potential when compared to structured data (Tan, 1999; Chen, Mao & Liu, 2014). As an example, companies hold and/or can access a diverse set of documents that may be of value, such as emails, social media posts and comments and surveys' answers (Gandomi & Haider, 2015).

Hearst (1999) recognized the hidden promises that this type of data beholds, but also acknowledged the difficulty to uncover them automatically. Text Analytics, or Text Mining, is "*the process of extracting interesting and non-trivial patterns or knowledge from text documents*" (Tan, 1990; p.65). It comprises statistical analysis, computational analysis and machine learning. Several methods are available to extract information from text, such as **(1) Information Extraction**, that identifies and collect structured data from unstructured text, **(2) Text Summarization**, which enables the production of a summary of one or multiple documents and allows the user to get a brief overview, **(3) Question Answering**, that seeks to provide answers to questions in natural language and **(4) Sentiment Analysis** (otherwise known as Opinion Mining), which aims to analyze text containing opinions about entities (Gandomi & Haider, 2015). Although this does not intend to be a comprehensive list, **(5) Topic Tracking**, a technique mostly that is applied

to suggest/predict possible interests by keeping track of a user profile, **(6) Clustering**, used to group analogous documents without using predefined topics; **(7) Concept Linkage**, that links documents by discovering shared concepts, and finally **(8) Categorization**, that identifies the main theme of a document, are also worth a mention (Fan et al., 2006).

While machines are prepared to process structured databases automatically, text was meant to be read by people (Hearst, 2003). "*The key to text mining is creating technology that combines a human's linguistic capabilities with the speed and accuracy of a computer*." (Fan et al., 2006; p.78), and Natural Language Processing (NLP) has appeared to produce technologies able to teach natural language to computers and allow them to comprehend, analyze and even produce it (Fan et al., 2006).

2.3. Natural Language Processing

Text can be generated in any language and have different modes and genres. In fact, there is only one requirement to text creation, and it is that it must be in a language used by humans as a form of communication to one another. Most of the times, the text subject to analysis was not even created with that purpose (Liddy, 2001).

A considerable group of Text Mining products are based on NLP (Tan, 1999), so understanding this concept is primordial. NLP, also known as Computational Linguistics (Liddy, 2001), goes back to the early 50's and was the result of the intersection of artificial intelligence and linguistics (Nadkarni, Ohno-Machado & Chapman, 2011). "*Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things*" (Chowdhury, 2003; p.51) and its goal is to develop techniques and tools with that in mind (Chowdhury, 2003). At the moment, it is mainly a data-driven field that uses statistical and probabilistic computations alongside with machine learning techniques (Otter, Medina & Kalita, 2019), but its foundations lie in a variety of disciplines, them being: computer and information sciences, mathematics, linguistics, electrical and electronic engineering, robotics, artificial intelligence and psychology. (Chowdhury, 2003).

According to Chowdhury (2003), creating programs that are able to understand natural language involves three major challenges. The first is related to thought processes, the second linked with meaning and representation of linguistic inputs, and finally, there is the issue of world knowledge. Both Liddy (1998; 2001) and Feldman (1999) defend

the existence of different levels of linguistic analysis, from where it is possible to extract meaning: **(1) Phonological**, related to speech and the way words are pronounced (relevant in systems with spoken inputs), **(2) Morphological**, which deals with the smallest units of meaning of a word (called morphemes and may include suffixes, prefixes and roots), **(3) Lexical**, referring to the word level analysis, as well as parts of speech, **(4) Syntactic**, that analyzes the words that compose a sentence in order to uncover its grammatical structure, **(5) Semantic**, which aims to determine possible meanings of a sentence, **(6) Discourse**, that interprets the structure and meaning that texts larger than a single sentence beholds, and **(7) Pragmatic**, a level that depends on knowledge about the outside world and is concerned with context and the situational use of language. This is important, as not every NLP system tackle every level. A system can involve all, or just some of the previously exposed levels of analysis (Liddy, 2001; Chowdhury, 2003).

The process for knowledge discovery through NLP document processing requires an essential step called Pre-Processing, which consists in the application of several techniques to prepare the text for its analysis (Gharehchopogh & Khalifelu, 2011). The process may involve a series of different steps that could include expanding abbreviations (e.g., "*asap*" means "*as soon as possible*"), removing the stop words (e.g., "*the*" and "*a*", which hold no significant value) (Haddi, Liu & Shi, 2013) and **Stemming** (removing the suffixes, e.g., "*consulting*" and "*consultant*" would both be considered "*consult*") (Symeonidis, Effrosynidis & Arampatzis, 2018). As an example, data extracted from the web is usually noisy and contain a lot of uninformative parts, so pre-processing plays a major role (Haddi, Liu & Shi, 2013). NLP also has some other widely known benchmark tasks. For instance, Collobert et al (2011) expose the most common approaches to syntactic and semantic information. Among others, in the first case, there is **Part-of-speech** (POS), which identifies every word and designate them with an indicative tag of their syntactic role (e.g., within a sentence it would determine the subject, verbs, pronouns, etc.). Regarding semantic information, the authors include, for example, **Named Entity Recognition** (NER), that tries to categorize the different elements that compose a sentence (e.g., categories might include "*person*" or "*event*"). **Tokenization** is also a crucial technique for the majority of NLP applications. It consists in splitting a sentence into tokens (e.g., "*I love this*" Would become "*I* | "*love*" | "*this*"). It could also be applied to documents and the tokens would therefore be the different sentences comprised within (Sun, Luo & Chen, 2017).

Although words have an ambiguous nature and linguistic variation poses as a challenge to NLP, its methods to process textual information are considered efficient (Gharehchopogh & Khalifelu, 2011). Recently, the field of NLP has witnessed several advances and much progress has been made. Hopefully, in the future even more of its potential will be unlocked (Hirschberg & Manning, 2015) and, as it gains further relevance, it is expected that software and tools used in this area become commodities and achieve a high user friendliness (Nadkarni, Ohno-Machado & Chapman, 2011). According to Cambria and White (2014), the answers for evolution in the field may rely on trying to teach NLP systems not only how to handle factual knowledge, but also to understand cultural nuances and emotions.

2.4. Sentiment Analysis

2.4.1. Concept

One of the most challenging NLP research topics within the Text Mining realm, is SA (Liu, 2012). SA, also commonly referred to as Opinion Mining, is the "*computational study of people's opinions, attitudes and emotions toward an entity*" (Medhat, Hassan & Korashy, 2014; p.1903). As a title of example, an entity may represent a topic, an event, an individual, a brand or organization or a product. Although SA and Opinion Mining are generally used with equal meaning, some authors defend that the terms are quite different and do not express the same. SA can also be considered a classification process (Medhat, Hassan & Korashy, 2014). An opinion can be seen as a quintuple (e_i, a_{ij}, s_{ijkl}, h_k, t_l), where e_i represents an entity, a_{ij} aspects of that entity, s_{ijkl} the sentiments regarding those aspects, h_k the opinion holder and t_l the time when the opinion was expressed (Liu, 2012). Liu (2012) distinguishes distinct types of opinions. It is commonly agreed upon that there are two types of sentences: objective sentences that refer to facts and subjective ones that contain opinions, beliefs, and perspectives about an entity (Feldman, 2013). Opinions can be regular or comparative (e.g., "*The audio sounds great on those speakers*." vs. "*X's speakers are better than Y's*."), explicit or implicit (e.g., "*This cable's quality is bad*." vs. "*I bought this cable last week and it broke already*.") and direct or indirect (e.g., "*This drink tastes awful*." vs. "*After drinking the juice, I felt sick*."). It is important to mention that the categories of direct/indirect opinions are sub-domains of regular opinions, and both regular and comparative can be considered explicit/implicit.

In a broad sense, authors such as Liu (2012), Feldman (2013) and Medhat , Hassan and Korashy (2014) consider the existence of three levels of SA: **(1) Document Level**, a type of analysis that assumes there is only a single entity to be evaluated and expresses the overall predominant sentiment on the entire document, **(2) Sentence Level**, a more in depth analysis that relies on the assumption that each sentence contains one general opinion towards an entity **(3) Entity/Aspect Level**, the most complex of all, and the analysis that reaches the most detailed results. On the latter level of analysis, every entity and its aspects are identified and classified accordingly (e.g., "*The audio of this phone is good, but the image does not have great quality*", the device's audio would have a positive sentiment associated, while the image would have a negative one). Usually, there are three possible sentiment classifications: **Negative**, **Neutral** or **Positive** (Feldman, 2013). There are authors that experiment with more categories, including Mohammad and Turney (2012) that explore the Plutchik's wheel of emotions (Appendix 1). This practice, despite being interesting and worth investigation, can generate too much complexity (Llombart, 2017).

"*Opinions are central to almost all human activities and are key influencers of our behaviors. Our beliefs and perceptions of reality, and the choices we make, are, to a considerable degree, conditioned upon how others see and evaluate the world. For this reason, when we need to make a decision, we often seek out the opinions of others. This is not only true for individuals but also true for organizations*" (Liu, 2012; p.5), so this topic rises with innumerous potential and opportunities. Although SA is being applied to almost every domain, from healthcare to social events, and its multiple commercial applications have boosted the industry (Liu, 2012), the majority of its use lies on Marketing, Finances, and Political and Social Sciences (Gandomi & Haider, 2015), as people are often keen to know opinions about a product before they buy it or interested in knowing the general opinion about a certain political candidate before they decide on their final vote. Brands also find extremely relevant to know the consumer's opinion in order to improve their products and increase sales (Pang & Lee, 2008; Liu, 2012), and in Politics, candidates are able to measure the impact of their campaigns and plan their next moves (Feldman, 2013). Albeit promising, there are several challenges that need consideration when it comes to SA. For instance, words can have different meanings depending on the context, sentiment words can be applied in sentences that express no opinion, or oppositely sentences with no sentiment words can bear opinions within, and

the use of sarcasm can mask the true intentions behind the text (Liu, 2012). Negation and domain dependence also appear as major issues (Hussein, 2018).

2.4.2. Approaches

To perform SA there are two major groups of approaches: **Machine Learning** and **Lexicon-Based** (Appendix 2). The first applies machine learning algorithms combined with the use of linguistic features, while the latter uses a collection of precompiled sentiment terms denominated as sentiment lexicon (Medhat, Hassan & Korashy, 2014).

Within the Machine Learning approach, it is possible to observe two different paths: **Supervised Learning** and **Unsupervised Learning**. The first relies on the existence of labeled documents for training. In the training process, the machine receives a properly labeled dataset with the desired outcomes and after tries to classify new unknown documents based on the examples given. Multiple classifiers, such as Decision Tree Classifiers, Linear Classifiers, Rule-Based Classifiers and Probabilistic Classifiers can be applied to train the algorithms (Medhat, Hassan & Korashy, 2014). Detailing how each of those work is, however, out of the scope of this project. Supervised classification methods have presented good results and achieved high accuracy, but they can be extremely sensitive to domain, as a classifier trained with a labeled with a document from one domain, usually does not perform well when applied to a different one (Taboada et al., 2011). They are adaptable and one can train a model for a specific purpose, nevertheless, labelled data is not always available and obtaining it can be costly (Gonçalves et al., 2013). The issue of domain specificity itself has been gaining attention and cross-domain approaches have been studied (Pang & Lee, 2008). For cases where annotated data is not available, unsupervised learning can be applied. Unsupervised techniques do not require annotated documents and instead rely on available unlabeled documents on which they try to identify patterns and similarities. Both supervised and unsupervised techniques can be combined (Medhat, Hassan & Korashy, 2014). Ultimately, one could say that a machine while learning through supervised methods learns by example and when it is through unsupervised techniques it learns by observation (Cambero, 2016).

When it comes to the Lexical approach, it can also be divided into two categories: **Dictionary-based** approach or **Corpus-Based** approach. Dictionary-based methods work by manually collecting a few opinion words with known orientations. Afterwards,

the set is grown by searching for their synonyms and antonyms, for instance, in thesaurus. As the new words are added, a new iteration starts, and the process is repeated until no more new words are found (Medhat, Hassan & Korashy, 2014). It is fairly easy to find a considerate list of words with their orientation, as well as manually correcting eventual existing errors. The dictionary approach is not able to deal with domain and context specific orientations, however corpus-based methods can help overcoming this challenge (Liu, 2012). These methods are dependent on syntactic patterns or patterns that occur alongside with a seed list of opinion words in order to find others in a large corpus. This can be done either using statistical or semantic approaches (Medhat, Hassan & Korashy, 2014). Using corpus-based techniques by themselves is usually not as effective as the dictionary-based approach because it is difficult to prepare massive corpora that cover the totality of the existing words pertaining in a language, however it has the advantage of being able to help find context specific words and deal with different domains (Liu & Zhang, 2012). Whilst Lexicon-based methods can be preferred when trying to simulate the effects of linguistic context (Taboada et al., 2011), they may be associated with low recall performance, as they depend on the presence of opinion words. Those words can be added, but expressions change rapidly, and new ones keep appearing as trends emerge. Consideration for domain dependent polarities must be taken as well (Zhang et al., 2011). Furthermore, there are many lexicons available and this multitude of options presents itself as a challenge when deciding which could perform better in the task at hands. Using information from all is not feasible, as some lexicons contradict themselves and scores for the same words are not typically equal (Hammer et al., 2015). One of the grand advantages is the fact that it does not rely on labelled data and consequently does not need to go through the training process (Gonçalves et al., 2013).

There is no overall conclusion on whether there is a better approach for each scenario, as they all have their weaknesses and advantages (Gonçalves et al., 2013). Ravi and Ravi (2015) defend that while Machine Learning provides maximum accuracy, approaches based on semantic orientation offer better generality. Authors such as Zhang, Wenyan and Jiang (2014) point to studies that show supervised machine learning methods having higher precision but lexicon-based being extremely competitive and not needing as much effort. They also refer the latter are not sensitive to the quality and quantity of the training dataset, which may pose as an advantage. In an attempt to gather the best of both worlds and attain better results, a **hybrid approach** that combines Lexicon-Based with Machine Learning is surfacing and several studies in this topic have already taken

place (Ravi & Ravi, 2015). On those, sentiment lexicons often play a major role (Medhat, Hassan & Korashy, 2014).

2.4.3. Sentiment Analysis on Twitter

Recently, opinionated postings published throughout social media have helped transform businesses and influence public sentiments in a way that ultimately impacted both the social and political systems (Liu, 2012). Hence, the importance of analyzing data originated on those sources has been increasing. Among the existing social networks and forums, there is one that has been gaining more attention due to its high potential: Twitter. Twitter is a social network founded in 2006 (Aslam, 2020) with about 330 million active users that are predominantly based on the United States, followed by Japan and India (Clement, 2019; 2020). Daily, there are around 500 million published posts (Aslam, 2020), which makes it an extremely rich source of information. To understand the Twitter dynamics, it is important to take a quick glance through it. Users have their own profiles and can follow and/or be followed by other profiles. Posts are called "*tweets*" and a user can like another user's post. They can also share them in their own page, which is denominated as a "*retweet*". Another interesting feature worth exposing is the ability to mention other users in their tweets (using "*@user*" syntax). *Hashtags* (#) to highlight and identify certain topics are a common practice and Twitter also allows user interaction, where people can engage in public conversations either through "*reply*" or via private message. This platform places a character-limit to the posts, which cannot surpass 280 characters (previously 140), although only around 5% of them surpass 140 (Rosen, 2017). The limitation imposed to the users motivate them to be straightforward and, due to the short length of the tweets, generally no bigger than one sentence, SA and other techniques can be performed at sentence level. Regarding to SA, there is also a common assumption that a tweet expresses an opinion about a single entity (Bravo-Marquez, Mendonza & Poblete 2014). Users can also upload images or videos and recently, a new upcoming feature of voice tweeting was introduced, allowing users to post up to 140 seconds of audio (Patterson & Bourgoin, 2020). Furthermore, Twitter is particularly interesting due to the range of personalities and organizations that use it. It goes from general users to big corporations, celebrities and even governors and legislators (Bharat & Murthy, 2016). This allows the gathering of information from different social and interest groups (Pak & Paroubek, 2010). Because of its wide availability and the fact that it does not require high end technological products to be used, anyone could potentially access Twitter, even on

less developed countries. Its low learning curve is also encouraging to new users (Murthy, 2011).

As of now, this specific platform has been the focus of research in very distinct domains. There are studies about how Twitter can be useful to predict crimes (Gerber, 2014), stock markets (Bollen, Mao & Zeng, 2010) and political tendencies (Pla & Hurtado, 2014), to detect influenza epidemics (Culotta, 2010; Aramaki, Maskawa & Morita, 2011) and track other public health issues (Paul & Dredze, 2011), to deal with scandals (Tse et al., 2016) and even to understand environmental concerns (Reyes-Menendez, Saura & Alvarez, 2018). A great part of those investigations falls in the SA dimension, in fact, despite the wide range of practices applied to Twitter, the SA community has taken a special interest in studying the platform (Martínez-Cámara et al., 2012). An encouraging study from O'Connor et al (2010) has shown a relatively strong correlation between public opinion measured with polls/surveys and sentiment expressed on this platform, hence, enterprises and entities are researching new methods to extract knowledge from this source (Koloumpis, Wilson & Moore, 2011). However, as expected, mining Twitter and performing SA bears its challenges. With the amount of data generated, the topics cover almost every domain, which might become overwhelming (Koloumpis, Wilson & Moore, 2011). The use of jargon and informal expressions, as well as misspelled and/or abbreviated words are also in need to be dealt with, not to mention the difficulty related to the lack of context some tweets might present, which is a problem for SA systems (Martínez-Cámara et al., 2012).

2.5. Public Image and Reputation

The rise of social media made it easier to disseminate information of every nature at a rate never seen before. Furthermore, social media has a special effect of presenting what is defined as a collective truth: users search, create and share information, and what was once singular and subjective becomes available and collective. However, the interactivity that defines social networks brings several implications for organizations, as they must adapt and have a proper strategy of communication. If they do not manage to do it properly, or if their actions are of a questionable nature, their reputation is threatened and put at risk, which may affect several aspects of their performance. For instance, it could affect their competitiveness, their positioning and even the trust and loyalty of stakeholders (Aula, 2010). Summarily, reputation has become a strategic asset and its loss can have astounding costs (Floreddu, Cabiddu & Evaristo, 2014), so the importance of

reputation management as a business function increases, much as Hutton et al (2001) once theorized early on.

But this is not only a reality for corporations and businesses, it is also for humans as well. Reputation plays a major role in people's life and is a central aspect of social identity that is shaped by discussion in social networks (Jazaeiri et al., 2018). Individuals pertaining in communities and/or groups are eager to share and collect accurate information about each other, as it is a facilitator of behaviour expectancy. Whether a person values his reputation or not, it is important from both a societal and an individual point of view (Cavazza, Pagliaro & Guidetti 2014). Studies have also shown that most people tend to counteract and try to reverse the situation when the public has constructed a negative or unfavourable impression of them (Baumeister, 1982) and that having a positive image is a universal worry (Cavazza, Pagliaro & Guidetti 2014). Now, with the advent of social platforms, managing one's online identity has become an even more important task. In fact, it is possible to state that reputation management has become a completely defining feature of online presence for many users, as research shows that people are getting more concerned about what kind of information about them, and also the ones around them, is shared and made publicly available (Madden & Smith, 2010). This has even more impact on public figures, whose public image and reputation may dictate their success or downfall.

It is also important to state that the social media dynamics regarding reputation is particularly interesting now, as the rise of both the Stan and Cancel Culture is witnessable. They are polarized behaviours on opposite sides of the spectrum, the first referring to the act of liking something greatly and/or being a zealous fan of someone, and the other to the withdrawal of support of public figures or organizations after they acted objectionably or offensively in some way (Dictionary.com, n.d). The latter often occurs in the form of public shaming and Ng (2020) highlights how that practice is proof of the way information is quickly propagated and how acts considered problematic generate fast and large-scale responses, originating big impacts on how those who practiced them are viewed.

3. Methodology

This project aims to create a system able to assess public figures' reputation through Twitter using SA. In its essence, this is a mining problem and the goal is to extract valuable information from sets of data. Therefore, following good practices and adequate

models/frameworks is essential. To mine data there are three well-known frameworks: KDD (Knowledge Discovery in Databases), SEMMA (Sample, Explore, Modify, Model, Assess) and CRISP-DM (Cross-Industry Standard Process for Data Mining). Fundamentally SEMMA and CRISP-DM are an application of the KDD process. They can be considered equivalent, although at a first glance CRISP-DM appears to be more complete than SEMMA (Azevedo & Santos, 2008). For this project, **CRISP-DM** will be the adopted methodology. Although it has its limitations, it is adaptable and is still one of the most used frameworks for this type of project (Piatetsky, 2014). The reference model, with all the phases that compose it, as well as its overall cycle can be found on Figure 1. Next to it, a brief explanation of each phase and how they can relate to this project will also be visible. Chapman et al (2000) have created a comprehensive step-by-step best practice guide for this framework, which will serve as a guideline for this project.

The phases, based on those authors (2000) include:

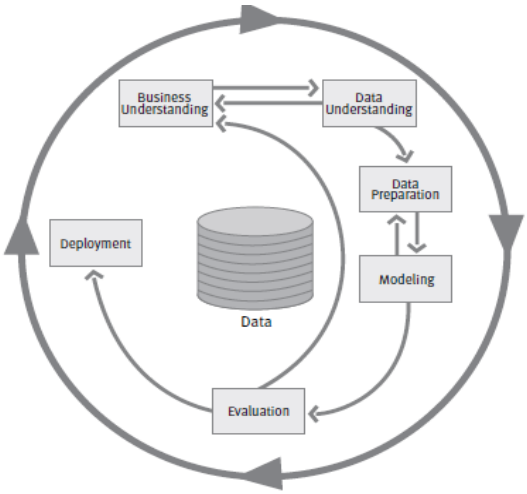


FIGURE 1 - CRISP DM Reference Model (Chapman et al., 2000, p.10)

- **Business Understanding** – Initial step that focuses on understanding the goals from a business perspective. This stage is clear earlier on this document, during the Literature Review where an overview of concepts is shown, and the relevance of this project is evidenced.
- **Data Understanding** – Refers to the initial data collection and the activities that enhance its most important characteristics. This stage will correspond to the elaboration of the dataset.
- **Data Preparation** – Includes all the tasks required to create the final dataset. On this project, it will also be related to the dataset, as it will need to be prepared to suit initial objectives and pre-processed for further analysis.
- **Modeling** – This phase corresponds to the selection of the techniques used and the calibration of their parameters. When it comes to this project, an appropriate SA approach and techniques within shall be chosen.
- **Evaluation** – After having built a model, it is important to evaluate its performance and assess if it corresponds with what was initially planned.

According to the techniques chosen for the project, metrics will be applied to measure how it performs.

- **Deployment** – Creating a model is usually not enough. It should be documented and presented properly. Final remarks and suggestions for future work will be included in the report.

The sequence of the phases is not strict and depends on the specific project and its outcomes (Chapman et al., 2000). The model, even on its early stages, has been tested and the results were positive (Wirth & Hipp, 2000).

Disregarding the hybrid that combines both, SA has two major groups of approaches, thus, to move forward, it is necessary to choose one as a starting point for this study. It is difficult to decide which method is more appropriate, as they both have their own strengths and weaknesses. Ultimately, one could say that there is no conclusion whether one method outperforms the other (Gonçalves et al., 2013). However, because it is known for being adaptive to changing inputs and its accuracy (Thakkar & Patel, 2015; Ahmad et al., 2017), and also for the fact research has shown that for Twitter SA its performance is better (Gonçalves et al., 2013), for this project the **Machine Learning Approach** will be the one adopted. Zhang et al (2011) also denote that Twitter data dynamics is also a problem for lexical approaches because there is an abundance of colloquial expressions, emoticons and abbreviations, among other features, that generally are not contemplated in opinion lexicons, which may affect the final results of its methods. The same authors point to the low recall problem of this method, that is dependent of the existence of opinion words, that could indeed be added but would still be considered problematic, as new trends and expressions are continuously appearing and their meanings could change within domains. Simultaneously, they also discuss the problem of domain dependency on machine learning methods, but this project aims to create a specifically created classifier for the purpose. Since the goal is to create domain specific corpora to train a classifier, the path chosen within the machine learning approach was the supervised. Due to the fact that tweets have a short character limitation and are usually no longer than a sentence, SA will be performed at sentence level and the assumption that only one entity is represented will be followed (Bravo-Marquez, Mendonza & Poblete 2014).

The project was divided in two different phases, the first corresponding to the dataset creation and classifier training and the latter to the final system elaboration (Figure 2).

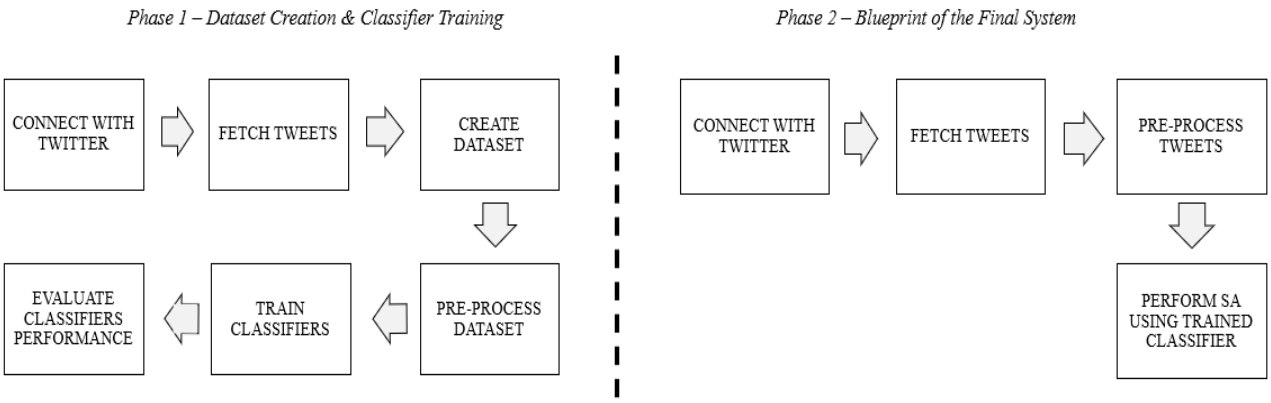


FIGURE 2 - Phases of the Project

The scripting language used throughout the project is Python[2]. Despite being a mature programming language, it is deemed as beginner friendly and easy to understand, as well as flexible (Igual & Seguí, 2017), which made it easier to choose from amongst the other available languages. Furthermore, it supports rapid prototyping and can be easily used to write structured object-oriented programs (Loper & Bird, 2002). As of now, and according to the TIOBE Index[3], Python ranks as the 3rd most popular programming language. Visual Studio Code[4] was the chosen IDE (Integrated Development Environment). The libraries and packages used, as well as other resources, will be referred to along the way.

4. System Creation

4.1. Connection to Twitter

In order to be able to stream and download tweets, a connection between Twitter and Python must be established. To achieve this, one must create a Twitter account and apply for it to be upgraded to a Twitter Developer account. After the application is successfully made, credentials to establish the connection become available. Finally, and using Tweepy[5], a library for Python that helps it access the Twitter API (Application Programming Interface), the link between both is made. Note that the Twitter API limits the number of requests.

4.2. Creation of the dataset

[2] https://www.python.org/
[3] https://www.tiobe.com/tiobe-index/
[4] https://code.visualstudio.com/
[5] https://www.tweepy.org/

As previously discussed, sentiment classification is very sensitive to domain, and classifiers trained using a dataset from one domain can perform poorly when used on other (Liu & Zhang, 2012). Although the process of labelling data ought to be a demanding and often time consuming (Liu & Zhang, 2012), labelled data is deemed a major contribute to the SA area, especially when it comes to supervised learning methods (Pang & Lee, 2008) and Owsley, Sood and Hammond (2006) have highlighted the promising results that domain specific corpora have for performing SA on text. One of the aspects that distinguish the system created from others available is the fact that it will be based on a dataset created specifically to the chosen domain, which would hopefully increase its performance there (Pang & Lee, 2008). It is also important to state that manually annotating data requires effort and the results from the trained classifier are sensitive not only to the quality and quantity of data available, but also to the existence of bias (Zhang, Gan & Jiang, 2014). The system is intended to function within the people's dimension and be able to assess the overall sentiment regarding a public figure as a means of knowing if they are in good or bad terms when it comes to their reputation. Hence, there was an effort to be as inclusive as possible in the types of public figure covered in the dataset. For that, the defined categories were: **Activist**, **Actor**, **Business**, **Celebrity**, **Internet Personality**, **Model**, **Politician**, **Religious Figure**, **Royalty**, **Singer**, **Sport** and **TV**. To avoid unnecessary complexity (Llombart, 2017), it was decided that there would only be three possible classifications: **positive**, **neutral** or **negative**. Hereby, this becomes a multi-class classification, and each input is to be classified into only one class. The classes do not overlap each other (Sokolova & Lapalme, 2009). As the dataset was going to be manually annotated, a set of labelling rules (Appendix 3) was elaborated to attempt consistency and cohesiveness throughout. However, there were cases difficult to interpret which could have had different classifications.

A script that allowed the stream of tweets and their placement into a data frame format was created. For that, two functions were necessary: **(1)** a function (*get_save_tweets*) to retrieve the tweets, that worked by prompting the user to insert the name of the public figure they wanted to search for, and **(2)** one to place the streamed tweets into a data frame structure (*tweets_to_df*). Both Jsonpickle[6], a Python library to work with JSON (JavaScript Object Notation) format files, and Pandas[7], a Python library for data analysis

[6] https://jsonpickle.github.io/
[7] https://pandas.pydata.org/

and manipulation, were used on this step. The primary focus was on the first function, as it had to have several details that improved the results obtained. A series of decisions were made:

- The fields collected are the id of the tweet, the text itself and the date/hour it was posted. Although it is possible to gather the poster's user and screen name, those were excluded in order to maintain their privacy.
- Every new search appended its results to the previous one, creating a single file with all the tweets.
- By default, the streamed tweets are a mix between the most recent and the most popular, and it was pertinent to maintain it that way.
- Although some translation tools are already available, their output is not always accurate. Thus, for this project only tweets in the English language were considered. Although they seldom appeared, some tweets that only contained some words in English (but whose main language was different) were included in the search results. This challenge is recognized as Multilingual Content and happens when a tweet is posted mixing more than one language (Giachanou & Crestani, 2016). When faced with this, and since the cases were minimal, the choice was to exclude them from the dataset.

Besides this, some crucial filters were added to improve the results. It is important to denote that in this step, trial and error was necessary to attain a combination of filters that would exclude and try to minimize tweets that did not add great value:

- First, it was decided that retweets would not be included, as the same tweets would later be classified the same way. Variety was valued in the training process. This benefits the training process, as it tries to avoid attributing specific tweets more weight (Go, Bhayani & Huang, 2009).
- Quotes are not an indicator of someone's reputation, as a person can like what was once said by somebody but not particularly like or dislike its author. Therefore, there was an effort to reduce their appearance. Tweets that had *#Quote* were filtered out, and those which did not, but were indeed quotes, were classified as neutral.
- Similarly, whenever the searched personality was related to music, there were several tweets that just contained a reference that the user was listening to a song. That was considered as not being indicative of the singer reputation. Consequently, the most common *hashtag* related to this scenario (*#NowPlaying)*

was designed to not be a part of the search results. As expected, this type of post, but without the *hashtag*, kept appearing. The decision was to classify them as neutral.

- Another discovery made in the process of creating the dataset was that most tweets that included links to another pages were news or, in the cases of public figures that own companies, promotions and/or campaigns for their products. For this reason, as they are objective and not indicative of someone's reputation, the automatic filtering excluded URLs (Uniform Resource Locator).

It is of extreme importance to note that filters are not 100% infallible, and sometimes tweets that should not have been included in the search results still appear. In these cases, they were labelled accordingly as to prepare the system to know the answer to this type of post. Finally, and to conclude the thought process of the creation of the document that would contain the tweets, the logic of not including retweets was followed, and duplicates were removed as well. For an easier data manipulation, the file, originally in JSON format, was converted into XLSX so it could be accessed and modified through Excel.

The creation of the dataset is a critical yet very challenging task. Sometimes it was hard to detect sarcasm and irony, and there were also cases where the interpretation would change considering the context. Although it organically contained tweets that were considered sarcastic, there was an effort to incorporate more by following the guidelines of González-Ibáñez, Muresan and Wacholder (2011), that state that there is no better judge to identify if something is sarcastic or not than the author of the tweet himself. Inspired by their work, some tweets that included the #sarcasm following the name of a public figure were fetched. This, however, was not as fruitful as expected, as there were not many tweets that contained both the name of someone and the pretended hashtag (*#sarcasm*). Besides, they were mostly related to political personalities, just as Feldman (2013) indicated in his paper. Nonetheless, the choice was to maintain them in the final dataset. There were also other difficulties throughout the labelling process. Although there is a common assumption that a tweet only expresses an opinion about a single entity (Bravo-Marquez, Mendonza & Poblete 2014), some posts contained more than one entity, each with their sentiment associated. Contemplate: "*Trump knows history will remember him to be a failure and a fraud. Trump also knows history will remember Greta Thunberg to be a hero.*"; here it is evident that there is a positive opinion about Greta Thunberg, but a negative linked to Donald Trump. Additionally, more than a few cases where the name of the public figure appeared, and the predominant sentiment was not attached to them,

emerged. Take the following the example in consideration: "*Before You Go by Lewis Capaldi will always remind me of my papa. That man was so loving and selfless. My life is so incomplete without him.*"; here, it is possible to assess that whoever posted this was probably sad and had a predominantly negative sentiment, but when it comes to the proper sentiment/opinion regarding the searched personality it is mostly neutral.

The final labelled dataset was composed by 2025 entries, equally distributed by the 3 desired types of classification, which means there were 675 positives, 675 neutrals and 675 negatives. On Appendix 4, it is possible to see how they are scattered between the previously defined groups of personalities. Bear in mind that categorizing public figures into types is not a linear process and, in most cases, they belong to more than one group (e.g., Cynthia Nixon is an actress but also majorly connected to politics and Miley Cyrus is mainly a singer but has acted too). Something that is also fundamental to denote is the fact that some types of personalities are not as tweeted about compared to others, hence the disparity in the number of tweets belonging to each group. This is particularly visible when regarding the religious figures and royalty members, at least at the time of the elaboration of this project. Collecting and categorizing tweets pertaining to different public figures was essentially done to diversify and dynamize the dataset, as it is interesting to understand the different dynamics of posts regarding different types of people. However, what matters the most is the balance between the classification groups (negative, neutral, positive), since having balanced classes in a dataset benefits the training process (Rahman & Davis, 2013) and classifiers often perform badly when faced with imbalanced datasets (Akbani, Kwek & Japkowicz, 2004). As previously stated, each class has an equal amount of posts associated (675).

To try to diminish the bias of the labelling process, the dataset was then sent to two fellow colleagues from the Information Systems Management Master. The initial evaluation of the author was then discussed by all and some changes were made. A snippet of the final dataset overall appearance can be found on Appendix 5 and more details are observable on Appendices 6 and 7. As previously indicated, this step of the system creation is directly connected to the Data Understanding process of the CRISP-DM Model.

4.3. Preparing the Data, Training and Testing

4.3.1. Pre-processing Techniques

After the creation of the dataset that will be used to train and test the classifier, the next step is to prepare it so it can be used and interpreted by machines. The pre-processor will also be used in the future, so that new unlabelled tweets can be interpreted as well. Pre-processing is an essential step for SA and even more relevant when applied to text originated on microblogging platforms such as Twitter (Symeonidis, Effrosynidis & Arampatzis, 2018). This process corresponds to CRISP-DM's Data Preparation.

Twitter has very interesting dynamics. People usually use a very informal language and tend to create their own words and terms, as well as inventing shortcuts/abbreviations and recurring to slang (Singh & Kumari, 2016). Users tend to not care about the correct use of grammar and the lack of context sometimes found is a strenuous problem that must be dealt with (Martínez-Cámara et al., 2014). That was clear when exploring the platform, as well as while creating and preparing the dataset. In fact, performing SA on Twitter data is considered a much harder task when compared to performing it on conventional text (Saif, He & Alani, 2012) and this has been a challenge the SA community has faced since 2009 (Martínez-Cámara et al., 2014). Terms associated with the "*Stan*" and "*Cancel*" culture mentioned in the Literature Review are a current common practice. Additionally, another characteristic noticed is that some words commonly associated with bad connotations are used to praise acts or individuals. Finally, one must take into consideration, that besides what was already been referred, the existence of bots posting is also something to recognize. On Twitter, there are also new trends and challenges that users tend to follow, so keeping up with the constant evolution of terms and expressions is very demanding. With this in mind, a few pre-processing techniques were chosen to transform the text:

1. **Lowercasing:** lowering all the words was the first step carried out on this project's pre-processing task. It means that, for instance "*high*", "*HIGH*", "*High*" would all be transformed and treated equally as "*high"*. This helps reducing the dimensionality (Symeonidis, Effrosynidis & Arampatzis, 2018). Note that uppercasing text usually has an underlying motive, but the system does not contemplate these features because it can overload it and induce other errors.
2. **Hashtag symbol removal**: the symbol # was removed, but the word attached to it was maintained. This could be useful to identify trending or new terms, or even

to maintain the meaning of the sentence (e.g., "*this is so #cool*" turns into "*this is so cool*").

3. **URL and hyperlinks removal**: URLs and hyperlinks just add noise to a text, so their presence was eliminated.
4. **Mentions/usernames removal**: mentions and usernames bear no value within, as they are used only to identify someone. Therefore, they were completely withdrawn (e.g., "*@person, see this!*" becomes "*, see this!*").
5. **Punctuation removal**: despite the use of some punctuation, such as exclamation points, may bear underlying motives and help conveying opinions (Symeonidis, Effrosynidis & Arampatzis, 2018), they add a lot of noise and hence, they were removed. The following list comprises all the characters eliminated from the text:***'!"#\$\%&\'()*+,-./:;<=>?@[\\]^_‘{|}~'***.
6. **Other symbols removal:** emoticons and emojis might convey sentiment and their importance has been the target of some research (Hogenboom et al., 2013). They can also be used sarcastically, which is an interesting feature. However, they can generate a lot of noise too and consequently will be out of the scope of this project by being removed from the text. Numbers were also chosen to be completely disregarded.
7. **Stop word removal:** stop words usually do not hold significant value, hence it is common to just eliminate them from texts (Saif, He & Alani, 2012). For instance, the expression "*It's a fruit*" would simply be reduced to the term "*fruit*". A complete list of the stop words considered can be found on Appendix 8.
8. **Length reduction:** some people exaggerate the number of letters to intensify and express opinions (e.g., "*amaaazing*"). Since this project does not use a spellcheck tool, which will be explained briefly afterwards, the word compression put into practice takes into consideration that the English language contains no word with more than two consecutive identical letters. That will certainly generate wrong terms, but it will diminish the number of different words considered (e.g., "*hellloooo*" to "*helloo*"). This kind of word length reduction can also be seen on Le and Nguyen's (2015) research. Another approach could have been keeping an extra character to identify words purposely augmented (Agarwal et al., 2011).
9. **Tokenization:** tokenizing text is a crucial step on the vast majority of NLP applications and consists of splitting the input text, usually sentences, into tokens

(Sun, Luo & Chen, 2017). As a title of example, the expression "*I like it*" would be transformed into a list of tokens ["*I*", "*like*", "*it*"].

10. **Stemming:** alongside the next technique, stemming is useful to reduce the number of different words with the same meaning processed by the machine, as well as to match similar text entries. Both base themselves in the notion that words have a root form and work in order to get the basic term meaning of a specific word (Igual & Seguí, 2017). Stemming functions by eliminating derivational suffixes and inflections (Balakrishnan & Lloyd-Yemoh, 2014). It can originate non-existing words (e.g., a stemmer could turn "*trouble*" into "*troubl*").
11. **Lemmatization:** lemmatization tries to return words to their dictionary form by removing inflectional endings (Balakrishnan & Lloyd-Yemoh, 2014). Unlike stemming, it always generates a real word (e.g., "*churches*" turns into "*church*").

Of course, there were also other tools and techniques that could have been added (Symeonidis, Effrosynidis & Arampatzis, 2018). It is important to refer that by not incorporating features such as POS tagging or NER, as a title of example, the machine learning model treats "*Will Smith*", which for humans is clearly recognized as a name, as ["*will*", "*smith*"]. Both words extracted from what was once a name can hold different meanings that completely change the context of the sentence they are in. By not identifying their role in the text structure, results can be compromised (Pietro, 2020). Those type of techniques are more refined and could even increase system's levels of linguistic analysis but are out of the scope of the project. Nevertheless, this acts as a baseline for further research. As of the pre-processing task carried out, a function for each of the mentioned techniques was created, followed by another one (*preprocess*) that combines all the ones which will be put to use and transform the text. At a first attempt at pre-processing, length reduction, stemming and lemmatization will not be applied. Only after an initial training and its respective tests, those shall be introduced in order to assess their impact on the classifier's performance and achieve an optimal combination of techniques. All the mentioned techniques were applied with the aim of Python's Regular Expressions module[8] and the NLTK library[9], which is a widely used library to deal with human natural language.

Not using a spellchecking tool to correct wrong words was a considered decision. There are some tools available for this purpose, however they are not completely infallible

[8] https://docs.python.org/3/library/re.html
[9] https://www.nltk.org/

and when dealing with this kind of platform, where users create their own words and expressions and sometimes deliberately misspell, the system could lose efficiency. Symeonidis, Effrosynidis and Arampatzis (2018) point towards the fact that a grand part of the devices used to post on social media benefit from spellcheck tool already, so errors are diminishing. They also conducted a series of tests, and the incorporation of word correction did not yield good results. Nonetheless, there are true errors that will not be corrected and that will affect the results and performance of the models. If the approach chosen was combined with a lexical one, a list with Twitter terms could be built to aid in this problem. Slang and jargon terms were also not excluded as they are recurrent on Twitter.

4.3.2. Text Representation

The previous text transformations are not enough to allow a machine to understand this type of data. To be able to process textual inputs, it must be properly represented. Note that this step also belongs to CRISP-DM's Data Preparation. For this task, it is important to contemplate the concept of word frequencies. There are several methods to pursue this step but using variants of the Bag-of-Words method is common (Igual & Seguí, 2017). With no intent to explain the mathematical details behind it, there is a variant of this concept denominated Term Frequency – Inverse Document Frequency (TF-IDF). Instead of just identifying, or not, the presence of a term, TF-IDF express the importance of that term in the document. Term Frequency reflects the number of occurrences of a particular word in the text, while the Inverse Document Frequency weigh the number of appearances of any in word on the corpus. The more the word appears, the bigger the TF-IDF value is (Tripathy, Agrawal & Rath, 2016).

Another aspect that requires consideration is the n-grams that are weighted. "*An n-gram is a contiguous sequence of n items from a given sequence of text*" (Igual & Seguí, 2017; p.191) and in this project, uni-grams (1-gram) were used, as they could provide good data coverage (Pak & Paroubek, 2010). Note that the expression "*I like apples*", while using uni-grams would be divided into ["*I*", "*like*", "apples"]. If bi-grams (2-grams) were used, the same expression would be transformed into ["*I like*", "*like apples*"].

4.3.3. Training and Testing

With the text ready to be processed and interpreted by machines, the next natural stage of the process is to train a model. Due the nature of this study, it is impractical to

train and test a multitude of classifiers and algorithms and have their performance compared. Amongst the available, Bharat and Murthy (2016) highlight the progresses and results made in classification tasks using methods such as Naïve Bayes (NB), Support Vector Machines (SVM) and Maximum Entropy (ME). Those classifiers are the most commonly used for the task at hands (Sun, Luo & Chen, 2017), thus they were the ones chosen to integrate this study. Following, a brief overview is carried out. NB is a simple yet powerful probabilistic classifier that relies on the Bayes Theorem and conditional probabilities (Medhat, Hassan & Korashy, 2014). For this project, a Multinomial NB was chosen, as studies imply that it provides good results with short documents and few training data, even outperforming SVM and ME (Wang & Manning, 2012). SVM has also proved itself to be highly efficient at text categorization tasks (Ye, Zhang & Law, 2009). Using encoding, SVM is a linear classifier that considers that features represent a position inside a hyperspace and tries to separate them into classes (Medhat, Hassan & Korashy, 2014; Llombart, 2017). Finally, ME also known as Logistic Regression (Yu, Huang & Lin, 2010), is a probabilistic classifier whose underlying principle is that data distribution, if not much is known, should be as uniform as possible. It functions by converting labelled features into vectors and then trying to calculate weights for each feature and combining them to determine the most probable label (Bhuta et al., 2014; Medhat, Hassan & Korashy, 2014).

Whereas the process of training the classifiers correspond to the Modeling phase of CRISP-DM's framework, the tests to assess their performance are connected to the Evaluation step. Another important thing to notice is that all tasks associated with this chapter were elaborated using Scikit-Learn[10], a machine learning library for Python. As already mentioned, the dataset purposely created for the creation of this system contains 2025 entries, distributed equally per the different types of classification. However, the entirety of the document will not be used for training. To be able to assess the classifiers' performance, it was split into a training dataset and a testing dataset. Although other divisions such as 80%-20% on the train-test split are more common (Brownlee, 2020), since the dataset may be deemed as small for machine learning purposes, the decision was to divide into 90%-10%. This way, the system shall have more examples to rely on. Considering this, and what was exposed before, the chosen classifiers were put under the train-test process.

[10] https://scikit-learn.org/stable/

The Evaluation step of the CRISP-DM's in this project comes immediately afterwards to assess how the trained models perform. Chawla (2005) states that a classifier is usually evaluated by a confusion matrix (Appendix 9). However, the common baseline only considers two classifications, and hence, four possible outcomes. They include True Negatives (TN), False Negatives (FN), True Positives (TP) and finally FP (False Positives). Since this project contemplates three possible classifications, the confusion matrix must be adapted to reflect the obtained results. For that, consider P=Positive, N=Negative and U=Neutral. In Table I some clarification is provided.

TABLE I

Confusion Matrix with 3 Labels (Adapted from Nakov et al., 2016)

	Predicted Positive	**Predicted Negative**	**Predicted Neutral**
Actual Positive	PP	PN	PU
Actual Negative	NP	NN	NU
Actual Neutral	UP	UN	UU

Additionally, trained models can be evaluated by several metrics (Hossin & Sulaiman, 2015), but the most used on SA are Accuracy, Precision, Recall and F-Measure (Giachanou & Crestani, 2016). Accuracy is the most predominantly used metric and evaluates how often the model made the correct prediction. To assess the exactness of the method being tested, Precision is used, as it calculates the ratio of instances of a class that were predicted correctly by the total number of instances that were predicted as belonging to that same class. On the other hand, Recall calculates the ratio of correctly predicted instances of a class and the number of instances that should have been predicted. Finally, and since Recall and Precision are usually considered not enough, a combination of both is used: F-Measure. This corresponds to the harmonic mean of the two previous metrics (Hossin & Sulaiman, 2015; Giachanou & Crestani, 2016).

$$(\mathbf{1})\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN}\ ; \qquad \textit{Equation 1}$$

$$(\mathbf{2})\ Precision = \frac{TP}{TP + FP}\ ; \qquad \textit{Equation 2}$$

$$(3)\ Recall = \frac{TP}{TP + FN}\ ; \qquad \textit{Equation 3}$$

$$(4)\ FMeasure = 2 \times \frac{Precision \times Recall}{Precision + Recall}\ . \qquad \textit{Equation 4}$$

Note that all, except one, are generic formulas for binary classification (assume there are only two possible outcomes, usually positive or negative), thus it is necessary to adapt to the project's scenario which comprises three classes (Sokolova & Lapalme, 2009). It is also interesting to mention that Precision, Recall and F-Measure can be calculated for each class, which is common practice for multiclass classification (Giachanou & Crestani, 2016). Although Prabowo and Thelwall (2009) state that ideally the performance of a classifier should be measured by the micro and macro scores of each of the previously mentioned metrics, for this project, only macro and weighted evaluations were considered, as micro-averaging results are more relevant when faced with imbalanced datasets (Prabowo & Thelwall, 2009). Scikit-learn does the calculations and provides the desired results. Note that the following values can only vary between 0 and 1, and values closer to 1 are desired.

TABLE II

Trained Classifiers' Metrics

		Precision					Recall					F-Measure				
	Accuracy	Weighted	Macro	Positive	Neutral	Negative	Weighted	Macro	Positive	Neutral	Negative	Weighted	Macro	Positive	Neutral	Negative
NB	0.76	0.76	0.76	0.80	0.73	0.74	0.75	0.76	0.76	0.67	0.84	0.75	0.76	0.78	0.70	0.79
SVM	0.75	0.77	0.75	0.87	0.65	0.75	0.75	0.76	0.68	0.81	0.78	0.75	0.75	0.76	0.72	0.76
ME	0.76	0.78	0.76	0.87	0.66	0.76	0.76	0.77	0.69	0.84	0.76	0.76	0.76	0.77	0.74	0.76

Confusion tables for each classifier can be analysed through Appendices 10 to 12. Overall, and although not ideal for it to be considered a good model, all classifiers have obtained satisfactory results (Table II). In comparison with the others, ME has the highest results in almost each parameter tested. Overall, the positive class shows the highest precision but the lowest recall. On the other hand, the neutral class appears to have the best recall and lowest precision, although this is not visible when using NB. Based on Accuracy and F-Measure of the present models, the one with better results is ME.

Besides, the typical train-test evaluation, a cross-validation test was performed in order to understand how the models would react with new, unseen, data. This usually results in a less biased and optimistic estimate when compared with the previous tests (Brownlee, 2018). There are several approaches and types of cross-validation, but in what

is called a k-fold cross-validation, the dataset is split, randomly, into k subsets of around the same size and that are mutually exclusive. The classifiers are then trained and tested k times (Kohavi, 1995). There is no formal rule for the choice of the k value, but the higher the value of k, the less bias is present. This comes with a tradeoff since the increase of k leads to impracticality and a computational onerous task. Choosing to perform a 10-k fold test is common, as it gives good results and presents a satisfactory computational efficiency (Kuhn & Johnson, 2013). Therefore, a 10-k cross-validation was carried out (Table III).

TABLE III

Classifiers' 10-k Cross-Validation

	10-K CROSS-VALIDATION	
	Accuracy	**Standard Deviation**
NB	0.47	+/- 0.21
SVM	0.47	+/- 0.19
ME	0.48	+/- 0.18

As observed, all models have a fairly poor performance, which may indicate that during the train they overfitted the training data. Overfitting is a common problem in classification tasks (Lever, Krzywinski & Altman, 2016) and it means that the classifiers might have learned to identify patterns and describe them, but not to understand the underlying relationships within, hence being incapable to predict accurate results on new data (Schaffer, 1993; Bronshtein, 2017). This is normal, as there is few training data and the models would benefit from more labelled inputs to achieve better results (Ray, 2015) This could also indicate that this kind of data, as suspected, is indeed difficult to classify (Shulga, 2018). There are other variances of the cross-validation test, perhaps conducting them otherwise would have provided different outcomes. In the future, it would be interesting to create a dataset exclusively to test and validate the trained classifiers and see if the results are similar to the ones obtained through the test carried out.

On chapter **4.3.1.** there was a reference on how some pre-processing techniques would be left out on the first train-test trial. After the initial attempt was carried out, the techniques that were not put in practice before were added individually to have their impact on the classifier measured. New accuracies, precisions, recalls and F-measures were tested, but confusion matrixes and cross-validations to the adjusted classifiers were not conducted. The first technique to be introduced was the **word length compression**,

which did not have a positive repercussion, lowering the overall results in every classifier, with EM still bearing the best outcomes (Table IV).

TABLE IV

Tests with Word Compression

		Precision					Recall					F-Measure				
	Accuracy	Weighted	Macro	Positive	Neutral	Negative	Weighted	Macro	Positive	Neutral	Negative	Weighted	Macro	Positive	Neutral	Negative
NB	0.75	0.75	0.75	0.79	0.73	0.74	0.75	0.75	0.76	0.65	0.84	0.75	0.75	0.77	0.69	0.79
SVM	0.73	0.75	0.74	0.85	0.63	0.74	0.73	0.74	0.67	0.77	0.78	0.74	0.73	0.75	0.69	0.76
ME	0.75	0.77	0.76	0.86	0.65	0.76	0.75	0.76	0.69	0.82	0.76	0.76	0.75	0.77	0.73	0.76

Lemmatization, too and in a generalized form, lowered the performance of all models (Table V). However, it slightly improved SVM's results, bringing its accuracy to 76% and outperforming the others.

TABLE V

Tests with Lemmatization

		Precision					Recall					F-Measure				
	Accuracy	Weighted	Macro	Positive	Neutral	Negative	Weighted	Macro	Positive	Neutral	Negative	Weighted	Macro	Positive	Neutral	Negative
NB	0.73	0.73	0.73	0.77	0.71	0.70	0.73	0.72	0.73	0.61	0.82	0.73	0.72	0.75	0.66	0.76
SVM	0.76	0.78	0.77	0.90	0.67	0.75	0.76	0.77	0.69	0.84	0.78	0.77	0.76	0.78	0.74	0.76
ME	0.75	0.76	0.75	0.86	0.65	0.75	0.75	0.75	0.69	0.79	0.78	0.75	0.75	0.77	0.71	0.76

Still regarding pre-processing tasks, the **stemming** tool was applied and that brought out a slight increase on the performance of all classifiers, with ME being the best again (Table VI). Consequently, at the end it might be included in the final model.

TABLE VI

Tests with Stemming

		Precision					Recall					F-Measure				
	Accuracy	Weighted	Macro	Positive	Neutral	Negative	Weighted	Macro	Positive	Neutral	Negative	Weighted	Macro	Positive	Neutral	Negative
NB	0.76	0.77	0.77	0.80	0.78	0.72	0.76	0.75	0.80	0.63	0.82	0.76	0.76	0.80	0.70	0.77
SVM	0.75	0.77	0.75	0.86	0.66	0.74	0.75	0.76	0.73	0.79	0.75	0.76	0.75	0.79	0.72	0.74
ME	0.77	0.78	0.77	0.86	0.67	0.78	0.77	0.77	0.73	0.82	0.76	0.77	0.77	0.79	0.74	0.77

The last pre-processing test carried out aimed to understand the real importance of the presence or absence of **stop words**. These types of words usually carry little to no meaning, hence it is common practice to exclude them (Saif, He & Alani, 2012), however the generally used lists of stop words can be unfit for SA on Twitter and there has been an effort to develop specific lists for that purpose (Giachanou & Crestani, 2016).

Ultimately, Saif, He and Alani (2012) performed a set of different analysis and concluded that maintaining the stop words could actually generate models that outperform the ones which choose to exclude them. Considering this, a trial was run, this time maintaining stop words instead of removing them, which led to a slight decrease on SVM's and ME's results, but generated encouraging to NB's, which achieved 79% on accuracy and f-measure (Table VII).

TABLE VII

Tests with Stop Words

		Precision					Recall					F-Measure				
	Accuracy	Weighted	Macro	Positive	Neutral	Negative	Weighted	Macro	Positive	Neutral	Negative	Weighted	Macro	Positive	Neutral	Negative
NB	0.79	0.79	0.79	0.80	0.79	0.79	0.79	0.79	0.81	0.72	0.84	0.79	0.79	0.80	0.75	0.81
SVM	0.74	0.75	0.74	0.85	0.64	0.74	0.74	0.74	0.71	0.77	0.75	0.74	0.74	0.77	0.70	0.74
ME	0.74	0.74	0.75	0.87	0.62	0.75	0.74	0.74	0.69	0.79	0.75	0.74	0.74	0.77	0.69	0.75

Encouraged by Wang and Manning's (2012) results, and going back to the baseline, the final feature that was changed to gather more insights about how the classifiers would react, was the use of **bi-grams** instead of the initial uni-grams. However, it also lowered the overall performance and generated the worst performance of every classifier (Table VIII). Hence vectorizing with uni-grams was maintained.

TABLE VIII

Tests with Bi-grams

		Precision					Recall					F-Measure				
	Accuracy	Weighted	Macro	Positive	Neutral	Negative	Weighted	Macro	Positive	Neutral	Negative	Weighted	Macro	Positive	Neutral	Negative
NB	0.63	0.63	0.63	0.62	0.55	0.71	0.63	0.62	0.65	0.53	0.69	0.63	0.63	0.64	0.54	0.70
SVM	0.66	0.70	0.68	0.82	0.51	0.72	0.66	0.67	0.59	0.77	0.65	0.67	0.66	0.69	0.62	0.68
ME	0.66	0.70	0.69	0.82	0.51	0.73	0.66	0.67	0.59	0.77	0.65	0.67	0.66	0.69	0.61	0.69

Stemming and **maintaining stop words** slightly increased some results generated by the trained models, so the last attempt was to gauge how they would both perform together. Although SVM did not achieved its best performance, it increased ME's accuracy and f-measure to 77% and NB's to 80%, which was so far the best result obtained (Table IX).

TABLE IX

Tests with Stemming and Stop Words

		Precision					Recall					F-Measure				
	Accuracy	Weighted	Macro	Positive	Neutral	Negative	Weighted	Macro	Positive	Neutral	Negative	Weighted	Macro	Positive	Neutral	Negative
NB	0.80	0.80	0.80	0.84	0.78	0.77	0.80	0.80	0.83	0.70	0.85	0.80	0.80	0.84	0.74	0.81
SVM	0.73	0.74	0.73	0.85	0.62	0.72	0.73	0.73	0.74	0.72	0.74	0.74	0.73	0.79	0.67	0.73
ME	0.77	0.78	0.77	0.87	0.65	0.79	0.77	0.77	0.77	0.77	0.76	0.77	0.77	0.82	0.70	0.78

Overall, the NB classifier with the added stemming and maintenance of the stop words showed the best results (Table IX) and the system will rely on it to predict the labels of new unseen data. A 10-k cross-validation on this model was also conducted, achieving an accuracy of 49%, with a standard deviation of +/-18%, values that are slightly more encouraging. A confusion matrix for this trained classifier can be found on Appendix 13.

4.4. Final script for public figure's public reputation assessment

Based on the previously results, the model shall use the trained NB classifier (with stemming and stop words included). Now that a classifier has been chosen and trained properly using the dataset that was created, it is time to prepare a script that allows the stream and classification of brand-new tweets. When collecting data for the dataset, two functions were made. That code can be reused with some adjustment in minor details. For instance, now retweets are important, as they show that the same opinion is shared by multiple people. There is also no need to append every new search into a file, so the script will no longer do that. Now, each new search overlaps the previous one and present its results. After that, the collected tweets go through the pre-processing phase, as well as the vectorizing, just as it was done with the training dataset. Subsequently, the formerly trained model is used to predict the categories in which the tweets belong and therefore, it will be possible to assess the overall sentiment associated with the public figure the user intended to research (Appendix 14). This allows to have a general sense of the online reputation of the individual.

The system works as follow: the user is prompted to insert the name of a public figure. It is important to note that the Twitter API limits the number of requests, so some searches might take some time or be interrupted if the limit is exceeded. The number of tweets that the author defined to be fetched per search is 500, but that number could be adjusted in the future if needed. This number presents a good compromise, as it poses both as a decent quantity to have a general assessment of what is being said about an individual and performs quite rapidly. The higher the number of streamed posts, the longer will take the system to capture, process and classify the data. Nonetheless, 500 tweets of an individual may not even be available, thus, once the posts are fetched, a message with how many were gathered is displayed. After that, the created model shows its predictions, indicating how many tweets there are per category and highlighting the category that contains the

majority. On Figure 3 an example of the system's behaviour can be found. Additionally, and in order to help visualize the results, a pie chart was included on the system's output. For that, Matplotlib[11] was used.

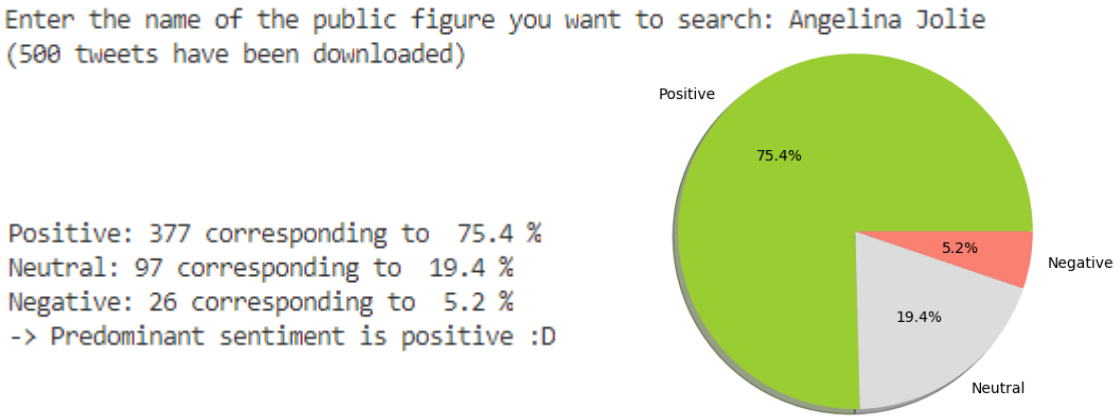


FIGURE 3 - System's Behaviour (Example)

There are, of course, other things that could be interesting to insert in the system's outputs. For instance, one could create word clouds that would show the most used words associated with a public figure search and the user could have an overall notion of what terms are associated with that entity (Appendix 15). The Wordcloud[12] package for Python could be used for that purpose.

5. Conclusions

Big Data has an immensurable number of applications and Sentiment Analysis also has a multitude of options and domains to explore within. This project is only an attempt to join both worlds and explore one of the major social concerns nowadays. It is undeniable that social media and social image, as well as the concern for online reputation, play an important role and that the speed at which data is generated poses a great challenge, even on this domain. Nevertheless, it also comes alongside several opportunities and surely this is where the focus should be and where the future relies.

The elaboration of the project provided a chance to deal with such relevant topics and also be in contact with one of the most used Data Mining frameworks: CRISP-DM. As mentioned, this study intended to investigate and intertwine emergent areas, adding a little more to the discussion around them and expanding the existing knowledge base.

[11] https://matplotlib.org/
[12] https://pypi.org/project/wordcloud/

Nonetheless, there were undoubtedly other routes that could have been taken in the creation of a system of this nature, and that is exactly the challenge that researchers, academics, and corporations face. Much of these realms was left out of the scope of this project, but that also means that there is room for improvement and further investigation. In the end, the models generated did not perform as well as desired, but the outcomes obtained could be explained by the lack of training data, the quality of the dataset, and of course the challenges this area comprises, which is completely understandable. Furthermore, a different set of tools and techniques, as well as other available classifiers could have produced different results. Nonetheless, this project intended to show the problems faced in these areas, as well as its potential, and that was successfully executed. The particularities of human communication and their constant evolution will pose as threats to successful analysis and SA systems, which means that, when it comes to this, continuous research and updates are of utmost importance. The research that lies within the elaboration of this project surely achieves its goal to expand the knowledge base and give a practical example of how available data can be used. Besides, if desired, further investigation can be carried out using this study as a baseline for improvement.

5.1. Contributions, Limitations and Future Work

At the end of the project there are two deliverables that can be used in additional research: a manually annotated dataset and a system able to collect new data, process it and apply a specifically trained model to predict the overall sentiment pertaining to an entity, thus globally assessing its online reputation. Although those are the biggest contributions, it is also possible to state that the present document, through the Literature Review, allows the reader to easily grasp some of the current most trending topics in this area. This project, besides adding to the knowledge base of the domains where it is comprised, can be the starting or comparison point for other studies.

It is also important to mention that some limitations impacted this project. The size of the dataset, that may be deemed as small for machine learning purposes, posed as a crucial factor in the results obtained. With more labelled content, perhaps the results might have been better. Still regarding the dataset, although it was revised by two fellow colleagues in an attempt to reduce bias and attain more consistency, it was manually annotated only by the author which may have induced bias, nonetheless. Errors may be present, and the rules chosen may have not been the most appropriate. Furthermore, and despite the fact that there was a clear effort to have diversity among the content used for training, the

system could have potentially benefited from more inclusivity regarding the entities it included, as well as more content, as some of the labelled posts were fairly similar. Another setback faced were the technical limitations, both from the hardware and applications limits, as wells as the lack of the author's experience with programming. Additionally, and although it could be interesting, due to the time and scope limitations, it was impractical to try more classifiers, or even altering the parameters within the chosen ones. There were other concerns, including the social media dynamics, as there is a multitude of tweets that contain sarcasm and irony, sometimes not easily detected even by humans, which affected the labelling task. Misspelling, slang, invented words, and the fact that context can change the entire meaning of a sentence are major challenges that were not only faced on this project, but also generally on the NLP, SA, BD communities. This, of course, had impact on the performance of the trained model. Finally, the system itself is limited in the sense that, by conducting a sentence-based analysis and assuming that the posts only contain a single entity with solely one predominant sentiment associated, might miss the bigger picture. On some posts, a reference to a certain individual may be found, but the overall sentiments and opinions displayed are not directly connected to it, and the system cannot identify nor understand the differences, thus providing deceitful results when faced with these cases.

Nonetheless, there is room for improvement and for further research on this topic. A first suggestion would be to add more entries to the dataset and see how the presented classifiers perform. Although not used for this specific project, the dataset creation function saves the date on which the tweets were posted. This could be used to perform a deeper analysis and for comparative purposes to assess, as an example, the evolution of a public figure's online reputation. The search input received by the system also gathers the date of the information, but as a new search is prompted, the file is overwritten with the new results. However, this detail could be easily modified if needed. Using other classifiers and a different set of pre-processing techniques could have also led to different results. Some authors also point to the promising results, progresses and potential of the use of Deep Learning and neural networks in NLP tasks (Li, 2017; Otter, Medina & Kalita, 2019). Python and specific libraries and modules designed for this programming language were used, but there are currently other tools available that could have been utilized. Choosing them might have also provided distinct outcomes. Also, instead of the initial assumption that there was only one entity in each tweet and the sentence-based analysis, trying to do an aspect-based analysis while identifying each attribute discussed

would be interesting. Possibly, that more fine-grained analysis would help tackle the cases where more than one entity was present, as well as when there were multiple or mixed opinions displayed. Forthcoming studies could potentially include another set of tests that were not undertaken in this project. Moreover, each and every trial, except the very last one, done to see if including or not a certain technique had any impact, was done individually. Perhaps running them simultaneously might have produced a different outcome. Additionally, one could also try to compare the other SA approaches to the one adopted on this project. A Lexical or Hybrid approach might have given better results, and even if not, it would be curious to explore that. Finally, the produced dataset could be combined with other existing datasets (Saif et al, 2013), much as Cambero (2016) did, and the artefacts the project originated could be interesting to investigate Cross-Domain SA (Yuan et al., 2018). Unfortunately, most of the research done on SA is in English, and this study contributes to that. Nonetheless, interest in adopting strategies for other languages is rising (Medhat, Hassan & Korashy, 214), thus studying this topic and building lexicon and classifiers for other languages should also be considered.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R. (2011). Sentiment Analysis of Twitter Data. In: *LSM '11: Proceedings of the Workshop on Languages in Social Media*. pp.30–38.

Agarwal, R. and Dhar, V. (2014). Editorial - Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. *Information Systems Research*, 25(3), pp.443–448.

Ahmad, M., Aftab, S., Muhammad, S.S. and Ahmad, S. (2017). Machine Learning Techniques for Sentiment Analysis: A Review. *International Journal Of Multidisciplinary Sciences And Engineering*, 8(3), pp.27–32.

Akbani, R., Kwek, S. and Japkowickz, N. (2004). Applying Support Vector Machines to Imbalanced Datasets. In: *Machine Learning: ECML 2004*. pp.39–50.

Aramaki, E., Maskawa, S. and Morita, M. (2011). Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. pp.1568–1576.

Aslam, S. (2020). *Twitter by the Numbers: Stats, Demographics & Fun Facts*. [online] Omnicore. Available at: **https://www.omnicoreagency.com/twitter-statistics/** [Accessed 27 Jul. 2020].

Aula, P. (2010). Social media, reputation risk and ambient publicity management. *Strategy & Leadership*, 38(6), pp.43–49.

Azevedo, A. and Santos, M.F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. In: *Proceedings of IADIS European Conference on Data Mining 2008*. IADIS European Conference on Data Mining 2008.

Balakrishnan, V. and Lloyd-Yemoh, E. (2014). Stemming and lemmatization: A comparison of retrieval performances. In: *Proceedings of SCEI Seoul Conferences*. pp.174–179.

Baumeister, R.F. (1982). Self-esteem, self-presentation, and future Interaction: A dilemma of reputation. *Journal of Personality*, 50(1), pp.29–45.

Bharat, A.V.L.P. and Murthy, K.S. (2016). Exploitation of Sentiment Analysis in Twitter Data utilizing Machine Learning Techniques. *IJRCCT*, 5(12), pp.595–601.

Bhuta, S., Doshi, A., Doshi, U. and Narvekar, M. (2014). A Review of Techniques for Sentiment Analysis Of Twitter Data. In: *2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*. pp.583–591.

Bjurstrom, S. and Plachkinova, M. (2015). Sentiment Analysis Methodology for Social Web Intelligence. In: *Proceedings of the Twenty-first Americas Conference on Information Systems*.

Bollen, J., Mao, H. and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), pp.1–8.

Bravo-Marquez, F., Mendoza, M. and Poblete, B. (2014). Meta-level sentiment models for big social data analysis. *Knowledge-Based Systems*, 69, pp.86–99.

Bronshtein. A. (2017). *Train/Test Split and Cross Validation in Python*. [online] Towards Data Science. Available at: **https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6** [Accessed 14 Oct. 2020].

Brownlee, J. (2018). *A Gentle Introduction to k-fold Cross-Validation*. [online] Machine Learning Mastery. Available at: **https://machinelearningmastery.com/k-fold-cross-validation/** [Accessed 14 Oct. 2020].

Brownlee, J. (2020). *Train-Test Split for Evaluating Machine Learning Algorithms*. [online] Machine Learning Mastery. Available at: **https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/** [Accessed 1 Oct. 2020].

Bulao, J. (2020). *How Much Data Is Created Every Day in 2020?* [online] TechJury. Available at: **https://techjury.net/blog/how-much-data-is-created-every-day/#gref** [Accessed 13 Sep. 2020].

Cambero, A. (2016). *A Comparative Study of Twitter Sentiment Analysis Methods for Live Applications*. Thesis.

Cambria, E. and White, B. (2014). Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. *IEEE Computational Intelligence Magazine*, 9(2), pp.48–57.

Cavazza, N., Pagliaro, S. and Guidetti, M. (2014). Antecedents of Concern for Personal Reputation: The Role of Group Entitativity and Fear of Social Exclusion. *Basic and Applied Social Psychology*, 36(4), pp.365–376.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). *CRISP-DM 1.0 Step-by-stet data mining guide*. SPSS.

Chawla, N.V. (2005). Data Mining For Imbalanced Datasets: An Overview. In: *Data Mining and Knowledge Discobvery Handbook*. Boston, MA: Springer, pp.853–867.

Chen, M., Mao, S. and Liu, Y. (2014). Big Data: A Survey. *Mobile Networks and Applications*, 19(2), pp.171–209.

Chikersal, P., Poria, S., Cambria, E., Gelbukh, A. and Siong, C.E. (2015). Modelling Public Sentiment in Twitter: Using Linguistic Patterns to Enhance Supervised Learning. *Computational Linguistics and Intelligent Text Processing*, pp.49–65.

Chowdhury, G.G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, 37(1), pp.51–89.

Clement, J. (2019). *Twitter: number of active users 2010-2018 | Statista*. [online] Statista. Available at: **https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/** [Accessed 20 Jul. 2020].

Clement, J. (2020). *Leading countries based on number of Twitter users as of July 2020 | Statistic*. [online] Statista. Available at: **https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/** [Accessed 22 Aug. 2020].

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12, pp.2493–2537.

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. In: *1st Workshop on Social Media Analytics (SOMA '10)*. pp.115–122.

Desjardins, J. (2019). *How much data is generated each day?* [online] World Economic Forum. Available at: **https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/** [Accessed 16 Jun. 2020].

Dictionary.com. (n.d.). *How Did Stan Come To Refer To A Super Fan?* [online] Available at: https://www.dictionary.com/e/slang/stan-2/ [Accessed 15 Aug. 2020a].

Dictionary.com. (n.d.). *What Does Cancel Culture Mean? | Pop Culture by Dictionary.com*. [online] Available at: **https://www.dictionary.com/e/pop-culture/cancel-culture/** [Accessed 15 Aug. 2020b].

Fan, W., Wallace, L., Rich, S. and Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM*, 49(9), pp.76–82.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4), pp.82–89.

Feldman, S. (1999). NLP meets the jabberwocky. *Online*, 23, pp.62–72.

Floreddu, P.B., Cabiddu, F. and Evaristo, R. (2014). Inside your social media ring: How to optimize online corporate reputation. *Business Horizons*, 57(6), pp.737–745.

Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G. and Gnanzou, D. (2015). How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, 165, pp.234–246.

Gandomi, A. and Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), pp.137–144.

George, G., Haas, M.R. and Pentland, A. (2014). Big Data and Management. *Academy of Management Journal*, 57(2), pp.321–326.

Gerber, M.S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61, pp.115–125.

Gharehchopogh, F.S. and Khalifelu, Z.A. (2011). Analysis and Evaluation of Unstructured Data: Text Mining versus Natural Language Processing. In: *2011 5th International Conference on Application of Information and Communication Technologies (AICT)*.

Giachanou, A. and Crestani, F. (2016). Like It or Not: A Survey of Twitter Sentiment Analysis Methods. *ACM Computing Surveys*, 49(2), pp.1–41.

Go, A., Bhayani, R. and Huang, L. (2009). *Twitter Sentiment Classification using Distant Supervision*. Stanford: CS224N Project Report, pp.1–6.

Gonçalves, P., Araújo, M., Benevenuto, F. and Cha, M. (2013). Comparing and Combining Sentiment Analysis Methods. In: *Proceedings of the first ACM conference on Online social networks - COSN '13*.

González-Ibáñez, R., Muresan, S. and Wacholder, N. (2011). Identifying Sarcasm in Twitter: A Closer Look. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Short Papers*. Association for Computational Linguistics, pp.581–586.

Grimes, S. (2008). *Unstructured Data and the 80 Percent Rule*. [online] Breakthrough Analysis. Available at: **http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/** [Accessed 19 Mar. 2020].

Haddi, E., Liu, X. and Shi, Y. (2013). The Role of Text Pre-processing in Sentiment Analysis. *Procedia Computer Science*, 17, pp.26–32.

Hammer, H., Yazidi, A., Bai, A. and Engelstad, P. (2015). Building Domain Specific Sentiment Lexicons Combining Information from Many Sentiment Lexicons and a Domain Specific Corpus. In: *IFIP International Conference on Computer Science and its Applications*. pp.205–216.

Hearst, M.A. (1999). Untangling Text Data Mining. In: *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*. pp.3–10.

Hearst, M.A. (2003). *What Is Text Mining?* [online] UC Berkeley School of Information. Available at: **https://people.ischool.berkeley.edu/~hearst/text-mining.html** [Accessed 19 Mar. 2020].

Hirschberg, J. and Manning, C.D. (2015). Advances in natural language processing. *Science*, 349(6245), pp.261–266.

Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F. and Kaymak, U. (2013). Exploiting Emoticons in Sentiment Analysis. In: *SAC '13: Proceedings of the 28th Annual ACM Symposium on Applied Computing*. pp.703–710.

Hossin, M. b. and Sulaiman, M.N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2), pp.01–11.

Hussein, D.M.E.-D.M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences*, 30(4), pp.330–338.

Hutton, J.G., Goodman, M.B., Alexander, J.B. and Genest, C.M. (2001). Reputation management: the new face of corporate public relations? *Public Relations Review*, 27(3), pp.247–261.

IBM. (n.d.). *The Four V's of Big Data*. [online] Available at: **https://www.ibmbigdatahub.com/infographic/four-vs-big-data** [Accessed 9 Mar. 2020].

Igual, L. and Seguí, S. (2017). *Introduction to Data Science: a Python Approach to Concepts, Techniques and Applications*. Cham, Switzerland: Springer.

John Walker, S. (2014). Big Data: A Revolution That Will Transform How We Live, Work, and Think. *International Journal of Advertising*, 33(1), pp.181–183.

Kaplan, A.M. and Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1), pp.59–68.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: *IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*.

Kouloumpis, E., Wilson, T. and Moore, J. (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG! In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. AAAI Press, pp.538–541.

Kuhn, M. and Johnson, K. (2016). *Applied predictive modeling*. New York: Springer.

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety*. [online] *Technical Report, META GROUP*. Available at: **https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.** **[**Accessed 9 Mar. 2020].

Le, B. and Nguyen, H. (2015). Twitter Sentiment Analysis Using Machine Learning Techniques. *Advanced Computational Methods for Knowledge Engineering*, 358, pp.279–289.

Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. *Business Horizons*, 60(3), pp.293–303.

Lever, J., Krzywinski, M. and Altman, N. (2016). Model selection and overfitting. *Nature Methods*, 13(9), pp.703–704.

Li, H. (2017). Deep learning for natural language processing: advantages and challenges. *National Science Review*, 5(1), pp.24–26.

Liddy, E.D. (1998). Enhanced Text Retrieval Using Natural Language Processing. *Bulletin of the American Society for Information Science and Technology*, 24(4), pp.14–16.

Liddy, E.D. (2001). Natural Language Processing. In: *Encyclopedia of Library and Information Science*. New York: Marcel Decker, Inc.

Liu, B. (2012). *Sentiment analysis and opinion mining*. San Rafael: Morgan And Claypool.

Liu, B. and Zhang, L. (2012). A Survey of Opinion Mining and Sentiment Analysis. In: *Mining Text Data*. Boston, MA: Springer, pp.415–463.

Llombart, Ò.R. (2017). *Using Machine Learning Techniques for Sentiment Analysis*. Final Project. Universitat Autònoma de Barcelona (UAB).

Loper, E. and Bird, S. (2002). NLTK: the Natural Language Toolkit. In: *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*.

Madden, M. and Smith, A. (2010). *Reputation Management and Social Media*. [online] Pew Research Center: Internet, Science & Tech. Available at: **https://www.pewresearch.org/internet/2010/05/26/reputation-management-and-social-media/** [Accessed 22 Jul. 2020].

Marr, B. (2019). *What Is Unstructured Data And Why Is It So Important To Businesses? An Easy Explanation For Anyone*. [online] Forbes. Available at: **https://www.forbes.com/sites/bernardmarr/2019/10/16/what-is-unstructured-data-and-why-is-it-so-important-to-businesses-an-easy-explanation-for-anyone/** [Accessed 17 Mar. 2020].

Martínez-Cámara, E., Martín-Valdivia, M.T., Ureña-López, L.A. and Montejo-Ráez, A. (2012). Sentiment analysis in Twitter. *Natural Language Engineering*, 20(1), pp.1–28.

McAfee, A. and Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Business Review*. [online] Oct. Available at: **https://hbr.org/2012/10/big-data-the-management-revolution** [Accessed 17 Mar. 2020].

Medhat, W., Hassan, A. and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), pp.1093–1113.

Misra, A., Sharma, A., Gulia, P. and Bana, A. (2014). Big Data: Challenges and Opportunities. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 4(2), pp.41–42.

Mohammad, S.M. and Turney, P.D. (2012). Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3), pp.436–465.

Murthy, D. (2011). Twitter: Microphone for the masses? *Media, Culture & Society*, 33(5), pp.779–789.

Nadkarni, P.M., Ohno-Machado, L. and Chapman, W.W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5), pp.544–551.

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F. and Stoyanov, V. (2016). SemEval-2016 Task 4: Sentiment Analysis in Twitter. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. pp.1–18.

Ng, E. (2020). No Grand Pronouncements Here...: Reflections on Cancel Culture and Digital Media Participation. *Television & New Media*, 21(6), pp.621–627.

O'Connor, B., Balasubramanyan, R., Routledge, B.R. and Smith, N.A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In: *Proceedings of the Fourth International Conference on Weblogs and Social Media*. pp.122–129.

Oracle. (n.d.). *What Is Big Data? | Oracle*. [online] Available at: **https://www.oracle.com/big-data/what-is-big-data.html#link3** [Accessed 9 Mar. 2020].

Otter, D.W., Medina, J.R. and Kalita, J.K. (2019). A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, pp.1–21.

Owsley, S., Sood, S. and Hammond, K.J. (2006). Domain specific affective classification of documents. In: *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*. pp.181–183.

Pak, A. and Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), pp.1320–1326.

Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), pp.1–135.

Patterson, M. and Bourgoin, R. (2020). *Your Tweet, your voice*. [online] Twitter. Available at: **https://blog.twitter.com/en_us/topics/product/2020/your-tweet-your-voice.html** [Accessed 27 Jul. 2020].

Paul, M.J. and Dredze, M. (2011). You Are What You Tweet: Analyzing Twitter for Public Health. In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. pp.265–272.

Philip Chen, C.L. and Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275, pp.314–347.

Piatetsky, G. (2014). *CRISP-DM, still the top methodology for analytics, data mining, or data science projects*. [online] KDnuggets. Available at: **https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html** [Accessed 13 May 2020].

Pietro, M.D. (2020). *Text Analysis & Feature Engineering with NLP*. [online] Medium. Available at: **https://towardsdatascience.com/text-analysis-feature-engineering-with-nlp-502d6ea9225d** [Accessed 20 Sep. 2020].

Pla, F. and Hurtado, L.-F. (2014). Political Tendency Identification in Twitter using Sentiment Analysis Techniques. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. pp.183–192.

Prabowo, R. and Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2), pp.143–157.

Provost, F. and Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, 1(1), pp.51–59.

Rahman, M.M. and Davis, D.N. (2013). Addressing the Class Imbalance Problem in Medical Datasets. *International Journal of Machine Learning and Computing*, 3(2), pp.224–228.

Ravi, K. and Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89, pp.14–46.

Ray, S. (2015). *8 Proven Ways for boosting the "Accuracy" of a Machine Learning Model*. [online] Analytics Vidhya. Available at: **https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/** [Accessed 14 Oct. 2020].

Reinsel, D., Gantz, J. and Rydning, J. (2018). *The Digitization of the World From Edge to Core*. [online] *Seagate*. Available at: **https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.**

Reyes-Menendez, A., Saura, J. and Alvarez-Alonso, C. (2018). Understanding #WorldEnvironmentDay User Opinions in Twitter: A Topic-Based Sentiment Analysis Approach. *International Journal of Environmental Research and Public Health*, 15(11), p.2537.

Rodríguez-Mazahua, L., Rodríguez-Enríquez, C.-A., Sánchez-Cervantes, J.L., Cervantes, J., García-Alcaraz, J.L. and Alor-Hernández, G. (2015). A general perspective of Big Data: applications, tools, challenges and trends. *The Journal of Supercomputing*, 72(8), pp.3073–3113.

Rogers, A. (2019). *Council Post: The 80% Blind Spot: Are You Ignoring Unstructured Organizational Data?* [online] Forbes. Available at: **https://www.forbes.com/sites/forbestechcouncil/2019/01/29/the-80-blind-spot-are-you-ignoring-unstructured-organizational-data/** [Accessed 9 Mar. 2020].

Rosen, A. (2017). *Tweeting Made Easier*. [online] Twitter. Available at: **https://blog.twitter.com/en_us/topics/product/2017/tweetingmadeeasier.html** [Accessed 21 Jul. 2020].

Rossum, P. (2011). *Big Data Analytics*. TDWI Best Practices Report.

Sagiroglu, S. and Sinanc, D. (2013). Big Data: A Review. In: *2013 International Conference on Collaboration Technologies and Systems (CTS)*. pp.42–47.

Saif, H., Fernandez, M., He, Y. and Alani, H. (2013). Evaluation Datasets for Twitter Sentiment Analysis: a survey and a new dataset, the STS-Gold. In: *1st International Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM 2013)*.

Saif, H., He, Y. and Alani, H. (2012). Semantic Sentiment Analysis of Twitter. In: *Proceedings of the 11th international conference on The Semantic Web - Volume Part I*. pp.508–524.

SAS. (n.d.). *What is Big Data and why it matters*. [online] Available at: **https://www.sas.com/en_us/insights/big-data/what-is-big-data.html** [Accessed 9 Mar. 2020].

Schaffer, C. (1993). Overfitting Avoidance as Bias. *Machine Learning*, 10(2), pp.153–178.

Schneider, C. (2016). *The biggest data challenges that you might not even know you have*. [online] Watson Blog. Available at: **https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/** [Accessed 10 Feb. 2020].

Shulga, D. (2018). *5 Reasons why you should use Cross-Validation in your Data Science Projects*. [online] Towards Data Science. Available at: **https://towardsdatascience.com/5-reasons-why-you-should-use-cross-validation-in-your-data-science-project-8163311a1e79** [Accessed 14 Oct. 2020].

Singh, T. and Kumari, M. (2016). Role of Text Pre-processing in Twitter Sentiment Analysis. *Procedia Computer Science*, 89, pp.549–554.

Sivarajah, U., Kamal, M.M., Irani, Z. and Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, 70, pp.263–286.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), pp.427–437.

Sun, S., Luo, C. and Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion*, 36, pp.10–25.

Symeonidis, S., Effrosynidis, D. and Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications*, 110, pp.298–310.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2), pp.267–307.

Tan, A.-H. (1999). Text Mining: The state of the art and the challenges. In: *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*. pp.65–70.

Thakkar, H. and Patel, D. (2015). *Approaches for Sentiment Analysis on Twitter: A State-of-Art study*.

Tripathy, A., Agrawal, A. and Rath, S.K. (2016). Classification of Sentiment Reviews using N-gram Machine Learning Approach. *Expert Systems with Applications*, 57, pp.117–126.

Tse, Y.K., Zhang, M., Doherty, B., Chappell, P. and Garnett, P. (2016). Insight from the horsemeat scandal. *Industrial Management & Data Systems*, 116(6), pp.1178–1200.

Wang, S. and Manning, C.D. (2012). Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. In: *ACL '12: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*. pp.90–94.

Wirth, R. and Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. In: *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*.

Yaqoob, I., Hashem, I.A.T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N.B. and Vasilakos, A.V. (2016). Big data: From beginning to future. *International Journal of Information Management*, 36(6), pp.1231–1247.

Ye, Q., Zhang, Z. and Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, 36(3), pp.6527–6535.

Yu, H.-F., Huang, F.-L. and Lin, C.-J. (2010). Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1–2), pp.41–75.

Yuan, Z., Wu, S., Wu, F., Liu, J. and Huang, Y. (2018). Domain attention model for multi-domain sentiment classification. *Knowledge-Based Systems*, 155, pp.1–10.

Zazaieri, H., Logli Allison, M., Campos, B., Young, R.C. and Keltner, D. (2018). Content, structure, and dynamics of personal reputation: The role of trust and status potential within social networks. *Group Processes & Intergroup Relations*, 22(7), pp.964–983.

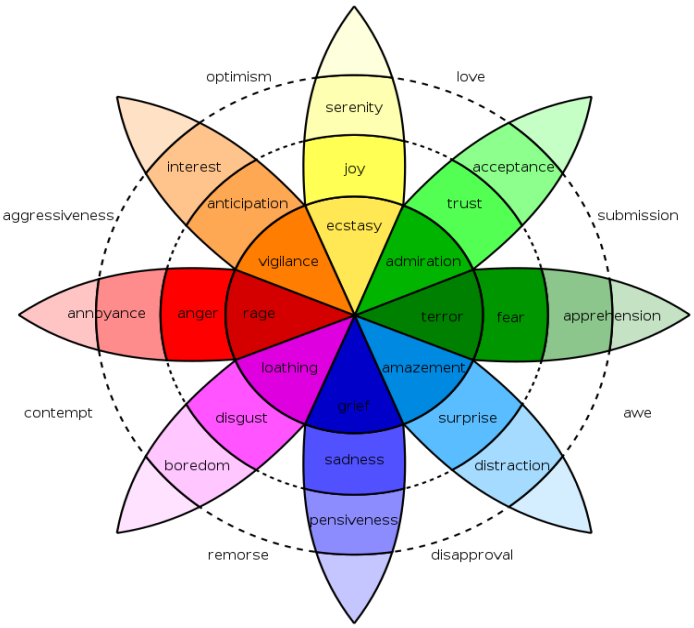
Zhang, H., Gan, W. and Jiang, B. (2014). Machine Learning and Lexicon Based Methods for Sentiment Classification: A Survey. In: *Proceedings of the 11th Web Information System and Application Conference*. pp.262–265.

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M. and Liu, B. (2011). Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis.

Appendices

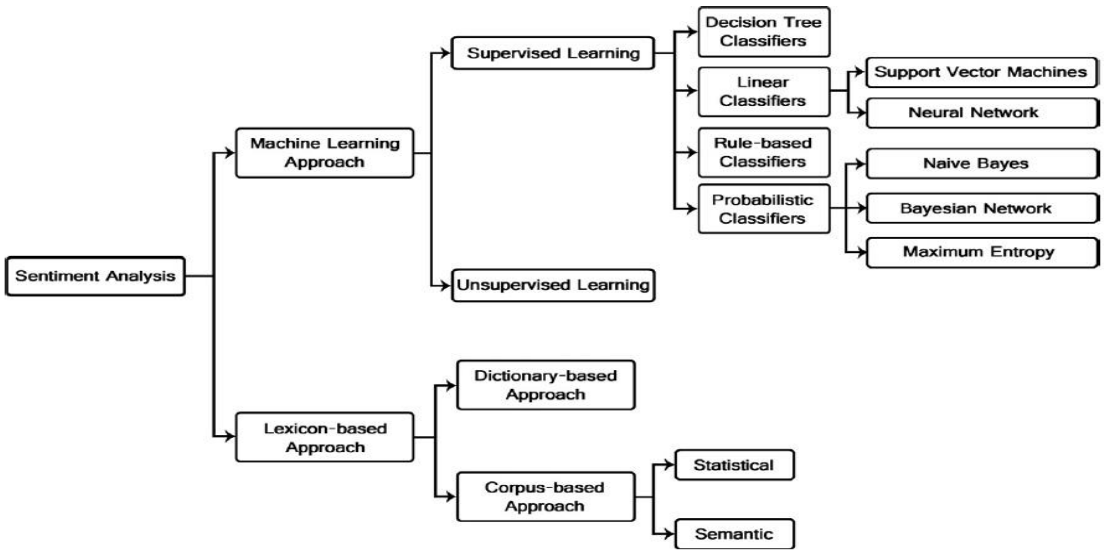
Appendix 1 – Plutchik's Wheel of Emotions

- Plutchik's Wheel of Emotions (from Mohammad &Turney,2012)



Appendix 2 – Sentiment Classification Techniques

- Sentiment Classification Techniques (from Medhat, Hassan & Korashy, 2014)

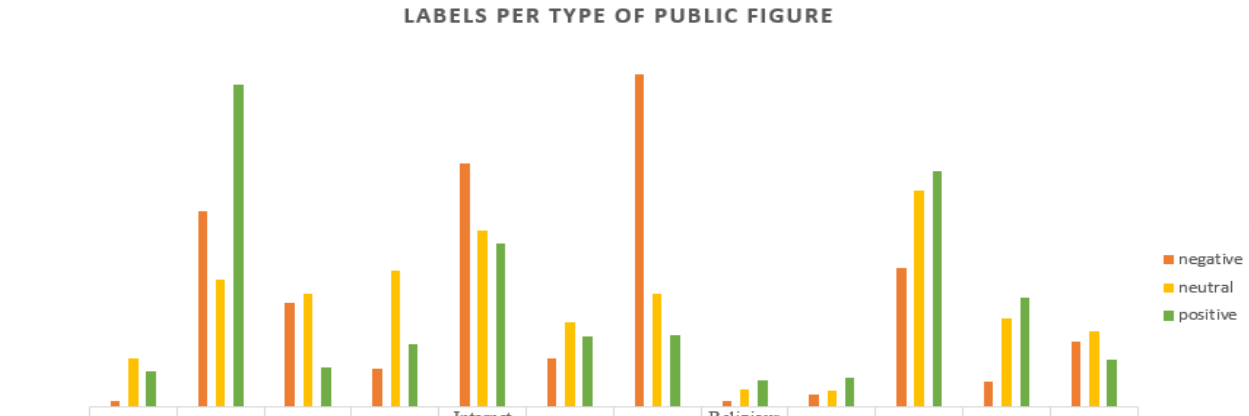


Appendix 3 – Classification Rules

Tweet Classification Guidelines

1. There are only three possible categories for classification: Negative, Positive and Neutral.
2. When the tweet is ambiguous classify it as Neutral.
3. When the tweet expresses mixed sentiments, classify it as Neutral (e.g. "*I don't like X's song but he's funny.*"), unless the sentiment towards the entity is extremely strong and evident (e.g. "*I was not a fan Y's last launch, but the other products are amazing and I absolutely love him!*"), classify as Neutral.
4. When the tweet is objective, classify as it as Neutral (e.g. "I *think Y and W are similar*").
5. Whenever there's a clear context needed to understand the opinion displayed and it is not known, classify the tweet as Neutral.
6. When faced with tweets about products, either it be songs, films or actual items of a brand they own or collaborate with, classify it as if it were the entity itself.
7. When presented with more than one entity, classify the tweet according to the specific public figure that was searched.

Appendix 4 – Labels per Type of Public Figure



	Activist	Actor	Business	Celebrity	Internet Personality	Model	Politician	Religious Figure	Royalty	Singer	Sport	TV
negative	3	109	58	21	135	27	185	3	7	77	14	36
neutral	27	71	63	76	98	47	63	10	9	120	49	42
positive	20	179	22	35	91	39	40	15	16	131	61	26

Appendix 5 –Dataset Snippet

text	weekday	month	day	hour	Public Figure	Label	Type of Public Figure
@Jack_Septic_Eye I wanna be to I'd be Irish Greta thunberg	Tue	Sep	24	17	Greta Thunberg	positive	Activist
If I have a kid like Greta Thunberg one day, I'd be so lucky	Tue	Sep	24	17	Greta Thunberg	positive	Activist
@GretaThunberg You are loved. You are respected. You are a gift to us all. Those who mock you, like... https://t.co/4N0eilSaSF	Tue	Sep	24	17	Greta Thunberg	positive	Activist
Miss Emilia Isobel Euphemia Rose Clarke better win an Emmy this weekend cause if she doesn't I'm coming for y'all @ television aca	Thu	Sep	19	17	Emilia Clarke	positive	Actor
Had a dream I met bill gates	Thu	Sep	19	0	Bill Gates	neutral	Business
soo bill gates and i have the same favorite food..	Thu	Sep	19	0	Bill Gates	neutral	Business
Bill and Melinda Gates weren't born billionaires	Thu	Sep	19	0	Bill Gates	neutral	Business
I am a person who never compromises my morals and dignity. I cannot stand Jeff Bezos, therefore I have never in my... https://t.co/DW	Tue	Apr	14	14	Jeff Bezos	negative	Business
And besides, I don't like Jeff Bezos.	Tue	Apr	14	14	Jeff Bezos	negative	Business
Stop buying things on amazon Jeff Bezos is evil	Tue	Apr	14	13	Jeff Bezos	negative	Business
Can someone put a sock in Justin Trudeau mouth to shut him up. He's a joker.	Tue	Sep	24	17	Justin Trudeau	negative	Politician
@DrJacobsRad Justin Trudeau is a liar. He is not being open.	Tue	Sep	24	17	Justin Trudeau	negative	Politician

Appendix 6 – Count of tweets per Public Figure

Type of Public Figure	Count
Activist	50
Actor	359
Business	143
Celebrity	132
Internet Personality	324
Model	113
Politician	288
Religious Figure	28
Royalty	32
Singer	328
Sport	124
TV	104
Total Count	2025

Appendix 7 – Public Figure per Type of Public Figure

	Tweet Count			
Type of Public Figure	**negative**	**neutral**	**positive**	**Total**
Activist	**3**	**27**	**20**	**50**
Greta Thunberg	3	20	18	41
Malala Yousafzai	0	7	2	9
Actor	**109**	**71**	**179**	**359**
Amber Heard	33	0	0	33
Angelina Jolie	3	15	36	54
Cynthia Nixon	0	0	2	2
Daisy Edgar Jones	0	2	9	11
Dwayne Johnson	0	6	2	8
Emilia Clarke	0	2	37	39
Evan Rachel Wood	23	0	0	23
Ezra Miller	7	3	0	10
Felicity Huffman	11	4	0	15
Jennifer Aniston	3	5	21	29
Joaquin Phoenix	0	0	1	1
Johnny Depp	2	0	9	11
Kevin Spacey	8	0	1	9
Lena Headey	0	1	9	10
Ricky Gervais	1	0	3	4
Robert Pattinson	0	17	9	26
Scarlett Johansson	0	4	0	4
Scott Baio	2	0	3	5
Tara Reid	0	1	0	1
Timothée Chalamet	2	8	20	30
Vanessa Hudgens	14	0	2	16
Will Smith	0	3	9	12
Zendaya	0	0	6	6

	Tweet Count			
Type of Public Figure	**negative**	**neutral**	**positive**	**Total**
Business	**58**	**63**	**22**	**143**
Bill Gates	8	15	3	26
Elon Musk	20	28	17	65
Elon Musk #sarcasm	1	0	0	1
Elon Musk sarcasm	0	0	1	1
George Soros	10	2	0	12
Jeff Bezos	10	4	0	14
Mark Zuckerberg	9	14	1	24
Celebrity	**21**	**76**	**35**	**132**
Chrissy Teigen	4	1	11	16
Kim Kardashian		6	1	7
Kylie Jenner	15	47	17	79
Oprah Winfrey	0	11	4	15
Paris Hilton	1	11	2	14
Vanessa Hudgens	1	0	0	1
Internet Personality	**135**	**98**	**91**	**324**
Austin McBroom	42	8	0	50
Jaclyn Hill	18	10	4	32
Jake Paul	7	0	2	9
James Charles	31	22	6	59
Jeffree Star	22	26	20	68
Nikkie	1	18	29	48
PewDiePie	5	7	1	13
Shane Dawson	6	5	26	37
Tati Westbrook	3	2	3	8

	Tweet Count			
Type of Public Figure	negative	neutral	positive	Total
Religious Figure	**3**	**10**	**15**	**28**
Pope Francis	3	10	15	28
Royalty	**7**	**9**	**16**	**32**
Meghan Markle	3	3	9	15
Prince Harry	0	6	6	12
Prince William	4	0	1	5
Singer	**77**	**120**	**131**	**328**
Aaron Carter	21	10	3	34
Adele		11	16	27
Ariana Grande	5	8	13	26
Beyoncé	6	19	26	51
Britney Spears	0	8	3	11
Dermot Kennedy	0	1	6	7
Drake	0	0	3	3
Harry Styles	0	0	2	2
Jennifer Lopez	1	4	14	19
Justin Bieber	30	31	20	81
Kanye West	6	0	0	6
Lewis Capaldi	0	4	4	8
Michael Jackson	1	0	4	5
Miley Cyrus	1	13	2	16
Sam Smith	3	3	9	15
Taylor Swift	3	8	6	17

	Tweet Count			
Type of Public Figure	negative	neutral	positive	Total
Sport	**14**	**49**	**61**	**124**
Cristiano Ronaldo	4	14	24	42
Gerard Piqué	2	8	7	17
João Félix	3	2	3	8
Lebron James	0	1	3	4
Mbappé	0	7	5	12
Messi	0	2	8	10
Michael Jordan	0	0	2	2
Nadal	3	14	3	20
Serena Williams	2	1	6	9
TV	**36**	**42**	**26**	**104**
Ellen DeGeneres	16	10	1	27
James Corden	0	3	8	11
Kevin Hart	2	1	1	4
Matt Preston	0	0	1	1
Oprah #sarcasm	1	0	0	1
Oprah sarcasm	1	0	0	1
Piers Morgan	9	5	2	16
RuPaul	0	10	5	15
Simon Cowell	4	2	2	8
Steve Harvey	3	11	6	20

	Tweet Count			
Type of Public Figure	negative	neutral	positive	Total
Model	**27**	**47**	**39**	**113**
Bella Hadid	1	1	11	13
Cara Delevingne	1	5	0	6
Gigi Hadid	1	11	9	21
Gisele Bündchen	0	1	3	4
Heidi Klum	4	6	4	14
Naomi Campbell	3	1	4	8
Sara Sampaio	0	4	1	5
Tyra Banks	17	18	7	42
Politician	**185**	**63**	**40**	**288**
Bernie #sarcasm	2	0	0	2
Bernie Sanders	7	8	30	45
Biden #sarcasm	2	1		3
Bolsonaro	10	0	0	10
Boris Johnson	22	12	1	35
Dominic Cummings	8	2	0	10
Donald Trump	44	11	1	56
Jeremy Corbyn	2	1	3	6
Joe Biden	16	9	1	26
Justin Trudeau	20	2	1	23
Kanye #sarcasm	1	0	0	1
Kylie sarcasm	0	1	0	1
Michelle Obama	0	2	1	3
Mike Pence	0	5	0	5
Obama	1	0	2	3
Obama #sarcasm	0	1	0	1
Robert Mueller	6	4	0	10
Theresa May	4	3	0	7
Trump	3	0	0	3
Trump #sarcasm	37	1	0	38

Appendix 8 – NLTK's English Stop Words

['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]

Appendix 9 – Confusion Matrix

- Confusion Matrix (Adapted from Chawla, 2005)

	Predicted Negatives	**Predicted Positives**
Actual Negatives	TN	FP
Actual Positives	FN	TP

Appendix 10 – NB's Confusion Matrix

	NB		
	Predicted Positives	**Predicted Neutrals**	**Predicted Negatives**
Actual Positives	59	8	11
Actual Neutrals	10	38	9
Actual Negatives	5	6	57

Appendix 11 – SVM's Confusion Matrix

	SVM		
	Predicted Positives	**Predicted Neutrals**	**Predicted Negatives**
Actual Positives	53	14	11
Actual Neutrals	4	46	7
Actual Negatives	4	11	53

Appendix 12 – ME's Confusion Matrix

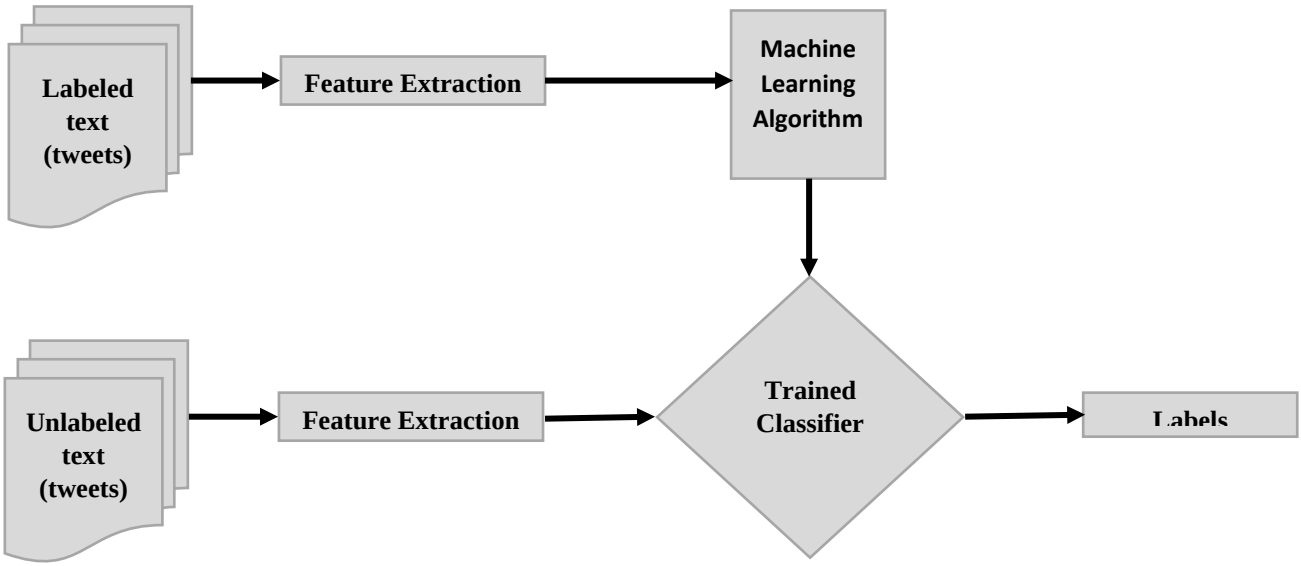
	ME		
	Predicted Positives	**Predicted Neutrals**	**Predicted Negatives**
Actual Positives	54	13	11
Actual Neutrals	4	48	5
Actual Negatives	4	12	52

Appendix 13 – Final Model's Confusion Matrix

	NB with Stemming and Stop Words		
	Predicted Positives	**Predicted Neutrals**	**Predicted Negatives**
Actual Positives	65	5	8
Actual Neutrals	8	40	9
Actual Negatives	4	6	58

Appendix 14 – Scheme of SA's System Process

- (Adapted from Giachanou & Crestani, 2016)



Appendix 15 – Word Clouds