

CITY UNIVERSITY LONDON
MSc IN BUSINESS SYSTEMS ANALYSIS AND DESIGN
PROJECT REPORT
2013

Sentiment Analysis using Software

Emphasis on healthcare information services

Despo Georgiou
Supervised by: Dr. Andrew MacFarlane
10th January 2014

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

Signed:

ABSTRACT

The aim of the project was to assess the performance and accuracy of systems in extracting sentiment, in order to validate their use in the field. Whether the configuration of the algorithm used improves accuracy was examined through a comparison of commercial and non-commercial tools. Focus was given on the domain of healthcare information services. Four systems were used to analyse sentiment from 137 responses resulted from an online survey. The data were provided by UX Labs. TheySay and Semantria represented the commercial group, while WEKA and Google Prediction API represented the non-commercial group. Different approaches were followed for each tool to determine the polarity of each response. A neutral class was also included in the analysis. Negation, punctuation usage and capital letters as well as categories and POS were some of the features provided by the tools. The accuracy of the systems was compared mainly by calculating the percentage of correctly classified responses. The performance of WEKA was found to be the most suitable for the healthcare domain, by achieving 82.35% accuracy. Next was Google Prediction API which led to the assumption that non-commercial tools are preferred. Due to the variety of features between the tools, recommendations for the use of each tool were made according to the requirements of the user. Single-sentence responses were isolated and tested for more precise results based on the assumption that they are more clearly associated with a single sentiment polarity. The findings provide evidence that verify that assumption. Further work can be done to establish a relationship. Exploring more tools and using more data will allow the establishment of generalisation.

Keywords: sentiment analysis, machine learning, software, classification, healthcare

TABLE OF CONTENTS

Abstract	2
List of Figures	6
List of Tables.....	8
1 Introduction	12
1.1 The Language	12
1.2 Text Mining, Sentiment Analysis and Applications	12
1.3 Aims and Objectives.....	16
1.4 Project Outline	17
2 Scope	18
3 Academic Context	19
3.1 Sentiment Analysis Problem	19
3.2 Supervised and Unsupervised Learning	32
3.2.1 Evaluation Methods for Supervised Machine Learning Classification	34
3.3 Sentiment Classification Problem and Techniques	38
3.4 Computerised Tools for Sentiment Analysis	44
4 Methods	47

4.1	Data	48
4.2	Project Procedure	49
4.2.1	Commercial Tools.....	49
4.3	Non-Commercial Tools.....	54
4.3.1	Google Prediction API	55
4.3.2	WEKA.....	58
4.4	Evaluation Techniques	61
5	Results.....	63
5.1	Human Observations.....	63
5.2	Commercial Tools.....	65
5.2.1	Semantria	65
5.2.2	TheySay	74
5.2.3	Comparison between Commercial Tools	79
5.3	Non-Commercial Tools.....	85
5.3.1	Google Prediction API	85
5.3.2	WEKA.....	94
5.3.3	Comparison between Non-Commercial Tools	103

6	Evaluation of Systems with respect to the Healthcare Information Service Domain.....	105
7	Discussion.....	109
8	Evaluation and Conclusions	117
9	Reflections.....	119
	Glossary.....	121
	References.....	123
	Appendices.....	A.1
A	Project Proposal	A.1
B	Semantria: Conversion Results	B.1
C	Data Labelling.....	C.2
D	WEKA Simple CLI	D.1
E	Semantria: Collection Analysis	E.1
F	Commercial Tools: Precision, Recall and F-measure.....	F.1
F.1	Semantria	F.1
F.2	TheySay	F.4
G	Google Prediction API: Precision, Recall and F-measure	G.1
H	Summary of WEKA results	H.1

LIST OF FIGURES

Figure 3.1: <i>Supervised classification</i> (Bird, Klein & Loper, 2009, p.222).....	33
Figure 3.2: Classifier's performance (Pang, Lee & Vaithyanathan, 2002, p.83).....	34
Figure 3.3: A 10-fold Cross Validation.....	35
Figure 4.1: Semantria's control panel in Microsoft Excel.	50
Figure 4.2: The three steps needed to create a project in Semantria.	51
Figure 4.3: TheySay's console box.	51
Figure 5.1: Quantitative analysis of the data.	64
Figure 5.2: Semantria's categorisation results.....	65
Figure 5.3: Examples of entities identified by Semantria.	66
Figure 5.4: Classification results of single-sentence responses by Semantria.....	71
Figure 5.5: TheySay's classification results.	74
Figure 5.6: Results for response " <i>cant find guidance...</i> " to examine the case sensitivity of TheySay.	76
Figure 5.7: Observing errors during verification/clarification.	77
Figure 5.8: Classification results of single-sentence responses by TheySay.....	78
Figure 5.9: Single-sentence responses classified by Semantria and TheySay.....	84

Figure 5.10: Total number of positive, neutral and negative responses used for training the model via Google Prediction API. They represent the 75% of the dataset.....	85
Figure 5.11: Classification results using the whole dataset. Google Prediction API was tested on the 25% of unseen responses in addition to the already trained responses.....	86
Figure 5.12: Number of single- sentence responses used for training purposes for Google Prediction API.....	91
Figure 5.13: Results of single-sentence classification analysis by Google Prediction API. The 75% of single sentences was used for training and 25% was used for testing.....	91
Figure 5.14: The three classes used for the training along with the number of responses contained in each of them. They represent the 75% of each class.	94
Figure 5.15: Number of single- sentence responses used for training purposes for WEKA.....	100
Figure 6.1: Comparison of precision among the four tools.	108
Figure 6.2: Comparison of recall among the four tools.	108

LIST OF TABLES

Table 3.1: Types of Comparatives (Jindal & Liu, 2006b).	24
Table 3.2: "Yeah right" in different speech acts (Tepperman, Traum & Narayanan, 2006).	28
Table 3.3: <i>"Percentage of sentences with some main conditional connectives"</i> (Narayanan, Liu & Choudhary, 2006).	30
Table 3.4: The three classification methods used in the Narayanan, Liu & Choudhary (2009) paper to categorise conditional sentences in any of the three classes: positive, negative or neutral.	31
Table 3.5: Example of Confusion Matrix	36
Table 3.6: Table of confusion for Peaches	37
Table 3.7: Sentiment polarity differences due to change of domain (Bollegata, Weir & Carroll, 2013, p.1721)	40
Table 3.8: Patterns of POS tags for extracting two word phrases (Turney 2002, in Liu, 2012, p.35).	44
Table 4.1: Survey questions.	48
Table 4.2: The available features of Semantria.	52
Table 4.3: The available features of TheySay.	53
Table 4.4: Scores assigned to each class.	56

Table 4.5: The available features of WEKA.	58
Table 4.6: The process needed to achieve a common output format for comparison purposes.	62
Table 5.1: Average Scores associated with the quantitative data.	64
Table 5.2: Examples of classification errors by Semantria.	67
Table 5.3: Comparison between keywords identified by the human and Semantria.	68
Table 5.4: Default categories.	69
Table 5.5: Custom categories.	69
Table 5.6: Results of the analysis based on the custom profile created.	70
Table 5.7: Different examples of neutral responses using Semantria.	73
Table 5.8: Examination of TheySay's features.	75
Table 5.9: Errors found in single-sentence responses when analysed by TheySay.	79
Table 5.10: Comparison between Commercial Tools	80
Table 5.11: Correctly Classified Responses - Commercial Tools	81
Table 5.12: Confusion Matrix for Semantria.	81
Table 5.13: Assessment of the reliability of the algorithm used by Semantria.	82
Table 5.14: Confusion Matrix for TheySay.	83

Table 5.15: Assessment of the reliability of the algorithm used by TheySay.	83
Table 5.16: Correctly classified single-sentence responses – Commercial Tools.	84
Table 5.17: Examination of the different training samples used for sentiment analysis using Google Prediction API.	86
Table 5.18: Cross validation training sets.	87
Table 5.19: Details of ‘Train 1’ model.	87
Table 5.20: Details of ‘Train 2’ model.	88
Table 5.21: Details of ‘Train 3’ model.	88
Table 5.22: Details of ‘Train 4’ model.	88
Table 5.23: Statistical evaluation of the model built in Google Prediction API (based on cross validation).	89
Table 5.24: Confusion Matrix for Google Prediction API.	89
Table 5.25: Assessment of the reliability of the model.	90
Table 5.26: Statistical evaluation of Google Prediction API’s results when tested on unseen data (testing set).	92
Table 5.27: Details of the four folds for single-sentence responses analysis	92
Table 5.28: Details of ‘Single Train 1’ model.....	92
Table 5.29: Details of ‘Single Train 2’ model.....	93

Table 5.30: Details of 'Single Train 3' model.....	93
Table 5.31: Details of 'Single Train 4' model.....	93
Table 5.32: Attribute and resampling experimentation	97
Table 5.33: Assessment of final model's performance using statistical measures.....	98
Table 5.34: Confusion Matrix for WEKA.	99
Table 5.35: Assessment of final model's reliability.....	99
Table 5.36: Correctly classified instances using Naïve Bayes classifier with 4-fold cross validation: attribute selection & resampling.	101
Table 5.37: Correctly classified instances using Naïve Bayes classifier with 4-fold cross validation: resampling & attribute selection.	101
Table 5.38: Experimentation with different attribute combinations along with resampling...	102
Table 5.39: Comparison between non-commercial tools.....	103
Table 5.40: Accuracy of the classifier for non-commercial tools.....	104
Table 5.41: Comparison between non-commercial tools using statistical measures.....	104
Table 6.1: Comparison among all four tools for sentiment classification.	106
Table 6.2: Statistical comparison of the four tools.	107

1 INTRODUCTION

1.1 The Language

Language is a communication tool used by people all over the world, either verbally or in writing. As the world progressed and technology was embedded in our everyday life, our main communication tool had to be adjusted to these changes. Despite the psychological need to understand how the human language is used, Natural Language Processing (NLP) has been widely studied to gain new insights in formulating algorithmic language representations. The goal is to promote artificial language by providing an accurate replication with the aid of machines.

Natural Language is a difficult concept to transfer to an artificial intelligence system. This is because humans are characterised by their individuality and their different perception skills. This involves opinions, sentiments, evaluations, attitudes and emotions, which can affect each person's decision-making process (Liu, 2012). Research emphasises on discovering text mining techniques, as well as possible combinations of these techniques that can help achieve a progress in the NLP field. Individuality can only be expressed in an unstructured way and with the help of text mining, hidden knowledge can be uncovered.

1.2 Text Mining, Sentiment Analysis and Applications

Analysing text can be more useful than one might think. It is well known that knowledge comes with information and more knowledge is always an advantage (Mack et. al., 2004). While the importance of discrete data was never underestimated, it must be pointed out that the combination of structured and unstructured data can provide new insights in finding patterns and trends (USA. IBM Corporation. 2012).

Sentiment analysis is concerned with investigation of opinions and thoughts. It is only reasonable that such information can be vital to the decision-making process (Pang & Lee, 2008). It can create enormous opportunities for knowledge, which will in turn influence productivity, efficiency and so on (QuestionPro. 2011; USA. IBM Corporation. 2012; Cameron, Bhagwan & Sheth, 2012). A number of methods for analysing sentiment exist that involve both artificial intelligence and the human element and they are widely used in a variety of fields. Examples of some methods and their application on specific fields can be seen below:

- **Surveys:** Explore and build models (USA. IBM Corporation. 2012), as well as analyse surveys that contain open-ended questions (QuestionPro. 2011).
- **Consumer Feedback:** Analyse reviews with the aim to understand and satisfy customers. The reviews are used to evaluate a product (Pang & Lee, 2008). Moreover, an effort to understand consumer opinions about products and services is made. New innovation opportunities and competitive advantage (Keke et. al., 2008) would be eventually created as a result.
- **Social Media:** Detect sensitive content, inappropriate for ads (Pang & Lee, 2008) is one example. Furthermore, social media is deep in our lives and people tend to use such tools to express themselves. This could be found useful for all the fields that have an interest in sentiment analysis.
- **Business and Government:** Provide clear, immediate and actionable insights into current performance and the ability to predict future activities. Also, text mining can provide consistent and accurate information that can be trusted to make more complete decisions in order to improve business performance (USA. IBM Corporation. 2012). Identifying key strategies or market shifts using a trend analysis can also be useful, as well as fraud detection capabilities. Governments can monitor the sources for increases in hostile or negative communication (Pang & Lee, 2008).

- **Politics:** Understand voters' way of thinking and in return provide a more clear view to voters about the politicians' position, as well as enhance the quality of information they have access to (Pang & Lee, 2008).
- **Health (Mack et. al., 2004):** Support problem-solving in life science by managing and analysing biomedical text. The aim is to treat diseases and enhance the health of humans. Analysing treatment effects as reported in various forums is one way to go. It has been found that drug-discovery can become faster with the aid of sentiment analysis.

However, the challenge in this field is to analyse people's beliefs, thoughts and opinions. Societies are based around people and individuality is a key aspect to be studied. The most common way to retrieve public opinions is questionnaires. It is up to the researcher to decide, but it had been found that a combination of closed-ended and open-ended questions give a more accurate idea of the interviewee's replies (Rogers, Sharp and Preece, 2011). Closed-ended questions are usually used as an introduction to the field questioned and later, open-ended questions follow to give the research a 'chunk' of useful unstructured data to analyse (QuestionPro. 2011). The data used for this project are retrieved from an online survey containing both quantitative (closed-ended) and qualitative (open-ended) responses, as we will see later. They were provided by UX Labs.

Sentiment analysis can also aid business and governments to understand their public and be better prepared for the future. Nowadays, where social media are ruling the world, advanced linguistic technologies and Natural Language Processing (NLP) can *"rapidly process a large variety of unstructured data and extract and organise the key concepts"* (USA. IBM Corporation. 2012, p.iii). They are a speed and cost effective means of finding the meaning in text, with a higher degree of accuracy (USA. IBM Corporation. 2012).

Systems aid in trend analysis and predictive modelling, by plotting promotional activities against historic map of on-going customer responses, for example (USA. SAS Institute Inc. 2010). This is done by converting source data to a standard format and then identifying candidate terms necessary for the creation of classes. Experimentation with different techniques is often inevitable to see which procedure best suits which company (USA. IBM Corporation. 2012). Technology can provide a competitive advantage (Spangler et. al., 2010) to all business in *“attracting, retaining and growing customers, while reducing fraud and mitigating risks”* (USA. IBM Corporation. 2012, p.iii), if used wisely, by trying to understand the customer’s behaviour (USA. IBM Corporation. 2012). However, the IBM Corporation (2012) sets an important question: *“Can machines understand human communication?”*

The answer lies in the combination of the two; humans and machines working closely together. Computers can calculate whether a given phrase is positive or negative with a certain degree of accuracy and humans can review the low-confidence results and advice the machine how to grade them. Pang, Lee & Vaithyanathan (2002) provided evidence that could support the collaboration between machines and humans.

Due to the complexity of biomedical texts, all the abbreviations and the frequent misspellings, health information services is one of the most difficult and challenging fields for text analytics to be applied. However, it is also one of the most rewarding fields to work in. Text analytics can provide *“new insights that could impact diagnosis, treatment and overall patient care”* (Cameron et. al., 2012, p.240). There are systems that combine a patient’s symptoms, medication and history record, along with their family history, and a confident diagnosis (USA. IBM Corporation. 2011).

To be more specific, Byrd et. al. (2013) developed a system that *“accurately identifies and labels affirmations and denials of Framingham diagnostic criteria (a symptom of heart failure)*

in primary care clinical notes and may help in the attempt to improve the early detection of a heart failure” (Byrd et. al., 2013, p.1).

1.3 Aims and Objectives

Is it possible to provide an accurate replication of language using machines? Promoting artificial language resulted in some commercial tools that will be used for this project. We examine whether machines can recognise feelings and hence are able to assist in our everyday life.

As far as we are aware of, text in the healthcare domain has not been investigated from a software perspective. Healthcare information services require extra work due to the complexity associated. But, how much more work is essentially needed? The intention is to show whether sentiment can be accurately extracted by systems. Existing software will be used and survey responses from a website evaluation will represent our dataset. The classification polarity assigned to the responses will also be investigated.

Due to the continuous increase of commercial tools for sentiment analysis in the market, they were our first resource. We wanted to examine whether they can cope with our dataset and evaluate their capabilities as far as the healthcare service is concerned. Further examination was undertaken using some customisable tools, referred to as non-commercial tools in this report. The latter are usually more accurate due to the variety of configurations that can be done to the algorithm used for classification. It was believed that this could add value to our project and help in making a more informed conclusion.

It is expected that commercial products should be at least equally good as non-commercial, since they often come with a price. We essentially investigate the use of four tools (two of each group) in the healthcare domain and whether commercial products are indeed worth

paying for. Even though, the demo version was used, no implications are anticipated during the project. The performance of the commercial tools should remain unaffected. On the other hand, we are comparing two very different things. Non-commercial products are trained on the specific area of this project, which makes comparison unfair. Nevertheless, human errors are inevitable with manual customisation.

1.4 Project Outline

This project report will refer to the related literature existing for sentiment analysis (Section 3). Different techniques and methods examined by researchers will be identified and explained. At the end of Section 3, it is expected that the reader will have an informed view of the sentiment analysis problem and complexities faced by researchers that would help understand the procedures followed in this project.

The project is focused on a healthcare information service website aimed for professionals and the general public. Unfortunately, the name of the website cannot be disclosed in this report due to confidentiality. After a recent update to the website, an online survey was conducted that resulted to 167 responses. These were provided by UX Labs Ltd for the purpose of this project. The aim was to find an appropriate tool available on the market that could accurately extract the sentiment analysis involved in the responses available.

The methods used to achieve accurate extraction on each tool can be found in Section 4. Moreover, the results of this study are outlined (Section 5) along with comparisons of the four chosen tools (Section 6). Later a discussion of the results is available where the results are explained in more detail (Section 7). In Section 8, a summary of the project is provided, to refresh the reader's memory of what was included in this report and the conclusions of this study are outlined.

2 SCOPE

Sentiment analysis is a very broad field. It is impossible to consider every aspect in the timeframe of an MSc project. Hence, our attention was diverted towards software analysis, focusing on the healthcare domain. Commercial and non-commercial products were the resources used for this project. Due to the time limitations, no more than two tools per group could be examined. Furthermore, certain features associated with complex tools were omitted from the analysis. In general, an initial investigation of software applications for sentiment analysis will be made. However, it is expected that the results found cannot give definite answers regarding healthcare data since only four tools are being evaluated and due to the small sample size. Further research is anticipated beyond this report, where more software can be evaluated for the healthcare domain in particular, but also for other fields other than healthcare.

3 ACADEMIC CONTEXT

A lot of research has been done in the field of Natural Language Processing (NLP) and linguistics. Since NLP is a very broad field, most research papers emphasize on one level of analysis – some examine documents or sentences, while others examine the aspects and words related to the documents and sentences. The aim is to identify possible patterns that will help the classification problem and hence enable better understanding of the use of human language.

Analysing text can be more useful than one might think. Its uses and contributions are endless, as already explained in the previous section. Knowledge comes with information; text analytics, as well as data analytics, enhance the information available to us (Mack et. al., 2004). Text mining tasks, such as text categorization, sentiment analysis and text clustering are widely used for text analysis. In this report we will focus on sentiment analysis, which is concerned with the investigation of the opinions, thoughts, attitudes and sentiments expressed by humans. It aims to determine the thoughts of the speaker or writer regarding a specific subject or topic (Li & Dash Wu, 2009) or simply identify the overall polarity of a document (classification). In other words, it extracts and retrieves information from unstructured raw data, which are usually presented in the form of judgement or evaluation and reflect any kind of emotion.

3.1 Sentiment Analysis Problem

According to Liu (2012), an opinion is defined by a quintuple,

$$(e_i, a_{ij}, s_{ijkl}, h_k, t_l) \quad \text{Equation 1}$$

where e_i is the name of an entity, a_{ij} is an aspect of e_i , s_{ijkl} is the sentiment on aspect a_{ij} of entity e_i , h_k is the opinion holder, and t_l is the time when the opinion is expressed by h_k . The sentiment s_{ijkl} is positive, negative, or neutral, or expressed with different strength levels.

The quintuple is used to transform unstructured text to structured data. It is important to note that all five components are essential to avoid any errors, since the definition above represents a hierarchy of parts that form an opinion. It has been found that simplification can result in information loss using the quintuple representation. However, the definition is considered sufficient for most applications. It provides fundamental information and the basis of the qualitative and quantitative summaries (Liu, 2012).

Based on this quintuple, Liu (2012) identified a sequence of six tasks for sentiment analysis. Taking a document, one must first extract and categorize the entities present and then extract all aspects of these entities. This is followed by noting the time when the opinion was expressed. In the end, each aspect's polarity is classified as positive, negative or neutral. All this information is used to form the quintuple. It is important to note that the definition of opinion provided by Liu (2012) assumes that only one aspect is responsible for the polarity. This type of opinion is called *regular* opinion.

The above method may be misleading as to the difficulties faced when analysing sentiment. Nevertheless, it is still considered a feasible approach towards sentiment analysis. Due to the complexity of the natural language, many have tried to find ways to make sense of it in a pattern-wise manner. In this report, we will examine most of these approaches and theories.

A good starting point is to consider subjectivity, which is closely related with forming an opinion, as we will see later. There are subjective and objective texts that can express opinion either explicitly or implicitly (based on the aspect expressions, i.e. words that appear in the

text). By definition, subjective texts are expected to explicitly express feelings and beliefs that construct an opinion (Benamara et al., 2011), while objective texts simply state facts (also known as factual statements). For example:

- **Subjective Sentences:** “I love this movie – I could literally watch it all the time.”, “The food was delicious and the service was excellent.”
- **Objective Sentences:** “He made a chocolate cake.”, “Today I visited my grandparents and later on I met with a friend.”

Due to the clear link between subjective sentences and opinions, researchers tend to ignore objectivity since it is believed that there is no significant loss of information. However, there are cases where the roles are reversed and subjective sentences do not express sentiment (Liu, 2012). A better understanding can be gained with the following cases:

- “*I think that he went home*” (Liu, 2012, p.27) is a subjective sentence, yet it expresses no sentiment
- “*The earphone broke in two days*” (Liu, 2012, p.27) states a fact about the earphone and while no sentiment is explicitly expressed, it clearly implies a negative sentiment about the earphone

Consequently, it is believed that an appropriate starting point is to distinguish opinionated and non-opinionated sentences, regardless of whether they are subjective or objective (Liu, 2012).

Benamara et al. (2011) provided evidence to justify that sentiment can exist in both subjective and objective sentences by examining four different combinations; two subjective, named *S* (subjective and evaluative) and *SN* (subjective and non-evaluative), and two objective, named *OO* (objective with opinion) and *O* (objective with no opinion).

Recognizing subjectivity is definitely important in the NLP field since it is linked with sentiment. In the case study of Bruce & Wiebe (2000), four human judges assigned subjectivity and objectivity tags to clauses from the 'Wall Street Journal', and assessed their uncertainty. The aim was to examine whether there was an agreement between the judges and understand any possible patterns that might have occurred. As a result, a more appropriate classification model could be defined. Latent class model was used to *"automatically assign to each clause the semantic tag that best explains the pattern of agreement among the judges' tags"* (Bruce & Wiebe, 2000, p.188). Using a confusion matrix, it is clear that subjectivity is more easily identified by all the judges. The classification differences found were later examined, and interestingly enough it was found that bias among judges existed. Overall, their approach can refine the reliability of human judgement in general.

Next, is the examination of the parts of speech. Being the basis for any sentence, POS is the most commonly used analysis method for extracting sentiment (Singh, Mukherjee and Mehta, 2011). It has been found that adjectives are more likely to be related with opinionated sentences (Nicholls & Song., 2009). In fact, research has shown that the presence of adjectives was highly correlated with subjective sentences (Pang & Lee, 2008). Nevertheless, nouns, verbs and adverbs, also contribute to the overall polarity of a sentence or a document (Nicholls & Song., 2009, Pang & Lee, 2008). Nicholls & Song (2009) used a Maximum Entropy Modelling (MEM) classifier to experiment with the optimal word weighting and developed a method that improves the overall sentiment classification. MEM was chosen due to its key principle: *"agree with everything that is known but carefully avoid assuming anything that is not known"* (Nicholls & Song, 2008, p.1593) and for not assuming feature independence. The classifier appeared to be 76.5% accurate at its lowest. According to Pang & Lee (2008), POS patterns offer a great assistance in the NLP field.

One step further, comparative expressions and opinion spams complicate things further. The first characterises a comparison between two or more entities, where usually no negative or positive sentiment is expressed for either of the entities in the sentence. Hence, comparative expressions can be found in subjective and objective sentences.

- **Subjective Comparison:** “The battery life of a Blackberry mobile phone is better than the battery life of an iPhone.”
- **Objective Comparison:** “Blackberry Z10 is bigger than the iPhone 5.”

Comparison appears regularly in sentiment analysis since people find it easier to evaluate an entity by comparing it with a similar and familiar to them entity. The familiar entity acts as a baseline for the new entity to be compared against, ideally making a more convincing review (Jindal & Liu, 2006a). This is very common especially in the business field as a result of competition (Jindal & Liu, 2006b).

Comparative sentences can be recognised if the words ‘*than*’ and ‘*as*’ are found in the sentence; these represent the standard form of comparison between entities. POS can be effectively used to distinguish comparative sentences, but exceptions to the rule do exist. Besides that, comparative expressions are used for creating a baseline for evaluating quality (Jindal & Liu, 2006b). The use of different forms of adjectives and adverbs, like words ending in ‘*-er/-est*’ and words like ‘*more/most*’ and ‘*less/least*’ (defined as comparatives and superlatives respectively) achieve this (Liu, 2012). However, there are irregular forms of comparative and superlatives words that do not follow the above pattern. Examples include: ‘*exceed*’, ‘*outperform*’, etc.

Due to the special nature of comparative expressions, different analysis techniques are required. Jindal & Liu (2006b) studied the mining of comparative sentences and identified such

techniques. Amongst others, they distinguish between gradable and non-gradable comparatives (refer to Table 3.1).

Table 3.1: Types of Comparatives (Jindal & Liu, 2006b).

Gradable Comparatives	Non-Gradable Comparatives
<ul style="list-style-type: none"> • Non-Equal Comparatives: Comparisons between two or more entities for the purpose of defining an ordering with regard to certain features. It is most common in sentiment analysis, especially in reviews, since it includes user preferences. • Equative: Comparisons where the aim is to identify equal entities with regards to certain features. • Superlative: Comparisons that rank/favour one entity over all others 	<ul style="list-style-type: none"> • Non-Gradable: Comparisons between features, but no explicitly grading is provided.

With a 98% recall, they can evidently suggest keywords are a good approach in identifying gradable comparatives. However, the precision of the classification was very low (32%) so Jindal & Liu (2006b) designed a new approach to improve learning. First, all non-comparatives were eliminated using the keywords and then supervised learning (see Section 3.2) classified the remaining into comparatives or non-comparatives. Keywords include words ending in ‘-er’, ‘-est’, etc. as defined above.

On the other hand, opinion spams describe the anonymous opinions that could be found anywhere a review might be required. Examples include surveys, social media tools, e-commerce tools, etc. This is of particular interest for our project, as survey responses may involve such opinions. Anonymity allows people with “*hidden agendas*” (Liu, 2012, p.14) to express fake opinions and reviews. This kind of behaviour is called *opinion spamming* and it is

very hard to be detected in a text due to the nature of the language used. Jindal & Liu (2008) studied the characteristics of reviews and the behaviour of reviewers found in amazon.com. They identified three forms of opinion spamming outlined below.

- **Deliberately Made Opinion Spamming (Type 1):** Reviewers giving undeserving positive and negative reviews, usually as a result of unknown incentives, in order to deliberately promote or damage the objects (aspects) of their interest. In order to disguise their reviews and make more convincing and trustworthy comments, they use certain features to make their reviews seem more real. While in real life people tend to detach themselves when they are trying to deceive someone by lying (using words such as '*she*', '*they*'), online opinion spammers do exactly the opposite. Exploiting anonymity, they use words such as '*I*', '*myself*', '*mine*', etc. to show their direct implication with the product and make their argument more believable (Liu, 2012).
- **Brand-Focused Opinion Spamming (Type 2):** Reviewers that disregard the object or service involved in the review and instead focus on the brand image; their intentions are either to promote it (acting on behalf of the company or brand-loyalty) or to demote it (competition or personal dislike). Unlike the other types of opinion spamming, reviews that fall under this category may actual reflect the reviewers' honest opinion. However, they are still considered spam as they are often prejudiced or biased and not concerned with the actual products (Jindal & Liu, 2008).
- **Non-Reviews (Type 3):** Advertisements and irrelevant non-opinionated reviews (e.g., questions, answers and random texts) are involved. Since they do not express actual user opinion, they are not strictly considered opinion spam. On the other hand, they do not offer any information about the specific products.

According to Jindal & Liu (2008), Type 2 and Type 3 can be identified by humans without any difficulty. Therefore, it also becomes fairly easy to be distinguished using supervised learning; 470 manually labelled spam reviews were able to achieve very good classification results (Jindal & Liu, 2008). On the contrary, Type 1 opinion spams are considered impossible to detect. Because of their nature, further information will need to be considered, to distinguish if the comments are genuine or not (Liu, 2012). Furthermore, it has been found that deliberately created spams can be harmful (Jindal & Liu, 2008). Hence, the challenge lies in finding ways to distinguish opinion spams from genuine opinions.

Jindal & Liu (2008) observed that opinion spamming is associated with duplicate and near-duplicate reviews. After some research, they concluded that it is more common for one person to give fake reviews to different products rather than group spamming. As a result, they developed a logistic regression model to deal with Type 1 spamming that used duplicate reviews as positive training examples and the rest as negative training examples. To construct their model, they considered individual (single reviewer registers multiple user-ids) and group spamming, and extracted information features from the reviewer, the content of the review and the product being reviewed. The model was able to accurately predict duplicate reviews, but the researchers wanted to take it one step further and see if it was able to recognise non-duplicate spam reviews. Using lift curves, the effectiveness of the model on outliers was examined and it was shown that it can predict non-duplicate spams to a good extend.

Opinion spamming is essentially deliberate lying. Similarly, people use sarcasm to say exactly the opposite of what they really mean (Liu, 2012). They do not actually lie, as they give their honest opinion for a specific product, unlike opinion spams, but they do it in a sarcastic way. Nevertheless, sarcasm is equally difficult to recognize in the text, especially when it comes to artificial intelligence systems. Humans are generally keener in recognising sarcasm, but due to

the unclear nature of sarcasm, even humans may be confused as to whether a comment is sarcastic or not (Tsur, Davidov & Rappoport, 2010). Studies have shown that people need more effort to understand sarcasm (McDonald, 1999).

Examples of sarcastic sentences include:

- “Don’t bother me. I’m living happily ever after.”
- “I work 40 hours a week to be this poor.”
- “Earth is full. Go home.”
- “Trees died for this book?”

Tsur, Davidov & Rappoport (2010) concentrated in finding an approach to identify sarcasm. A small set of labelled sentences was used as a starting point and expanded through Web search. The assumption that sarcasm is commonly used in texts where other sarcastic sentences exist was made by the researchers. The finalised set was then processed through pattern-based and punctuation-based features, i.e. high frequency words and punctuations were used, in order to build a semi-supervised model. Results showed an overall good precision and recall. It is also worth mentioning that punctuation usage decreased precision but had a slightly better recall.

Tepperman, Traum & Narayanan (2006), however, showed that the use of punctuation may be beneficial in recognising sarcasm. They studied the expression “*Yeah right*” which is a very commonly used phrase that can be found in many sarcastic sentences. The different uses and forms of this expression were examined (see Table 3.2), along with its position in the sentence and the gender behind the sarcastic comment. Overall, they concluded that if the “*Yeah right*” expression is accompanied with laughter, it is more likely to represent a sarcastic comment.

Table 3.2: "Yeah right" in different speech acts (Tepperman, Traum & Narayanan, 2006).

Name	Explanation	Example
Acknowledgment	Used as a response in a conversation to show understanding and acknowledgement of what the other person said.	Person A: "So, I continue straight ahead and turn on the first right." Person B: "Yeah right."
Agreement and Disagreement	This is self-explanatory. It is used to show a person's agreement or disagreement for a statement. In the case of disagreement, it is considered sarcastic.	Person A: "A thorn in my side: bureaucracies." Person B: "Yeah right, I agree."
Indirect Interpretation	Used as a sarcastic statement while telling a story. It is not involved in a dialogue.	Person A: She suggested we should buy a new house and I thought, "Yeah right, like we have the money!" Person B: [laughter]
Phase-Internal	When "Yeah right" is part of a longer comment or response.	Person A: "Park Plaza, Park Suites?" Person B: "Park Suites, yeah right across the street, yeah."

A similar concept to sarcasm is the use of negation; both are considered sentiment shifters as they have the ability to change the polarity of the sentence. Negation is usually accounted in computerised methods for sentiment analysis, where many researchers use a list of keywords to process their data. Such keywords include: 'no', 'not', '-n't', 'never', 'less', 'without', 'barely', 'hardly', 'rarely', 'no longer', 'no more', 'no way', 'no where', 'by no means', 'at no time', and 'not (...) anymore' (Hogenboom et. al., 2011). Hogenboom et. al. (2011) considered four different methods to determine negation.

- (1) Negate the whole sentences where a negation keyword exists.
- (2) Negate the sentiment of the first sentiment carrying word following or around a negation keyword.
- (3) Negate only the sentiment of the first word following a negation keyword.

Results showed that method (2) significantly improves accuracy and method (1) is the least favourable approach when dealing with negation.

As sentiment shifters, negation and sarcasm can change the entire meaning of the sentence with the use of negation words or by adding a sarcastic tone to the sentence, respectively (Liu, 2012). Conditional sentences (*conditionals*) are somewhat alike. They do not change the meaning of the sentence but they provide a condition that has to be met. They are separated into two parts, the condition and the consequence, which are depending on each other (Liu, 2012; Narayanan, Liu & Choudhary, 2009). When appeared in the text, they affect the overall sentiment polarity, which is very hard to identify due to the delicateness associated with this kind of sentences. Examples of conditional sentences can be seen below.

- If you go to school, I will buy you chocolate.
- If you run, you might catch the bus.

As we can see from the examples above, '*if*' is the key distinguishing word for conditionals. However, depending on the intentions of the speaker or writer, other words can be used (see Table 3.3). Most of them can replace the word '*if*' without any complications, but others tend to give a slight different meaning. A comparison between '*if*' and '*in case*' might clear things out.

- **Constructing a conditional sentence using 'if':** If it rains, I will take a jacket with me."
- **Constructing a conditional sentence using 'in case':** "I will take a jacket with me in case it rains."

It can be seen that 'if' is used to state a consequence if the condition is met; the person will take a jacket as a consequence of raining. On the other hand, 'in case' acts more like a precaution; the person is superstitious and takes the jacket anyway even though at the moment it is not raining.

Table 3.3: *"Percentage of sentences with some main conditional connectives"* (Narayanan, Liu & Choudhary, 2006).

Conditional Connective	% of sentences
If	6.42
Unless	0.32
Even if	0.17
Until	0.10
As (so) long as	0.09
Assuming/Supposing	0.04
In case	0.04
Only if	0.03

Narayanan, Liu & Choudhary (2009) studied the polarity differences caused by conditional sentences. Positive, negative and neutral classes were considered. Three classification methods (see Table 3.4) were proposed for identifying sentiment polarity. Afterwards, they built two models using Support Vector Machine (SVM): a two-class classifier model (negative and positive sentiment was involved) and a three-class classifier model (neutral sentiment was added). Experimenting with features such as POS, position of sentiment words, non-sentiment words, future and past tense, they tested the three proposed methods.

Surprisingly enough, the clause-based classifier was the one with the worst performance, while consequent-based and whole-sentence-based were very similar in their results. This provides supporting evidence for the observation of the researchers that the consequent part is key for identifying sentiment in a sentence (Narayanan, Liu & Choudhary, 2009). It is worth mentioning that that whole-sentence-based classifier had slightly better results than the others. Moreover, results showed that accounting for all the features improves accuracy, recall and precision. Interestingly enough, taking neutral sentences into consideration decreased the overall performance of the model by around 10%.

Table 3.4: The three classification methods used in the Narayanan, Liu & Choudhary (2009) paper to categorise conditional sentences in any of the three classes: positive, negative or neutral.

Clause-based classification	Consequent-based classification	Whole-sentence-based classification
Separate the two parts of a conditional sentence and classify them separately. Both use labelled training data which are then tested on unseen data. Furthermore, the topic of the sentence was predicted. At the end, the two classifiers were combined and the polarity was chosen based on the topic prediction.	Observing that the sentiment of the sentence is usually formed from the consequent part of conditionals, Narayanan, Liu & Choudhary (2009) classified only the consequent part using the same approach as on the clause-based classification.	A single classifier was used to determine the polarity of the whole sentence. Opinion weight and topic location was also included in this classification.

The complications of the natural language described above provide a justification of why NLP is such a comprehensive field. According to Narayanan, Liu & Choudhary (2009), realistically, it is improbable that one method will be able to be used for all different types of sentences, as sentiment is expressed in different ways. They suggested a divide-and-conquer approach, where researchers should focus their studies on each type of opinion sentences. This will essentially provide more informative methods on NLP. In Section 3.3, we will examine

classification methods suggested by researchers as an attempt to solve the sentiment analysis problem described in this section. A lot of them focused on a particular area of sentiment analysis, while others examined the problem in a more general manner. Gathering information from all the approaches and considering the different areas each researcher studied, shaped an informed decision on what the methodologies of this report should be.

3.2 Supervised and Unsupervised Learning

Supervised and unsupervised learning have been widely used over the years for different applications. It is not a surprise we encountered them in the NLP field. They are used to construct artificial intelligence systems that have a mind of their own, and can help in the processing of data. As far as sentiment analysis is concerned, supervised and unsupervised learning is mainly used for classification purposes.

As the name indicates, supervised learning methods, use pre-labelled data (usually this is done by a human) to train the model using a specific classifier and then predict the label of some unseen data of the same domain as the training data. Unseen data are also known as a testing set. It is important to separate the data sample beforehand so that the data supplied for testing (testing set) are not included in the training set. This way the accuracy and precision of the classifier can be assessed in order *“to establish the reliability of supervised machine learning methods”* (Grimmer & Steward, 2013, p.3). The usual ratio is 75% for training and 25% reserved for testing.

During training, the chosen classifier extracts certain features that can be used to distinguish all the labels provided. Researchers often focus on negative and positive labelling and chose to ignore the neutrality that might be associated; due to the difficulty to assign a neutral label. After the training is completed, testing data are entered into the built model. Results are

assessed afterwards and changes are made, if necessary. Figure 3.1 provides a visual interpretation of how supervised learning works in the domain of sentiment analysis.

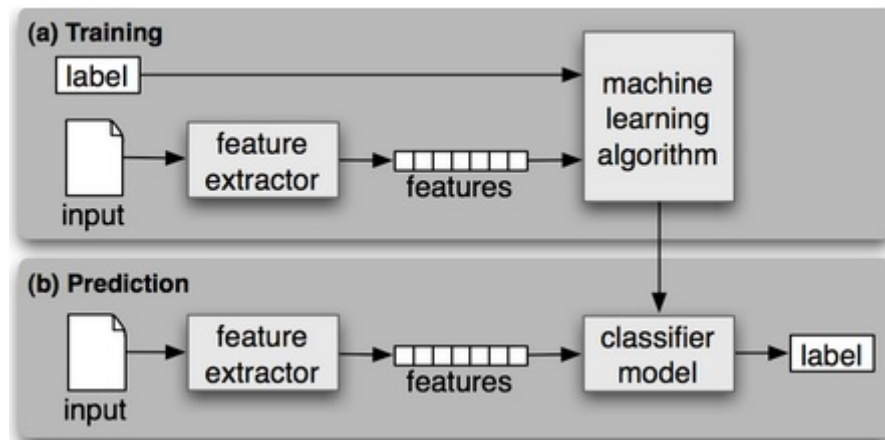


Figure 3.1: *Supervised classification* (Bird, Klein & Loper, 2009, p.222).

Bird, Klein & Loper (2009) explained supervised machine learning using Figure 3.1. Their description for the above figure was: “(a) During training, a feature extractor is used to convert each input value to a feature set. Pairs of feature sets and labels are fed into the machine learning algorithm to generate a model. (b) During prediction, the same feature extractor is used to convert unseen inputs to feature sets. These feature sets are then fed into the model, which generates predicted labels.” (Bird, Klein & Loper, 2009, p.222).

Unsupervised learning, on the other hand, does not need any labelled data. Actually, the classifier relies only on test data, i.e. it tries to extract information from unlabelled data in order to build a model that will essentially provide a classification label, in our case, a sentiment polarity.

In general, supervised learning methods are preferred due to the difficulty associated with unsupervised learning. Further to the learning methods, one must choose which classifier should be used for the sentiment analysis. Through the literature, the most commonly used classifiers are the following:

- **Naïve Bayes (NB).**
- **Support Vector Machine (SVM).**
- **Maximum Entropy (MaxEnt).**

Pang, Lee & Vaithyanathan (2002) showed that SVM and MaxEnt have more accurate classification than Naïve Bayes. Their conclusion came from the examination of different features among the three classifiers (see Figure 3.2).

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	78.7	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	82.9
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	82.7
(4)	bigrams	16165	pres.	77.3	77.4	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	81.9
(6)	adjectives	2633	pres.	77.0	77.7	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	81.4
(8)	unigrams+position	22430	pres.	81.0	80.1	81.6

Figure 3.2: Classifier's performance (Pang, Lee & Vaithyanathan, 2002, p.83).

3.2.1 Evaluation Methods for Supervised Machine Learning Classification

When supervised learning methods are in use, validation is also recommended together with testing to assess the performance of the classifier. Cross validation is a usual measure for this kind of task. A k -fold cross validation divides the data into k equal groups and performs cross validation where every fold is used as a testing set in the k different iterations. Usually a 10-fold cross validation is used (Figure 3.3). It can also be used to overcome the chance of overfitting the model. Overfitting occurs when the system memorizes the data provided instead of learning from them. Hence, generalization accuracy decreases (Mitchell, 1997).

Furthermore, the size of the data sample can result to under-training or over-training the classifier.

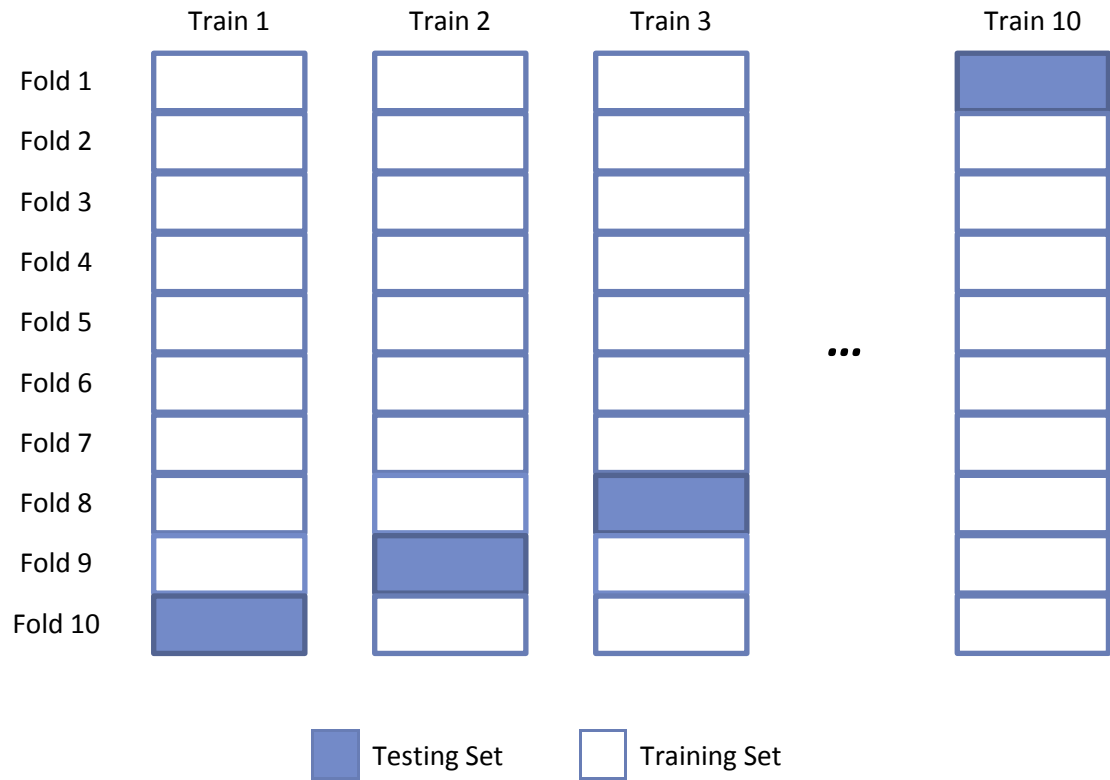


Figure 3.3: A 10-fold Cross Validation.

To compute the total accuracy of a k -fold cross validation, the average of each fold's/train's accuracy is measured (Equation 2).

$$Total\ Accuracy = \frac{Accuracy\ 1 + Accuracy\ 2 + \dots + Accuracy\ k}{k} \quad \text{Equation 2}$$

Since supervised learning works with human labelled data, there needs to be a measure to ensure the reliability of these labels. In other words, an examination on whether other people understand the labels given by the human. A common measure is the Cohen's kappa statistic. It measures the agreement of the predicted data and the actual data, taking into consideration

that agreement can occur by chance. The resulting score ranges from 0 to 1 and it should be expected anything more than 0 to ensure good classification reliability. A close-to-zero result shows that any agreement is more likely to be due to chance rather than correct prediction. A confusion matrix is used to calculate the Cohen's kappa statistic.

Table 3.5: Example of Confusion Matrix

		<i>Predicted Class</i>			<i>TOTAL</i>
		Peach	Apple	Orange	
<i>Actual Class</i>	Peach	7	3	0	10
	Apple	0	10	0	10
	Orange	1	3	6	10
	TOTAL	8	16	6	30

The confusion matrix measures the classification performance of the algorithm. For example, the confusion matrix in Table 3.5 suggests that from a total of ten oranges, only 6 were categorised correctly as oranges. The remaining four were wrongly predicted as three apples and one peach. Similarly, tables of confusion can be created for each class independently, as shown in Table 3.6. A guideline of how the table of confusion was created is shown below:

- **True Positives (TP):** The number of peaches that were correctly classified as peaches.
- **True Negatives (TN):** The number of all the remaining fruits that were correctly classified as non-peaches.
- **False Positives (FP):** The number of all the remaining fruits that were wrongly classified as peaches.
- **False Negatives (FN):** The number of peaches that was wrongly classified as any of the remaining two fruits.

Table 3.6: Table of confusion for Peaches

	Positives	Negatives
True	7	19
False	1	3

Evaluation of the classifier is required. Calculating the precision and recall of the classifier, as well as the accuracy, are usual metrics to assess the reliability of the learning method. Accuracy denotes the proportion of correctly classified texts (Equation 3).

$$Accuracy = \frac{\text{Number of Correctly Classified Texts}}{\text{Total Number of Texts Available}} \quad \text{Equation 3}$$

Even though it might seem like an appropriate measure to evaluate a classifier, it is advisable to consider the precision and recall rates of the classifier together with its accuracy score. Precision is the probability of accurately predicting the category a particular text belongs to (Equation 4). Recall refers to the memory of the classifier (Equation 5). Given a labelled dataset (supervised learning), recall represents the probability that the label would be correctly identified by the system (Grimmer & Steward, 2013). The F-measure is often used to combine precision and recall in one score (Equation 6).

$$Precision = \frac{TP}{TP + FP} \quad \text{Equation 4}$$

$$Recall = \frac{TP}{TP + FN} \quad \text{Equation 5}$$

$$F - \text{measure} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \text{Equation 6}$$

It is worth mentioning that in this project precision and recall rates are directly associated since the categories used for classification are the labels the machine has to learn to recognise. Measuring the accuracy gives a general idea of the model's performance. This is enhanced

with the recall and precision rates as well as the F-measure. However, these do not take into consideration the true negatives. A high number of TN ensures the reliability of the classifier. Cohen's kappa statistic is used to solve this problem and compensate for the ignorance of TN. The combination of these measures enables a well-rounded classification analysis.

3.3 Sentiment Classification Problem and Techniques

Having already defined the major difficulties faced in sentiment analysis; in this section we will examine ways to overcome those challenges. The aim is to outline techniques and methods used for classification purposes that have resulted from research on various topics related with NLP.

Liu (2012) studied several forms of sentiment classification in the book "Sentiment Analysis and Opinion Mining". It was pointed out that the usual assumption of researchers that a sentence only contains a single opinion is rarely true and outlined ways to overcome this. In more detail, a sentence can contain both positive and negative sentiment, each associated with a different aspect described in the sentence. In the same way, it does not make sense to assign a single label to a document, which is nothing more than a lot of sentences combined together. McDonald et al. (2007) examined both document-level and sentence-level classification in an attempt to correct this (in Liu 2012). They assigned labels to both sentences in the reviews and to the whole document, containing all the reviews. The labelled text was then used to train the classifier. The outcome of using both levels was more accurate results for both levels of classification (Liu, 2012).

The successful combination of document-level and sentence-level encourages proceeding to other measures to improve accuracy even more. It is suggested that being able to identify whether a sentence, or a document for that matter, is positive, negative or neutral is not

sufficient if you do not know the cause of the polarity. Ding, Liu & Yu (2008), as summarized by Liu (2012, p.60), introduced a lexicon-based approach to define sentiment orientation for aspects that ended up performing quite well in many cases. They considered sentiment shifters, like negation and sarcasm, and contrary words, which significantly influence polarity.

Lexicon models are a common approach in sentiment analysis. Many researchers have approached the sentiment analysis problem by creating lexicon lists with little iteration each time. The basis is to distinguish between negative and positive words and assign a score to each. Then, the data are pre-processed to the interest of the researcher – some remove punctuation and some do not, for example – and a machine learning approach is followed. The aim is to produce a near-perfect combination that will improve the accuracy and reliability of the classifier as much as possible. Some would argue that this should be fairly easy as NLP has been studied for several years now. Conversely, a perfect combination of features and processes might not work in a different domain than the original one. Adding to the already challenging environment of sentiment analysis, language differences play an important role since most of the research is based on the English language. Translation is maybe the obvious option; however, not an ideal one as it may change the meaning of a sentence and subsequently the meaning of the document as a whole. This is due to the existence of different vocabularies. In most cases, the overall meaning does not change, but the choice of words during the translation may result to minor inconsistencies, which are crucial in the NLP field. According to Liu (2012), current research proposed three main strategies for the resolution of language differences:

- Translate test sentences in the target language into the source language and classify them using a source language classifier.
- Translate the source language training corpus into the target language and build a corpus-based classifier in the target language.
- Translate a sentiment or subjectivity lexicon in the source language to the target language and build a lexicon-based classifier in the target language.

Kim & Hovy (2006) tested these strategies on sentiment orientation and subjectivity using German, English and Romanian (in Liu 2012, p.53). Surprisingly enough, predictions were relatively precise, but recall was relatively poor. Nowadays, tools exist that examine cross-language subjectivity and cross-domain subjectivity. Nevertheless, due to the extensive research on the English language, cross-domain is more easily found in research papers. This might be due to the fact the cross-domain implicates the whole document and any differences can change the meaning and cause severe problems to the classification (see Table 3.7). Language differences usually influence the weight of the words and do not affect the overall polarity of the document.

Table 3.7: Sentiment polarity differences due to change of domain (Bollegata, Weir & Carroll, 2013, p.1721)

	Books Domain	Kitchen Domain
Positive expressions	'interesting', 'high quality', 'well researched'	'professional', 'high quality', 'delicious'
Negative expressions	'boring', 'lengthy', 'badly written', 'disappointed'	'disappointed', 'rust', 'disgusting'

As shown in the above table, the term '*high quality*' is always used to express a positive sentiment. The same applies for the negativity of the term '*disappointed*'. Despite the expected language similarities between the two domains, words like '*delicious*' and '*rust*' are

not likely, if not impossible, to be present in a books domain document. Essentially, a classifier trained on data from a books domain will not be able to understand the meaning of the words '*delicious*' and '*rust*' in order to properly assign the sentiment polarity.

As an attempt to resolve the different-domain problem, Aue & Gamon (2005), as outlined by Liu (2012, p.38) showed that combining small amounts of pre-labelled data in the training phase, improved the performance of the model since both labelled and unlabelled data could be used (Liu, 2012). This was followed by other studies that focus on the common features of the different domains (Blitzer et. al., 2007, in Liu 2012, p.39) or on topic identification (He et. al., 2011, in Liu 2012, p.40).

Keim & Oelke (2007) attempted to combine characteristics of the opinion holder through authorship. Approaching sentiment from another perspective, they examined different levels of classification such as words, sentences and documents as a means to create a blueprint that can provide more information about the text. Using visualisation, they applied their technique on books of two authors in order to examine authorship attribution. Authorship attribution is used with texts where there is doubt about the author or the author is unknown, by using specialised features, aiming to identify the rightful author. Average sentence length was one of the characteristics used to see the similarities between the books of each author. Even though their research was directed towards authorship attribution, their visual blueprint can be applied to any text. The combination of visualisation and feature selection proved useful and provided promising functions that could improve reliability in the future.

A more traditional approach requires processing the data prior to training to enhance the classifier's performance. This involves eliminating punctuations, examining frequency of words or even taking into consideration negation. It is not necessary to include everything when investigating new techniques into text classification. Usually the researchers choose the

appropriate features depending on the purpose of the report and the data provided. Most of these features will be explored in order to understand their use and eventually make the most appropriate feature selection combination to fulfil the objective of this project.

Review data are often preferred for sentiment analysis since the use of informal language enables freedom in writing and is the closest form to the spoken language. They are considered as potentially invaluable sources of information as suggested by Dey & Haque (2008). Excessive use of exclamations or symbols that intend to show the emotions of the reviewer (emojis) are examples of the informality, which can be found very useful when analysing sentiment (Dey & Haque, 2008). Nevertheless, in order to build an accurate model, pre-processing is necessary. One must choose the features involved in the training depending on the project in order to minimise the noise in the text (Liu, 2012). Researchers must decide whether, to ignore punctuation, transform text in lowercase or not, introduce stemming words or if the frequency of words (common and uncommon words) is worth examining. Then, depending on the project subject, other techniques might be added to account for subjectivity, negation or even comparative sentences. It is up to the researcher to choose.

In general, it has been found that punctuation, capital letters and stopwords do not offer extra information to the data as far as the machine learning method is concerned (Aggarwal & Zhai, n.d.). They are often eliminated in the pre-processing phase (Nicholls & Song, 2009) to provide greater accuracy, but there are studies that chose to include them (Pang, Lee & Vaithyanathan, 2002).

A lot of debate exists on whether frequent and infrequent words should be eliminated from the training data. It all comes down to the field examined. Aggarwal & Zhai (n.d.) studied similarities between documents and suggested that using inverse document frequency reduces the importance of the most frequent words in the document. This ensures that the

matching of the documents is more influenced by more discriminative words. Likewise, infrequent words, which are often a result of misspelling or typographical errors, do not offer value to the similarity (Aggarwal & Zhai, n.d.) and are also eliminated. All these can be *“leveraged in order to improve the quality of results”* (Aggarwal & Zhai, n.d., p.78). On the other hand, Dey & Haque (2008) believed that misspelling should not be ignored and should instead be replaced with the most correct word from the dictionary.

Pre-processing also includes accounting for concepts that could affect the overall classification algorithm. Most of them were outlined in Section 3.1 along with approaches to deal with them. In this section, we will focus more on the attributes that form the basis of NLP, such as POS and punctuation-based features.

Maks & Vossen (2012) examined the semantic categorisation of verbs, nouns and adjectives in an attempt to create a lexicon model for the purpose of sentiment analysis. They were amongst the few who drew their attention to the opinion holder since they believed it can play an important part to the sentiment expressed. Indeed, attitude holders can change the reliability of the results.

Furthermore, Turney's (2002) unsupervised technique extracted sentiment using specific phrases or syntactic patterns that were more likely to be expressing a specific opinion (in Liu, 2012). His research was based on the Part of Speech (POS) of words. By examining the patterns between adjectives (JJ), adverb (RB), comparative adverb (RBR), and superlative adverb (RBS) words, an algorithm was constructed that extracted two consecutive words which satisfied the patterns shown in Table 3.8. The result was a *pointwise mutual information* (PMI) equation that measured the degree of statistical dependence between the two terms.

$$PMI(term_1, term_2) = \log_2 \left(\frac{\Pr(term_1 \wedge term_2)}{\Pr(term_1) \Pr(term_2)} \right) \quad \text{Equation 7}$$

Table 3.8: Patterns of POS tags for extracting two word phrases (Turney 2002, in Liu, 2012, p.35).

	First word	Second word	Third word (not extracted)
1	JJ	NN or NNS	Anything
2	RB, RBR, or RBS	JJ	not NN nor NNS
3	JJ	JJ	not NN nor NNS
4	NN or NNS	JJ	not NN nor NNS
5	RB, RBR, or RBS	VB, VBD, VBN, or VBG	Anything

Afterwards, the algorithm calculates the average *SO* of all extracted information and classifies the document as positive or negative, depending on the polarity of the average *SO*.

It is important to note that neutrality is often ignored as a polarity class. This is because it is considered very difficult to define neutrality as it can take various different forms. For example, it can indicate an equal amount of negative and positive opinions, in which case the text can be characterised as neither positive nor negative and hence a neutral label is assigned, or it can imply that no opinion is held. Subsequently, many of the research papers ignore neutrality, making the problem easier. (Liu, 2012). In this report, neutrality expressed by people was considered as part of the analysis.

3.4 Computerised Tools for Sentiment Analysis

In the age of information, computerised methods are continuously increasing. If used wisely, technology can provide a competitive advantage (Spangler et. al., 2010) to all business. However, IBM Corporation (2012) sets an important question: *“Can machines understand human communication?”*

The answer lies in the combination of the two; humans and machines working closely together. Computers can calculate whether a given phrase is positive or negative with a certain degree of confidence and humans can review the low-confidence results and advise the machine how to grade them. Pang & Lee (2008) provided evidence that could support the collaboration between machines and humans. With time, computers will absorb more and more and results will be more accurate and useful (machine learning).

Many companies launched commercial tools that use the latest technologies and are able to analyse sentiment. They essentially use algorithms and features described in the previous sections, but all the user has to do is “press a button” to initiate the analysis. Usually no customisation is available, i.e. the user of the tool cannot change the settings of the algorithm. Depending on the package chosen, the user may have the ability to experiment with different features of the tool. Freedom of how to present the data is often given to the user. TheySay and Semantria (add-in Excel tool) are examples of this kind of tools. Both come in a free trial version where some functions are not accessible.

Customisation can be increased when programming languages are involved. Some tools use predefined platform, like Python or Java, where the user can construct their own algorithm. The most commonly used tool for sentiment analysis is NLTK which uses the Python platform. Even simpler, there are tools which do not require any programming language knowledge but the user is still given full customisation capabilities. For example, WEKA enables the user to pre-process the data fed to the chosen classifier.

Crowdsourcing is the act where data are collected from the public. Essentially, the source is the public, as the name suggests. Online surveys are usually used to gather customer's opinions, thoughts and ideas about the products and services provided. These are then analysed. It is an efficient way to achieve customer satisfaction. These fall under the category

of sentiment analysis since data gathered is often in the form of unstructured text, resulting from open-ended questions. People are encouraged to give their feedback by giving them the freedom of expression. Furthermore, as (Brew, Greene & Cunningham, 2010) suggested, crowdsourcing can be used to tackle the time-consuming labelling of data needed for supervised machine learning. It is usually an effective and easy way to collect a large amount of data. This was partly used for the aim of this project. The dataset provided by UX Labs was a result of public responses to an online survey. As we will see later, the scoring provided by the respondents was the main source for labelling the data.

In this report, we set the question on whether commercial tools perform better than a more customisable tool. The latter are referred to as non-commercial tools. We examined the accuracy of the predictions of each tool. Supervised learning was used. To the best of our knowledge, there is no literature to examine this area.

4 METHODS

The aim of this project is to investigate and assess the differences (if any) of sentiment analysis using commercial and non-commercial tools. The focus was turned to the healthcare information service domain. An online survey to evaluate a healthcare advisory website was used as the dataset to extract sentiment by the respondents. This was an attempt to gain insights and discover patterns that could improve sentiment analysis in the particular field.

After examining different tools, four were chosen to be used in this project: two commercial and two non-commercial. Non-commercial tools were based on supervised machine learning algorithms, while commercial tools used an unknown algorithm, pre-defined by the tool itself, since no configuration was available. Also, commercial tools did not require any kind of pre-labelling to perform the analysis.

In this section, details of the methods and procedures used for the purpose of the project are outlined and explained. A sentiment classification analysis of all the available responses was made using the four tools. Further examination on the data was done by using only single-sentence responses, after observing that short text was usually used to express a single opinion. An assumption was made that single sentences would be more clearly categorised to a specific class and therefore easier for any software to extract their polarity. This contradicts Liu's (2012) observation. According to Liu (2012), a sentence is not necessarily associated with a single label. Hence, it is not advised to assign a single label to a response that normally contains more than one sentence. Taking this into consideration, we wanted to examine the sentiment classification of single-sentence responses only and test whether our assumption is valid for the dataset provided. It was also assumed that training a model using only single sentences might improve the accuracy and reliability of the results.

4.1 Data

An online survey was conducted to evaluate a healthcare information service website that has recently been updated. A questionnaire consisting of eight questions, which resulted in a total of 165 responses, was used. It included both quantitative questions, to help with the analysis, and qualitative questions, to welcome any feedback with respect to the website and its features (see Table 4.1).

Table 4.1: Survey questions.

	Question
1	How easy is it to navigate the website?
2	How easy is it to find the information you are looking for on the website?
3	How visually appealing is the website?
4	How often do you use the website?
5	What is your feedback on the changes to the website?
6	Overall, are you satisfied with the website, neither satisfied nor dissatisfied with it, or dissatisfied with it?
7	What improvements would you like to see?
8	How likely are you to recommend the website to others?

None of the questions were mandatory. Therefore, information was missing for some questions; some did not share their feedback (Questions 5 and 8) and others did not provide a rating for certain questions. For instance, only 137 people responded to Question 5. Survey participants were asked to rate, amongst other things, the navigation and design of the website as well as their overall satisfaction using a scale of 1-5, 1 representing the most positive feedback for the relevant question. Open-ended questions were included, one for general feedback and one for improvement suggestions (Questions 5 and 7), to give people the freedom to express their personal opinion. In short, the survey acted as a comparison between

the old and the new version of the healthcare website. This project was mainly based on the feedback gathered by the 137 respondents of the online survey, i.e. it was focused on the responses of Question 5. The data used for this project were provided by UX Labs.

The website for which the survey was conducted is an advisory healthcare website aimed to provide necessary information to both professionals in the field and the general public. However, the website cannot be named in this report, due to confidentiality.

4.2 Project Procedure

A quantitative analysis of the data was first made to identify possible patterns and trends. Afterwards, eyeballing browsing was executed to better understand the material – mostly qualitative data were involved here. This helped in forming an initial awareness of the sentiment hidden between the words. Using relevant studies, categories and groups were identified with the aid of keywords.

All of the above formed a general idea of the data from a human perspective. Following the human observation, two commercial products were chosen to be compared with two non-commercial products in order to satisfy the objective of this project.

4.2.1 Commercial Tools

Due to the infinite benefits resulting from analysing text and especially sentiment, there was a variety of products available in the market to choose from. The commercial software used for this project was Semantria, a Microsoft Excel add-in tool, and TheySay (www.theysay.io), an online sentiment analysis tool. As these are commercial tools, no processing of the data was required beforehand. However, due to format variations of the results, different measures had to be taken to evaluate each product.

At a beginning stage, the sentiment polarity of each response was found using a simple procedure for both tools. A “click of a button” was sufficient to produce the results. Screenshots presented below can be used to enhance the understanding of each system’s functionalities. Figures 4.1 and 4.2 refer to the Semantria software, while Figure 4.3 is associated with TheySay.

i. Semantria



Figure 4.1: Semantria's control panel in Microsoft Excel.

According to the output required, the relevant option is chosen from the ‘Output’ section. For a basic sentiment analysis, ‘Sentiment’ was chosen as indicated. The ‘Analytics’ section highlighted in Figure 4.1 represents the basic functions needed to create a project within the Semantria tool. A detailed description of the process Semantria followed to produce results is shown in Figure 4.2.

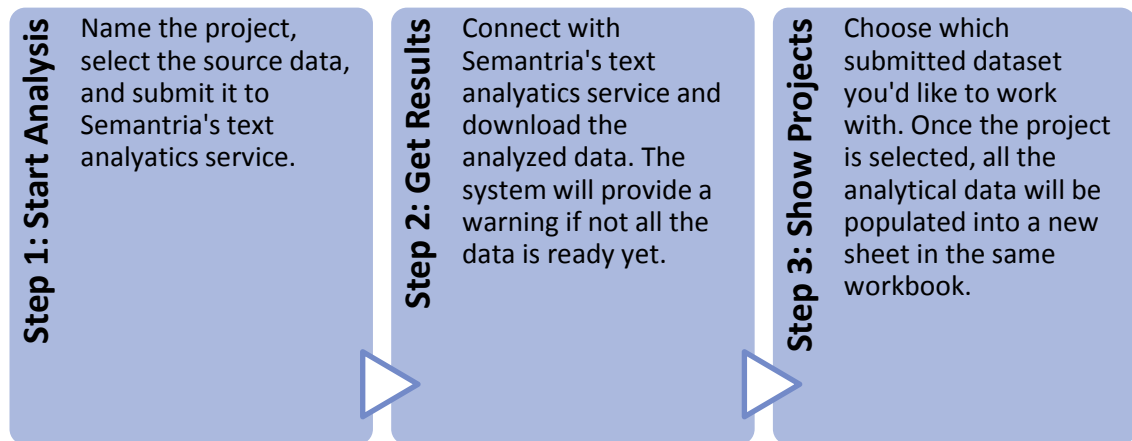


Figure 4.2: The three steps needed to create a project in Semantria.

ii. **TheySay**

Paste some text here and then click on the "Analyse" button...

Analyse Settings

0 characters

Figure 4.3: TheySay's console box.

TheySay functions in a simpler way from Semantria. The text to be analysed is entered in the box provided (see Figure 4.3). The text is then sent to the PreCeive API for analysis. Results are shown almost instantly in the same page.

Furthermore, each tool's features were examined to produce a more detailed analysis of the data. In the case of Semantria, it was necessary to manipulate the features to meet the needs of the project (healthcare domain). This was done by working through the 'Advanced' section of the control panel (Figure 4.1). For the purpose of this project, the free demo of the products was used. This meant that some features were not available. However, the performance of the

systems was not significantly affected. In order to achieve the best results possible, a thorough analysis was undertaken using every accessible feature of both tools. Table 4.2 and Table 4.3 show the available features of Semantria and TheySay respectively, along with a description for each feature.

Table 4.2: The available features of Semantria.

Feature	Description
Sentiment Analysis	Semantria's internal algorithm classifies each response as negative, neutral or positive by assigning a single score to it. Scores range from -2 to 2 to represent the two ends: negative and positive respectively.
Collection Analysis	The frequency of words is examined in the text and the most frequent words, called <i>facets</i> , are found. These are accompanied with the number of times they were found in the text and their polarity each time.
Categories	<p>Semantria has a library of 40 default categories consisting of appropriate keywords. A categorical analysis is performed by the tool and each response is assigned to the appropriate category.</p> <p>The user has the ability to build new custom categories.</p>
Queries	Manually constructed queries that are usually used when the attention needs to be turned to a particular topic. They are built using keywords and gather text according to the relevancy of the keywords.

Table 4.3: The available features of TheySay.

Feature	Description
Document Sentiment	<p>An overall sentiment analysis is provided. TheySay detects any reference to positive, neutral and negative sentiments, opinions, emotions and subjective points of views.</p> <p>The text is given three different values to represent the negativity, neutrality and positivity associated with it.</p>
Sentence Sentiment	Each sentence is considered individually and is characterised by a single polarity.
Part of Speech Recognition	The part of speech of all the words in the text is identified.
Comparison Detection	Comparative sentences are detected and shown in this section.
Humour Detection	Text that is believed to express any kind of humour is shown in this section along with a number to represent how humorous the text is. This includes sarcasm and irony.
Speculation Analysis	Speculative content is identified. Categories to distinguish the type of speculation exist and are presented in this section. Categories include among others requests and advices of the author.
Risk Analysis	Any references to risk are detected.
Intent Analysis	Expressions of intent are recognised and shown in this section.
Named Entity Recognition	The entity specified in the text is identified. Usually it is the entity which is responsible for the overall polarity of the sentence or even the whole text.
Shallow Chunk Parsing	Phrases are determined and analysed using POS tagging.
Dependency Parsing	A full grammatical analysis is completed.

It is worth mentioning that TheySay recently added a 'Gender Detection' feature where the genre of the author is detected. However, this was not present during the analysis of this project and hence not included in the results.

After individually evaluating each tool, a comparison between them was made. In order to compare the two tools, the results were transformed into a uniform format. TheySay provided three different percentages (adding-up to 100%) to represent how negative, neutral and positive each response is, while Semantria assigned only a single score, ranged between -2 and 2.

The outcome from Semantria was converted into a percentage (see Appendix B); in order to be compared with the dominant polarity number of each response from TheySay.

Comparison helped in understanding the internal procedures of the systems and in outlining possible malfunctions, as well as positive features.

4.3 Non-Commercial Tools

Most of the non-commercial products require a substantial amount of programming knowledge and aim at building a unique algorithm that matches the specific requirements and goals of the problem in question. However, in this report, rather than creating new systems, pre-existing systems will be evaluated.

WEKA (version 3.7.10) and Google Prediction API were selected. They are both open-source software that use supervised machine learning methods, i.e. labelling of the responses beforehand was needed. WEKA also enables the use of unsupervised learning but this was not used for the current project. We tried to maintain a consistent format throughout the project, in order for the comparison to be fair and accurate.

The steps needed to perform a sentiment analysis along with a brief description of the features and capabilities of each tool are listed in the subsections below.

4.3.1 Google Prediction API

Google Prediction API is assessed by RESTful interface. It could simply be described as a prediction tool. The data are imported to create a model which is then trained. A prediction then follows, to test the accuracy and performance of the model. Restructuring the problem into a format that the API can answer is necessary. Instructions to help the analysis process of specific use cases are available to the user. The use cases currently available are: sentiment analysis, purchase prediction and spam detection. Depending on the nature of the project, models, called *hosted models*, already exist in the software for a quicker analysis. However, only the demo version can be used. Due to the emphasis given on the healthcare domain, the sentiment analysis hosted model was not suitable for our analysis.

It is worth mentioning that the tool uses several algorithms for training. The API runs the algorithms on the dataset provided and automatically chooses which algorithms and parameters work best for the data. However, this is not yet visible to the user.

Consulting the sentiment analysis use case guidelines, an appropriate model for our project was conducted. The steps followed to analyse the sentiment found in the dataset are shown below.

1. **Pre-processing the data:** Punctuation and capitals were removed. All strings were inserted into quotes and where quotes already existed, double quotes were used. Any words that had to remain together and not split by the system were connected using an underscore, (_). The latter technique was used to account for negation. Negation words like 'not' were connected using an underscore with the following word (Hogenboom et. al., 2011, method 3).

2. **Labelling the data:** Supervised learning requires labelling the data before training. To assign accurate labels, the quantitative part of the survey was used. The average of all the numeric responses of each respondent was calculated. Then, using a pre-defined scale shown in Table 4.4, responses were identified as negative, neutral or positive accordingly (Appendix C). These were then reviewed by the human to assess their accuracy; i.e. whether the labels assigned represented a reasonable sentiment classification with respect to the written response. Due to the different topics associated with the closed-ended questions (refer to Table 4.1), it was assumed that the concluded label did not necessarily refer to all the features in the respondent's feedback (Question 5). The human intervention was made to resolve any errors that may have occurred. Labels were changed based on the human's intuitive knowledge of the language. The guidelines followed for the neutral classification can be found in the Appendix C, along with the final classification results.

Table 4.4: Scores assigned to each class.

Class	Score Range
Positive	1 – 2.7
Neutral	2.8 – 4.2
Negative	4.3 – 5

3. **Separating data into two sets:** A training set (75%) and a testing set (25%) were created to be used for training and testing respectively.
4. **Creating a csv file:** Google Prediction API works better with a csv file. Instructions for the appropriate format of the file are given in the tool's website. One response was required per row and the first column needed to include the label of the row (refer to

<https://developers.google.com/prediction/docs/developer-guide#data-format> for the required training data format).

5. **Training the model:** A model was created by training it using the training set available.
6. **Testing the model:** Using the testing set, the performance of the model was assessed.
7. **Performing cross validation (refer to Section 3.2):** This was done manually. The dataset containing the 137 responses was divided into four equal folds. Afterwards, three folds were used for training and the remaining one was used for testing. Through rotation, all folds were used as a test set once.

During testing, each response was given three different values (adding-up to 1) to represent how negative, neutral and positive each unseen response was. The predictions were then matched with the actual labels to examine the accuracy of the classifier. The same procedure was followed while performing cross validation tests to assess the reliability of the model. At the end, the percentage of correctly labelled responses was calculated both for testing and for cross validation purposes.

4.3.2 WEKA

Table 4.5: The available features of WEKA.

Features	Description
Data Pre-Processing	<p>With over 75 data pre-processing tools, WEKA is able to analyse data of various formats. The removal of punctuation, capitals and stopwords is resolved at this stage.</p> <p><i>Attribute Relation File Format (arff) is the default file type used in WEKA.</i></p>
Classification	<p>Classification algorithms exceed 100 in number. They are separated into groups to distinguish the nature of the classifier. WEKA includes amongst others, Bayesian methods (Naïve Bayes), tree learners (C4.5) and function-based learners (SVMs).</p>
Clustering	<p>It is usually used to support unsupervised machine learning. The provision of a variety of clustering algorithms allows this.</p>
Attribute Selection	<p>15 attribute evaluators and 10 search algorithms for feature selection exist to aid the classification performance. Attribute selection is an important aspect of the classification problem.</p>
Data Visualisation	<p>WEKA provides visual representations options to study the data. Plotting attributes against the class is one example. Also, words can be examined separately to define the most possible associative polarity based on frequency. Moreover, <i>“classified output can be compared to training data in order to detect outliers and observe classifier characteristics and decision boundaries”</i>.</p>

WEKA provides three graphical user interfaces through which the features are accessible.

These are:

- **The Explorer.** The most popular of the three graphical user interfaces. It enables the investigation of data with the support of the features identified in Table 4.5. Hence, pre-processing, attribute selection, classification learning and visualisation are available.
- **The Experimenter.** It is often used for testing and evaluating the classifier's performance. In other words, measure the efficiency of learning. Summaries are also provided in an easy-to-read format to allow easy comparison of performance.
- **The Knowledge Flow.** It is very similar to "*The Explorer*" but it includes some extra features. Modifying the parameters of the algorithms is one of the characteristics of this interface. The advantage over the 'Explorer' is that it easily prevents errors that might occur when dealing with manual processing. For example, attribute selection can be omitted using the 'Explorer' graphical user interface but not when using the Knowledge Flow interface.

It is also worth mentioning that extra classifiers can be added using a programming code. However, this was not needed for this project. Even though SVM and MaxEnt are preferred, Naïve Bayes produced an outstanding output and hence it was decided not to assess other classifiers due to time limitations.

"*The Explorer*" was used for this project. We believe that the 'Explorer' is much easier to comprehend and manipulate than the other two user interfaces. Even though WEKA's full capabilities were not essentially assessed, the aim of the project was still satisfied. In fact, it is believed that the use of The Explorer provides a more representative and comparable output with the commercial tools. Recalling that the goal was to assess the accuracy and validity of

commercial tools, non-commercial tools were used as added value to the analysis but not as a mean to overpower this project. Nevertheless, the other two graphical user interfaces should be assessed at another time for the healthcare domain. Due to time limitations, we leave that for future research.

The steps followed to analyse the sentiment found in the dataset are shown below.

1. **Labelling the data:** The same technique was used as for Google Prediction API.
2. **Separating the data:** The data were separated according to their polarity. Negative responses were grouped together as well as neutral and positive responses.
3. **Dividing the data into two sets:** Training and testing sets were created. 75% of each class was used for training and the remaining 25% was reserved for testing.
4. **Inserting the data into the system in the form of an arff file:** WEKA performs quicker when data are provided in an arff file format. Using WEKA Simple CLI (see Appendix D), data were transformed in the appropriate format required by the tool. Punctuation was removed and apostrophes were followed by a '/ '.
5. **Pre-processing the data within the tool:** Experimented with different combination as far as punctuation, stop-lists, lowercase, etc. are concerned.
6. **Selecting features:** Experimented with the different feature selection options in order to produce a more reliable model.
7. **Resampling:** Considered any resampling technique to balance the dataset.
8. **Classifying the model:** Chose the appropriate classifier from a variety of classifiers offered by WEKA. Cross validation was integrated in the classification results. A 4-fold cross validation was used as it was done for Google Prediction API.

Comparison between the two tools followed. Since the outputs produced were not in the same format, measures were taken to make them compatible with one another. Google Prediction API's outputs were converted into percentages by multiplying by 100 and the number of correctly classified responses was noted during cross validation for both tools. This was automatically provided by WEKA when the cross validation training was finished and was manually computed for Google Prediction API.

4.4 Evaluation Techniques

At a final stage, the four tools were evaluated in a consistent manner to establish their suitability for sentiment analysis in the healthcare information service domain. To achieve that, all results were converted into a common format, a single percentage of correctly classified responses. This was a simple way to examine the accuracy of the tools. While further statistical tests followed for a more advanced comparison and evaluation, percentages were used as a starting point. They represented the number of responses that were correctly predicted by each system in comparison with a baseline, which was then converted into a percentage using the equation below.

$$\frac{\text{Number of correctly predicted responses} \times 100}{\text{Total number of responses}} \quad \text{Equation 8}$$

The use of supervised learning by non-commercial tools forced the formation of a ground truth, i.e. the labels assigned to each response. As already explained, these labels were based on the quantitative responses of the survey and were reviewed by the human. Hence, they will be used as a baseline for comparison purposes between the different systems used in this project. The predictions of each tool were compared with the baseline and resulted to a single percentage of correctly classified responses (see Table 4.6). These percentages were used a comparative metric to assist in achieving the goal of this project.

Table 4.6: The process needed to achieve a common output format for comparison purposes.

Semantria	TheySay	Google Prediction API	WEKA
The single polarity suggested for each response was compared with the labels proposed by the baseline.	The dominant polarity of each response was compared with the “true” label which was based on the baseline.	Predictions were matched with the actual labels. This was done manually during testing and cross validation. The average percentage of the four models built for cross validation was used to represent the correct predictions.	No action was needed. Supervised learning was cross validated by the system itself. The output provided by WEKA is in the form of a single percentage representing the number of correct predictions.

5 RESULTS

In this section the results derived using each tool are discussed. Comparisons within the two groups were also made and are presented at the end of each subsection. As an introduction to the data and the problem we dealt with, human observations can be found at the beginning of this section to help the reader form a better understanding of the project. As already mentioned, sentiment classification was undertaken using all the available responses as well as only single-sentence responses. The performance of the tools was mainly measured through the accuracy of predictions. Statistical tests were also involved when analysing all the available responses. Due to the tight timeframe associated with the MSc project, statistics were not used to back-up the accuracy of the single sentences. It was believed that omitting the latter would not compromise the aim of this project. Instead, an initial examination was made and further work will be undertaken to compensate for any project limitations.

5.1 Human Observations

The data were quantitatively analysed at first by a human. Results are available in Figure 5.1. As we previously mentioned the score range is between 1 and 5, with 1 being the highest score possible.

Referring to Figure 5.1, the 165 respondents that participated to the online evaluation survey were not pleased with the website and its features. Excluding the appearance of the website, a typical person would be expected to have negative feedback for the website, based on the large volume of score 5 responses (Figure 5.1).

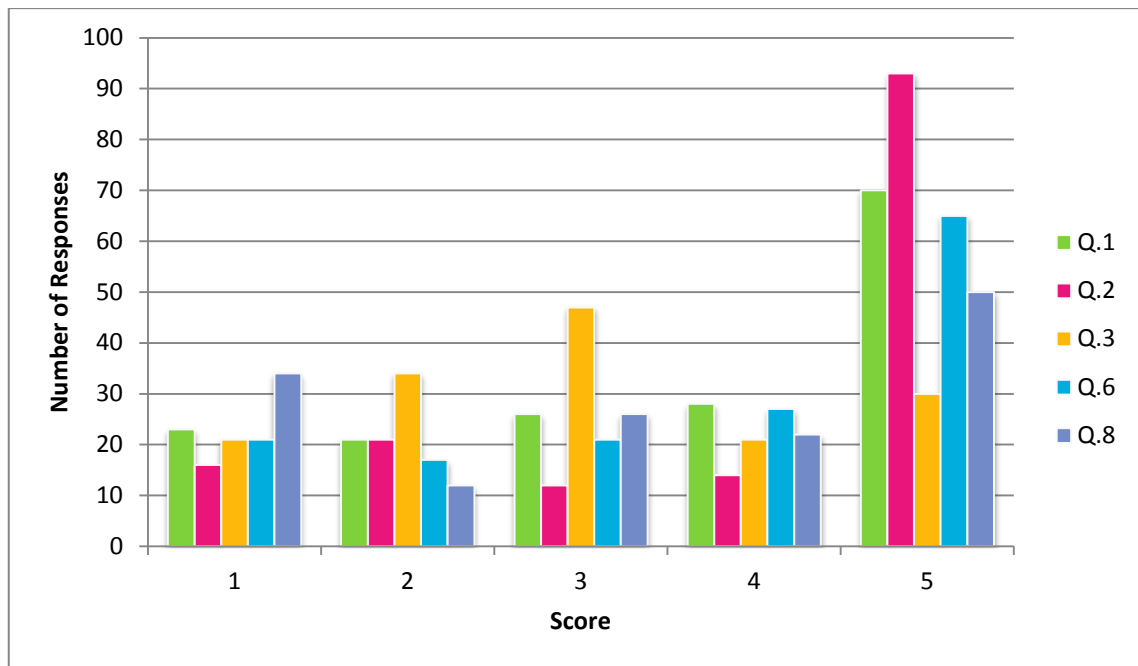


Figure 5.1: Quantitative analysis of the data.

Table 5.1: Average Scores associated with the quantitative data.

Question Number	Theme Questioned	Average Score
Q.1	Navigation	3.64
Q.2	Information	3.94
Q.3	Appearance	3.03
Q.6	Satisfaction	3.65
Q.8	Recommendation	3.29
Average		3.51

The overall feedback of the respondent was very close to negative, with the exception of the website's design (Figure 5.1). However, observing the average score of each question may suggest a different conclusion (Table 5.1). It was noticed that an average person would be more likely to rate the website by an overall score of approximately 3.51. Nevertheless, it is

worth mentioning that this was concluded from a total of 167 participants; a small number to enable accurate conclusions, not enough to enable generalisation of results.

Furthermore, eyeballing techniques were used to gain a more informative opinion. Human keywords were identified at this stage that would be used in the next section as a comparative feature with one of the commercial tools. Also, the author observed that the provision of guidance as well as navigation received strong negative feedback. Moreover, a substantial amount of sarcasm was used by the respondents and it was interesting to observe how well the systems chosen recognise this.

5.2 Commercial Tools

5.2.1 Semantria

5.2.1.1 *Sentiment analysis using all available responses*

i. Basic Sentiment Analysis with Entity Recognition

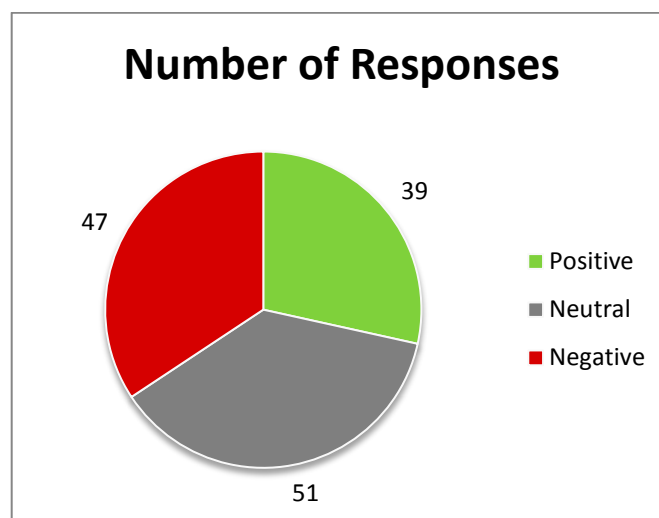


Figure 5.2: Semantria's categorisation results.

Figure 5.2 shows the number of responses that were categorised in each class by Semantria. Neutral classification was comprised of 51 responses compared to the 47 and 39 responses associated with the negative and positive polarity, respectively.

Furthermore, Semantria was able to identify entities responsible for the polarity of only 31 responses. Apart from this very low number, the entities identified were not very accurate or helpful (Figure 5.3). However, this might be due to the nature of the survey. Respondents were asked to state their feedback on the website, so the entities identified by Semantria were essentially specific features of the website. Also, the topic was predefined and most of the responses gave a direct review about the website by using the pronoun 'it' as a reference to the website. This can explain the low number of entities recognised.

Entity	Entity Type
"cancer"	Quote
"Chronic Fatigue Syndrome"	Quote
"Personality Disorders"	Quote

Figure 5.3: Examples of entities identified by Semantria.

Another observation was that a neutral label was usually given to responses stating facts, such as “CG 102 is not available”. While this is not wrong, we were expecting that responses containing both negative and positive feedback would belong to the neutral class. Instead, Semantria weighted each side (negative and positive) and assigned each response to the most appropriate class. As it was already mentioned before, Liu (2012) suggested that assuming the association of a sentence with a single polarity is rarely true. A sentence containing both negative and positive sentiment should be considered neither negative nor positive. It is more logical to us that a neutral label would better describe the situation of such sentences. Minor errors resulted from the sentiment classification and are shown in Table 5.2.

Table 5.2: Examples of classification errors by Semantria.

Response	Sentiment Polarity	Comments
Never seen it before.	-0.493858993053436 (negative)	Even though the sentiment is close to the benchmark, we believe that this particular response is more neutral than negative.
I am doing it for a friend with Anklosing Anklosis.	0.660000026226043 (positive)	Again, this is a vague statement and we believe it should have been categorised as neutral.

ii. Collection Analysis

We used the collection analysis feature available by Semantria to find facets related to the survey. In other words, frequently-occurring words were identified along with a description of how many times they appeared with negative, neutral or positive sentiment in the text (see Appendix E). These were compared with the keywords identified by the human at the beginning of the study.

Table 5.3: Comparison between keywords identified by the human and Semantria.

Keywords identified by Human	Facets identified by Semantria
Website	Website
Guidelines	Site
Guidance	Guidelines
Design	Guidance
Appearance	Time
Navigation	Version
Usability	Page
Information	One
Access	Screen
Availability	People
Links	Use
Updates	Information
Search	Access
Improvements	Search
	Link

Table 5.3 indicates that similar keywords were identified by both the human and the tool. As expected, Semantria was able to detect facets that a human did not consider. However, these were based on the frequency of the words. As suggested by Kim, Li & Lee (2009), the frequency of a word is not always so important for sentiment analysis purposes and sometimes it is worth examining whether a word is present or not in the text.

iii. Categories

Semantria's 40 default categories were used to better understand the data, yet most of the categories suggested were irrelevant and not representative of the whole responses. We continued our analysis by keeping the relevant Semantria categories and created more

appropriate categories that contained keywords and facets previously outlined (Table 5.3). The final set of categories is shown in the tables below.

Table 5.4: Default categories.

Semantria <i>Relevant</i> Categories	Associated Responses
Advertising	2
Beverages	
Business	3
Education	2
Fashion	
Guidance	72
Hardware	1
Health	10
Mobile devices	1
Politics	
Science	1
Social media	
Software & Internet	7
Space	
Technology	4

Table 5.5: Custom categories.

Human Categories	Associated Responses
Usability: navigation, website, site, page, usability, screen	60
Design: website, site, page, appearance, visual, screen, style, layout	15
Guidance: direction, counselling, guidance, guidelines	59
Version: version, new, old, update, improve, familiarity, change	19
Information: search, information, link, access, result, available	58

As expected, categories created to meet the needs of the survey (Table 5.5) absorbed more results. However, the default category ‘*Guidance*’ in the Semantria tool resulted in 72 related responses in comparison with the 59 responses associated with the custom-made human ‘*Guidance*’ category. It is worth mentioning that some responses were associated with more than one category. This is because a respondent can have an opinion for different attributes of the website. Semantria seemed to be able to categorise the data relatively well, taking into consideration that categories had to be created to fulfil the needs of the data (healthcare field). Categories can be found useful to give a detail analysis of the text. Even though the

word ‘guidance’ appeared only 19 times in the text (see Appendix E), according to the collection analysis, 59 responses were associated with the ‘Guidance’ category after customisation. It is interesting to see that categories extract more information.

iv. Survey analysis profile

Using the categories identified above, a profile was created to assess the website’s performance in an alternative way.

Table 5.6: Results of the analysis based on the custom profile created.

Categories	Positive	Neutral	Negative	Observations and Concerns
Design	20	9	23	The design of the website seems to attract equal amounts of positive and negative reviews. A closer examination on the negative reviews would help identify specific problems.
Guidance	15	24	25	This was expected as a lot of negative feedback was found regarding the guidance provided in the webpage.
Information	19	18	21	Respondents had problems with the availability of information. Also, some felt that necessary information was missing.
Usability	19	12	29	Usability is an important feature as far as websites are concerned. From a website perspective, this is not good.
Version	5	5	9	No particular preference of the version. Negative feedback might be due to the sudden change and time may be needed to familiarise with the new website. Complaints about the change are inevitable.

Overall, Semantria provided an accurate sentiment analysis. However, a lot of configuration was required to adjust the tool in the healthcare information service field. For a commercial

product, this is not ideal as it depends a lot on human error. Nevertheless, configurations are a positive reinforcement to use Semantria even though cases, such as healthcare, are not supported by the system itself. Generally, it is very easy to use and gives results very quickly.

5.2.1.2 Sentiment analysis using only single-sentence responses

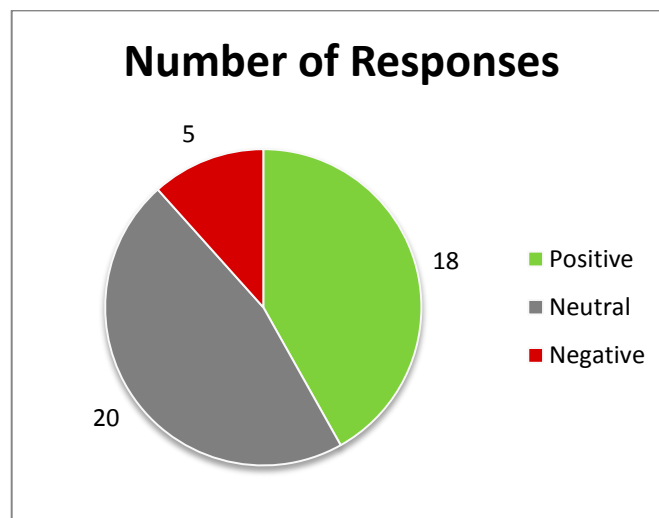


Figure 5.4: Classification results of single-sentence responses by Semantria.

Through the analysis, it was observed that some responses which were clearly positive (negative) in nature, were assigned a negative (positive) label by Semantria, or sometimes even a neutral label. Based on our assumption that single-sentence responses are more clearly associated with a single polarity, the examination of single-sentence responses could show the reason for these errors. The weaknesses of Semantria can be identified. Semantria was asked to categorize 43 single-sentence responses. The results are shown in Figure 5.4.

Surprisingly enough, negative responses are a minority in this sample (5 responses). This might suggest that people expressing negativity usually provide more analytical comments and hence there might be a relation between negativity and long responses. It might be interesting to

examine this further and maybe, with a larger dataset, establish new patterns. Also, neutral responses represent the majority of the classification results (20 responses). This was expected since single sentences are usually short texts that are more likely to state facts rather than opinions and thoughts. However, it was not expected that the difference between positive and neutral responses would not be noticeable. Examining this further, we concluded the following:

- Negativity is easily recognised. The human was able to confirm that all the negative labels were representative of the responses.
- Semantria's neutral class emphasizes on factual statements which can easily be found in single sentences. Hence the number of neutral responses dominates. However, there were some results (see Table 5.7) that caught our attention.

Table 5.7: Different examples of neutral responses using Semantria.

Response	Response Polarity	Comment
This is much better and easier to navigate directly to the vital information XXX provides.	0 (neutral)	A clearly positive response was categorized as neutral by Semantria. This is a comparative sentence, where the user compares the new version of the website to the old.
I looked for sarcoma and was disappointed to see the material was more than a year old.	0 (neutral)	In spite the factual element of the sentence, the user is “ <i>disappointed</i> ” with the website’s performance, hence an opinion was expressed.
I do like the new layout.	0.144063994 (neutral)	A clearly positive response
I am doing it for a friend with Anklosis Anklosing.	0.660000026 (positive)	A clearly neutral response

The errors that occurred during the examination of single-sentence responses were insignificant. It was observed that single-sentences are indeed more clearly associated with a specific polarity. But, the amount of single-sentences in our dataset was not enough to make a general statement that Semantria supports our assumption. This needs further investigation, with a larger dataset in place.

5.2.2 TheySay

5.2.2.1 Sentiment Analysis using all available responses

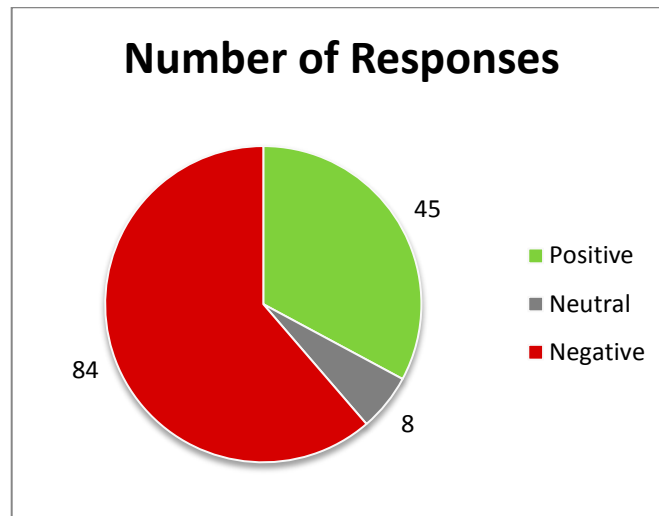


Figure 5.5: TheySay's classification results.

Looking at Figure 5.5, the classification results of TheySay differ a lot from the findings of Semantria (Figure 5.4). It seems that negativity and neutrality are classified differently by the two tools.

TheySay provides an in-depth analysis to sentiment. As we said before it considers document and sentence level sentiment. It also carries a POS analysis. A thorough examination of the tool's features resulted in interesting conclusions. Table 5.8 provides detailed feedback and observations for each feature used (with the exception of sentence and document sentiment).

Table 5.8: Examination of TheySay's features.

Feature	Feedback
Part of Speech Recognition	<p>Accurately recognises and defines a range of POS. This is extremely helpful for the academic literature of sentiment analysis. New patterns might be found.</p> <p>Our results support the general opinion that adjectives are the main form of sentiment but other parts of speech are equally important. For example, the phrase '<i>visually impaired</i>', when put into context, can affect the sentiment polarity.</p>
Comparison Detection	<p>Comparative expressions are recognised by the tool easily. However, in our survey responses comparisons between the old and the new version of the website are mentioned. The tool was not able to identify this kind of comparisons.</p>
Humour Detection	<p>Results were not very precise. But it seems to detect sarcasm and irony rather easily. For example:</p> <ul style="list-style-type: none"> I can't even see the whole of the first page without scrolling as your typeface is so LARGE. (score: 0.973) Do you think all users of the website are visually impaired? (score: 0.981) <p>Errors do exists of course:</p> <ul style="list-style-type: none"> I have just done a search for eczema and got back 5 pages of results. (score: 0.84)
Speculation Analysis	<p>A good feature to discover common issues presented in the dataset. It has several categories of speculation, one of them describing what the speaker/writer wants. This is named 'SPECULATION.WANTING' by TheySay and can be found extremely useful for gaining consumer insight.</p>
Risk Analysis	<p>These two features were not particularly popular through our analysis so no comments can be said.</p>
Intent Analysis	

Feature	Feedback
Named Entity Recognition	It is not very helpful since it only provides a list of all entities and terms mentioned in the text. Entities are indeed identified but in a very confusing manner.
Shallow Chunk Parsing	As already explained, Shallow Chunk Parsing determined phrases found in the text. Through the analysis, the performance of this feature was accurate.
Dependency Parsing	It is somewhat similar to Shallow Chunk Parsing as it mainly connects words that are dependent to each other. Most of the times these combinations form a phrase.

A couple of issues were observed during the analysis:

- TheySay appeared to be case-sensitive. A response included the word '*cant*' without the apostrophe. This is common in informal style responses such as survey responses. TheySay was not able to identify the negativity associated with the particular response. Adding the needed apostrophe changed the overall classification results.

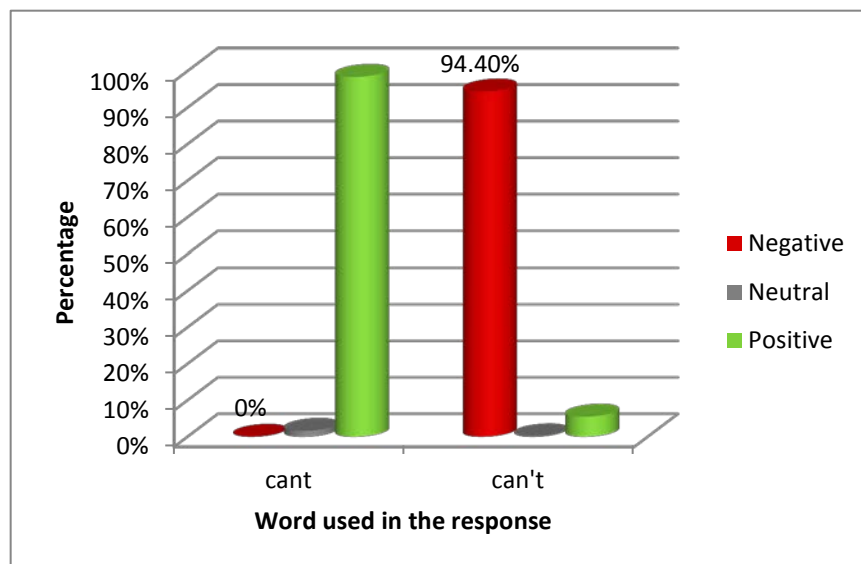


Figure 5.6: Results for response "*cant find guidance...*" to examine the case sensitivity of TheySay.

- Some responses were inputted into the system twice for clarification. A different result occurred the second time for a few of them pointing out that results may be inconsistent. An example is shown in Figure 5.7. However, it is worth mentioning that these minor inconsistencies were not enough to establish a serious problem of the tool.

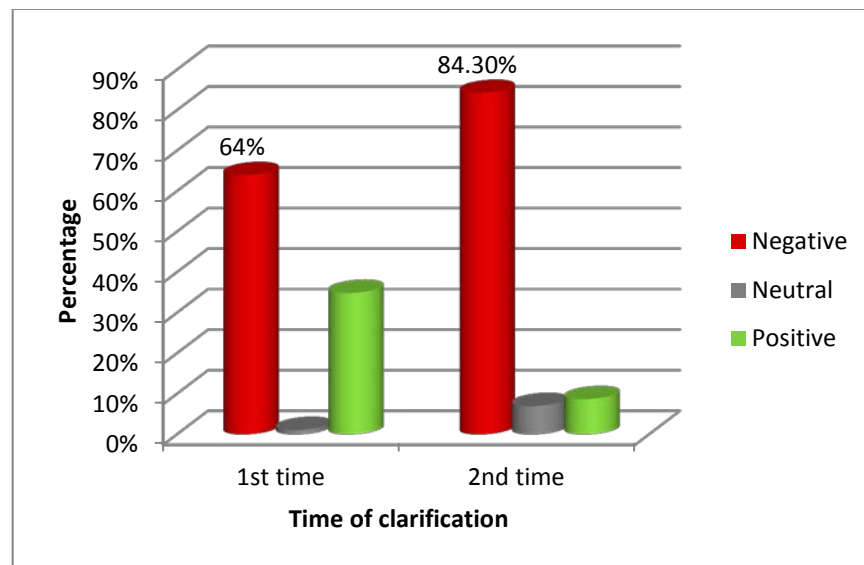


Figure 5.7: Observing errors during verification/clarification.

- Generally, the three percentages provided represent a more realistic analysis as most reviews, which are where sentiment is found, contain a mixture of negative and positive sentiment. This is quite helpful in analysing, but can sometimes be confusing. For example, the clearly negative response *“USELESS. Search box doesn’t work & half of the links don’t function.”* received a classification result of 64.30% negative, 4.20% neutral and 31.50% positive, while no positivity was found.
- It was observed that single-sentence responses were clearly described by only one label. This supports our assumption. However, it must be noted that the sample size

used for single-sentence analysis was much smaller and thus not sufficient for generalisation.

- The percentage of neutrality found in each response was usually associated to punctuation features. Overall, only 8 responses were assigned a neutral label (Figure 5.5), i.e. received a score of 100% in favour of the neutral class. The rest neutral percentages were ranged from 0-25%.

5.2.2.2 *Sentiment analysis using only single-sentence responses*

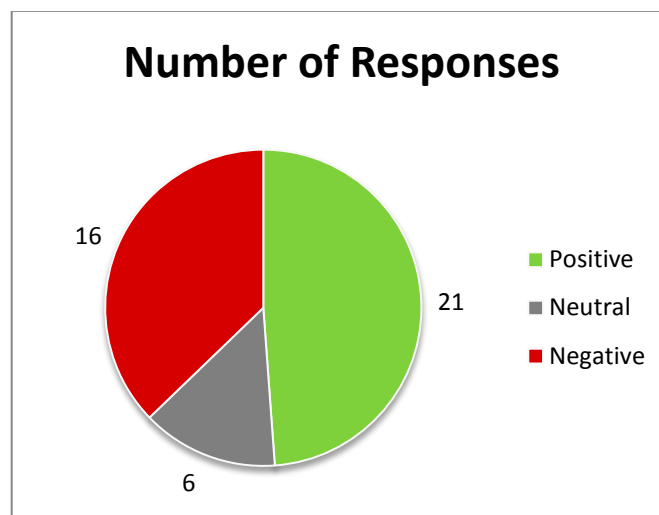


Figure 5.8: Classification results of single-sentence responses by TheySay.

The same set of single sentences was classified by TheySay. The results are shown in Figure 5.8. As expected, the neutral class represents the minority of responses. Furthermore, most of the responses (38 out of the 43 responses) were clearly associated with one of the three classes by having a percentage of 65% or above for a specific classification. This could provide supportive evidence for our assumption. Despite some negligible errors (see Table 5.9), sentiment classification was accurately predicted for single-sentence responses.

Table 5.9: Errors found in single-sentence responses when analysed by TheySay.

Response	Negative	Neutral	Positive
I am looking for a specific guideline but it keeps telling me the guidelines have been moved - and I can't find them!!	0%	2.90%	87.10%
I do like the new layout.	0%	100%	0%

5.2.3 Comparison between Commercial Tools

TheySay and Semantria were chosen to represent the commercial products category. The classification results, point out the difference between each tools' internal algorithm. It seems that negativity and neutrality are classified differently by the two tools. Semantria showed a clear tendency in assigning a neutral label to the responses; 51 out of the 137 were characterised as neutral (Figure 5.4). However, the negative class (48 responses) and positive class (39 responses) indicate that Semantria has a balanced classification algorithm. On the other hand, TheySay favoured the negative class by assigning a negative label to 84 responses. The positive class followed with 45 responses and the neutral class with only 8.

We investigated these classification differences in order to better understand the functions of the algorithm used by each tool. Table 5.10 shows a summary of this investigation and essentially acts as a comparative feature between the commercial tools.

Table 5.10: Comparison between Commercial Tools

Semantria	TheySay
<ul style="list-style-type: none"> • Emphasize on words to extract sentiment. • Due to the above, when a spelling mistake is in place, Semantria is not able to ignore it and focus on the overall meaning of the statement. • Problem with adverbs and adjectives-related sentiment. • It is more keen in identifying neutral responses, and hence in recognising suggestions. • The above leads in sometimes underestimating the non-obvious negative responses. • In some cases, Semantria, fails to see the sentiment and opinion held and tends to classify the response as neutral. • Phrase ignorance: lacks knowledge of important phrases that could identify the class of each response. 	<ul style="list-style-type: none"> • Looks at the document as a whole and hence it is able to better understand the respondent and produce a more accurate classification. • Due to the above, when a spelling mistake is in place, TheySay is able to ignore it and focus on the overall meaning of the statement. • More likely to understanding the human tone of voice; sarcasm and irony are relatively easy to identify. • TheySay gets confused with large, explanatory responses. • It certainly does not emphasize on words in the document. • It is able to identify irony and sarcasm most of the time, as well as hidden frustration. • It can identify non-direct negativity.

Overall, TheySay appears to have more powerful features that could result in a more accurate prediction. A closer look at the percentage of “correctly” classified responses might produce a clearer picture of the abilities of each tool. Assuming that the human labels are mostly correct, as they were extracted using the quantitative survey questions, they were be used as a baseline for verification. Further statistical analysis was undertaken to ensure the reliability of the commercial tools (see Appendix F). These were also used in Section 6, where a comparative evaluation of the four tools was conducted.

Table 5.11: Correctly Classified Responses - Commercial Tools

	Semantria	TheySay
Correctly Classified Responses	70	94
Total Responses	137	137
Percentage of Correctly Classified Responses	51.09%	68.61%
Kappa Statistic	0.2627	0.3886

It can be seen from Table 5.11 that TheySay's performance is slightly better (68.61%). This might suggest that the internal algorithm is more efficient than the one used by Semantria, or, simply that the inclusion of some more academic features (e.g.: POS, humour detection, etc.) give the advantage to TheySay. The kappa statistic, for both tools, indicates that there is a fair agreement between the predictions and the baseline. There is still a lot of space for improvement to achieve a near-perfect agreement.

Comparing the predictions of the commercial tools with the baseline provided, we constructed two confusion matrices for Semantria and TheySay, found in Table 5.12 and 5.14 respectively. These were used to calculate the Cohen's kappa statistic for the commercial tools via the online calculator (Vassstats, 2012) which is shown in Table 5.11.

Table 5.12: Confusion Matrix for Semantria.

		<i>Predicted Class</i>		
		Negative	Neutral	Positive
<i>Actual Class</i>	Negative	42	38	15
	Neutral	4	9	5
	Positive	1	4	19

The majority of each class was correctly predicted, as shown in Table 5.12. It can be observed that half of the neutral responses were incorrectly classified as non-neutrals, while the 38 negative responses wrongly predicted as neutral, support our previous observation. It seems that Semantria favours the neutral class.

Table 5.13: Assessment of the reliability of the algorithm used by Semantria.

	Negative	Neutral	Positive	Weighted Average
True Positive (TP) Rate	0.442	0.500	0.792	0.511
False Positive (FP) Rate	0.589	0.353	0.177	0.486
Precision	0.894	0.177	0.487	0.729
Recall	0.442	0.500	0.792	0.511
F-Measure	0.592	0.261	0.603	0.550

The high TP rate for positive classification was expected since 19 out of the 24 responses were correctly predicted. Also, the reverse relationship existing between precision and recall is clear. While Semantria is able to accurately identify the responses belonging to the negative class (precision = 0.894), it was not able to label all the responses from the negative class as belonging to the negative class (recall = 0.442).

Conversely, Negativity is easily recognised by TheySay as shown in the confusion matrix (Table 5.14) and hence received a high TP rate (0.758), along with a high precision (0.868) and recall (0.758). However, the low values associated with the neutral class (Table 5.15) indicate the difficulty in correctly identifying neutrality.

Table 5.14: Confusion Matrix for TheySay.

		<i>Predicted Class</i>		
		Negative	Neutral	Positive
<i>Actual Class</i>	Negative	72	4	19
	Neutral	7	3	8
	Positive	4	1	19

Table 5.15: Assessment of the reliability of the algorithm used by TheySay.

	Negative	Neutral	Positive	Weighted Average
True Positive (TP) Rate	0.758	0.167	0.792	0.686
False Positive (FP) Rate	0.324	0.042	0.239	0.272
Precision	0.868	0.375	0.413	0.724
Recall	0.758	0.167	0.792	0.686
F-Measure	0.809	0.231	0.492	0.678

5.2.3.1 Examination of single-sentence responses

The main difference between the two tools is the categorisation of neutral and negative responses as already pointed out. The results shown in Figure 5.9 support our previous observation that Semantria is very keen in assigning neutral labels. However, there were several issues with sentiment extraction that may contribute to the cause of this excessive neutral categorization.

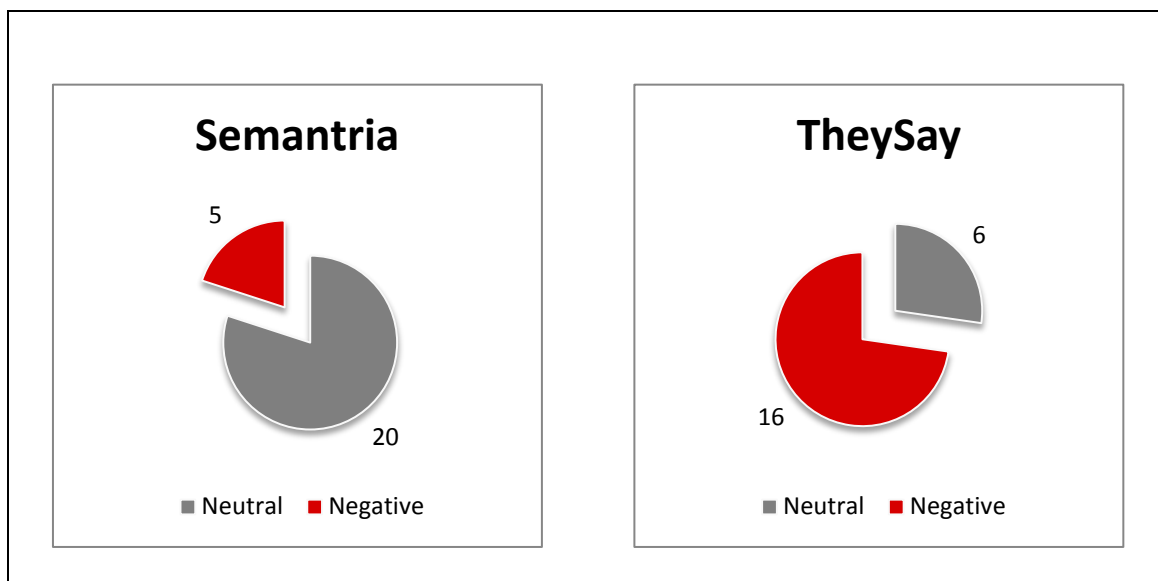


Figure 5.9: Single-sentence responses classified by Semantria and TheySay.

Furthermore, the assumption made from Semantria’s findings that negativity tends to appear more often in long sentences was overruled by TheySay. Examining this further might be useful.

Table 5.16: Correctly classified single-sentence responses – Commercial Tools.

	Semantria	TheySay
Correctly Classified Responses	23	31
Total Responses	43	43
Percentage of Correctly Classified Responses	53.49%	72.09%

According to this project, TheySay is the leading tool for sentiment classification in a healthcare domain. This was of course concluded by comparing it with Semantria only. A classification accuracy of 72% indicates the efficiency of the algorithm used. Furthermore, it

supports our assumption that the usage of only single sentences could provide improvements in performance.

5.3 Non-Commercial Tools

5.3.1 Google Prediction API

5.3.1.1 *Sentiment analysis using all the available responses*

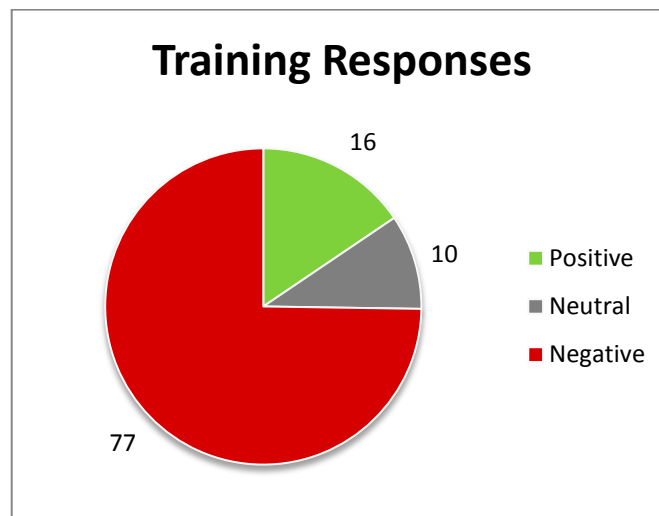


Figure 5.10: Total number of positive, neutral and negative responses used for training the model via Google Prediction API. They represent the 75% of the dataset.

Figure 5.10 represents the different class values associated with the 75% of the dataset that was used for training the model. There was a clear majority of negative responses, almost two and a half as much as the positive and the neutral together. Therefore, we constructed a new model that contained equal number of responses in each class (10 each) and investigated the differences on the classifier's performance (see Table 5.17).

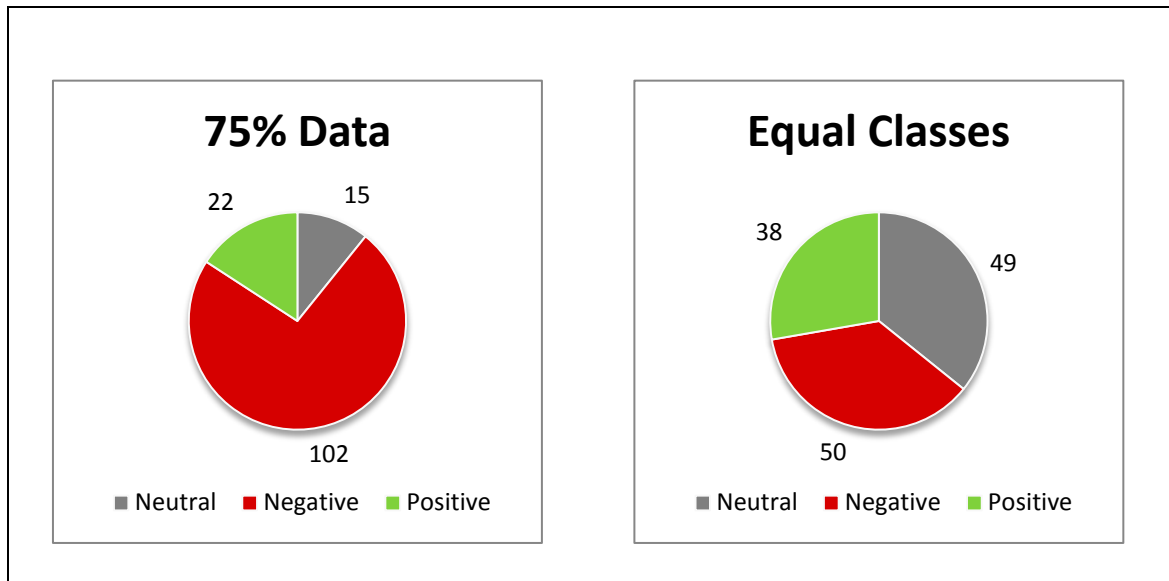


Figure 5.11: Classification results using the whole dataset. Google Prediction API was tested on the 25% of unseen responses in addition to the already trained responses.

Table 5.17: Examination of the different training samples used for sentiment analysis using Google Prediction API.

Statistical Evaluator	Results	
	75% Data Training Set	Equal Classes Training Set
Correctly Predicted Responses	89.78%	51.10%
Incorrectly Predicted Responses	10.22%	48.90%

As expected, using 75% of the data to train the model improved the accuracy of the predictions. Theoretically, equal classes should have performed better but the number of responses used for training was not enough to create an accurate model. Using 75% of the dataset for training is a common technique.

Since Google Prediction API does not provide a cross validation feature, this was done manually. A 4-fold cross validation was chosen. This meant that the data were separated into four equal parts. Four training models were created to perform cross validation. For each

model, three folds were used for training and one was reserved for testing. Details of the training folds can be seen in Table 5.18.

Table 5.18: Cross validation training sets.

	Train 1	Train 2	Train 3	Train 4
Fold 1 (34 responses)	√	√	√	
Fold 2 (34 responses)	√	√		√
Fold 3 (34 responses)	√		√	√
Fold 4 (35 responses)		√	√	√

Results of the four training models for cross validation purposes are shown below. The number of negative, neutral and positive responses used for training is identified in each table, along with the results after testing. The percentage of correctly classified responses was calculated.

Table 5.19: Details of ‘Train 1’ model.

Train 1			
Negative	Neutral	Positive	
68	14	20	
9 out of the 35 responses (FOLD 4) were wrongly predicted by Google Prediction API when model ‘Train 1’ was in use.			74.3% correctly classified responses

Train 1 was trained using Folds 1-3 and was tested using Fold 4.

Table 5.20: Details of 'Train 2' model.

Train 2			
Negative	Neutral	Positive	
73	13	17	
10 out of the 34 responses (FOLD 3) were wrongly predicted by the Google Prediction API when model 'Train 2' was in use.			70.6% correctly classified responses

Train 2 was trained using Folds 1, 2, 4 and was tested using Fold 3.

Table 5.21: Details of 'Train 3' model.

Train 3			
Negative	Neutral	Positive	
71	13	19	
11 out of the 34 responses (FOLD 2) were wrongly predicted by Google Prediction API when model 'Train 3' was in use.			67.6% correctly classified responses

Train 3 was trained using Folds 1, 3, 4 and was tested using Fold 2.

Table 5.22: Details of 'Train 4' model.

Train 4			
Negative	Neutral	Positive	
73	14	16	
8 out of the 34 responses (FOLD 1) were wrongly predicted by Google Prediction API when model 'Train 4' was in use.			76.5% correctly classified responses

Train 4 was trained using Folds 2-4 and was tested using Fold 1.

Most of the negative responses were accurately predicted, which indicates the accuracy associated with Google Prediction API. It was expected that neutral and positive responses would be predicted with less accurately, since the training set did not contain enough examples for the system to learn and identify them correctly. Negative responses were significantly more than the others. Overall, Google Prediction API uses a precise learning

algorithm since it was able to accurately predict approximately 72% of the responses, according to the cross validation results. This was calculated using Equation 2 (Section 3.2.1) as shown below.

$$Total\ Accuracy = \frac{74.3 + 70.6 + 67.6 + 76.5}{4} \times 100 = 72.25\% \quad Equation\ 9$$

A statistical analysis was made in order to further examine the reliability of the model and can be found in the following tables. An online calculator (Vassarstats, 2012) was used to find the Cohen's kappa statistic value based on the confusion matrix.

Table 5.23: Statistical evaluation of the model built in Google Prediction API (based on cross validation).

Statistical Evaluator	Results
Correctly Classified Responses	72.25%
Incorrectly Classified Responses	27.75%
Kappa statistic	0.2199

While the classifier seems to be relatively accurate (72.25%), the kappa statistic is surprisingly low (0.2199). This indicates that agreements found between the classifier and the baseline, are biased and it can be assumed that they are likely to have resulted by chance. It is possible that the software memorised the data.

Table 5.24: Confusion Matrix for Google Prediction API.

		<i>Predicted Class</i>		
		Negative	Neutral	Positive
<i>Actual Class</i>	Negative	89	3	3
	Neutral	16	2	0
	Positive	17	0	7

Table 5.25: Assessment of the reliability of the model.

	Negative	Neutral	Positive	Weighted Average
True Positive (TP) Rate	0.937	0.111	0.292	0.716
False Positive (FP) Rate	0.786	0.025	0.027	0.553
Precision	0.730	0.400	0.700	0.681
Recall	0.937	0.111	0.292	0.716
F-Measure	0.820	0.174	0.206	0.628

The confusion matrix in Table 5.24 provides an alternative way to assess the model's performance when tested on unseen data, i.e. on the 137 responses. Details on how these values were computed can be found in Appendix G. The prediction results support the assumption resulting from the kappa statistic (Table 5.23). It can be seen that almost all negative responses were predicted as negative, while neutral and positive classes suffered from clarity. This led us to believe that Google Prediction API may have memorised some of the responses and that majority class (negative class) was favoured. The majority of the responses from the neutral and positive class were categorised as negative after training. This can also be seen from the low recall values (Table 5.25). Knowledge of the classifier used could provide some insight for these conclusions but, as already said, this is not presently available.

The model is considered to be relatively precise (F-measure = 0.628). However, the low recall values of neutral and positive responses, led us to believe that the software was not able to learn neutral and positive classification. It is possible, that neutral and positive labels occurred by chance since the classifier did not find any indication of negativity. This can also be observed by the TP rate of neutral and positive responses.

5.3.1.2 Sentiment analysis using only single-sentence responses

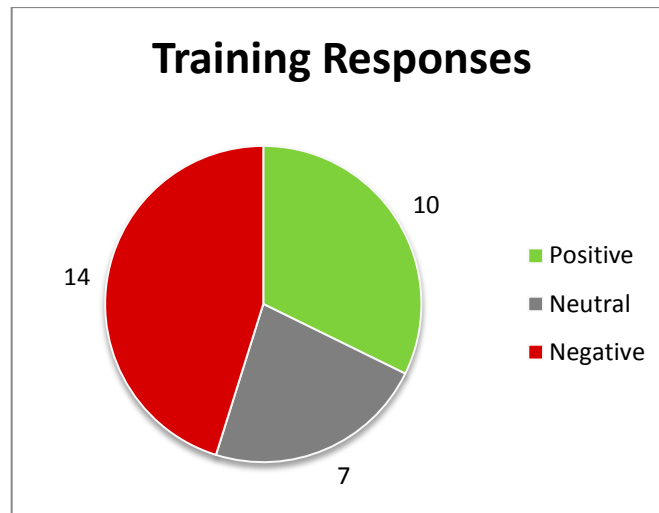


Figure 5.12: Number of single- sentence responses used for training purposes for Google Prediction API.

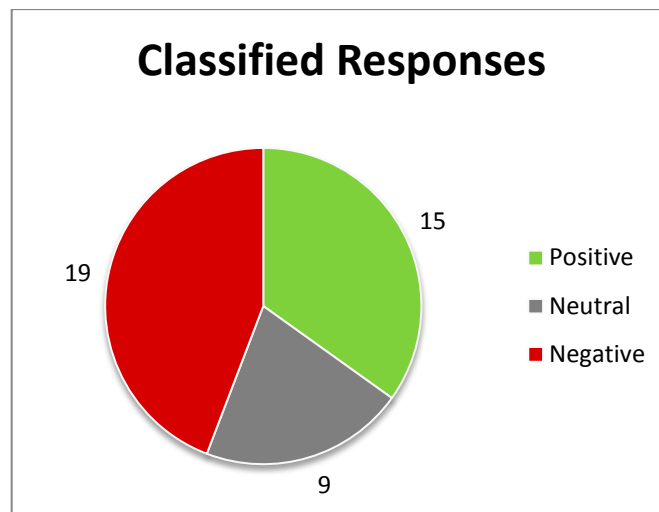


Figure 5.13: Results of single-sentence classification analysis by Google Prediction API. The 75% of single sentences was used for training and 25% was used for testing.

The responses containing only single sentences were used (same as for the commercial tools). Again, 75% of the data was used for training (Figure 5.12) and the remaining 25% was reserved for testing. We did not consider equal classes as it was proven unsuccessful and no extra value was added to our analysis. Moreover, the number of responses would not be representative of

the sample and results could not be generalized to the healthcare information service in general.

Table 5.26: Statistical evaluation of Google Prediction API's results when tested on unseen data (testing set).

Statistical Evaluator	Results
Correctly Classified Responses	91.67%
Incorrectly Classified Responses	8.33%

Looking at Table 5.26, 91% of the data was correctly predicted. We wanted to verify this result by validating the model. Cross validation was again done manually due to the limitations faced by Google Prediction API. As before, four training sets were created, each trained using three folds and tested on the remaining one. The difference here is the number of responses associated with each fold. Table 5.27 shows a new representation of the folds.

Table 5.27: Details of the four folds for single-sentence responses analysis

Fold 1	Fold 2	Fold 3	Fold 4
10 responses	10 responses	11 responses	12 responses

The results of the cross validation are shown below in the same format as before.

Table 5.28: Details of 'Single Train 1' model.

Single Train 1			
Negative	Neutral	Positive	
14	7	10	
8 out of the 12 responses (FOLD 4) were wrongly predicted by Google Prediction API when model 'Single Train 1' was in use.			33.3% correctly classified responses

Single Train 1 was trained using Folds 1-3 and tested using Fold 4.

Table 5.29: Details of 'Single Train 2' model.

Single Train 2			
Negative	Neutral	Positive	
15	5	11	
3 out of the 11 responses (FOLD 3) were wrongly predicted by the Google Prediction API when model 'Single Train 2' was in use.			72.7% correctly classified responses

Note: Two responses were given equal number of positive, negative and neutral classification.

Single Train 2 was trained using Folds 1, 2, 4 and was tested using Fold 3.

Table 5.30: Details of 'Single Train 3' model.

Single Train 3			
Negative	Neutral	Positive	
12	8	12	
5 out of the 10 responses (FOLD 2) were wrongly predicted by Google Prediction API when model 'Single Train 3' was in use.			50% correctly classified responses

Single Train 3 was trained using Folds 1, 3, 4 and was tested using Fold 2.

Table 5.31: Details of 'Single Train 4' model.

Single Train 4			
Negative	Neutral	Positive	
16	7	9	
4 out of the 10 responses (FOLD 1) were wrongly predicted by Google Prediction API when model 'Single Train 4' was in use.			60% correctly classified responses

Single Train 4 was trained using Folds 2-4 and was tested using Fold 1.

$$\text{Total Accuracy} = \frac{33.3 + 72.7 + 50 + 60}{4} \times 100 = 54\%$$

Equation 10

Google Prediction API showed no improvement when only single-sentence responses were used for training. Based on our assumption, single sentences were expected to be associated with single sentiment classification. However, looking at the results of the Google Prediction API, with a 54% of correct classification, it can be seen that this was not the case here. Google uses supervised machine learning, and this might be the reason for the decrease in performance when dealing with single sentences. In the case of machine learning, more information is better for the system since it has more data to train on and hence 'learn' to classify more accurately. It is possible that the single sentences caused the system to undertrain.

5.3.2 WEKA

5.3.2.1 Sentiment analysis using all the available responses

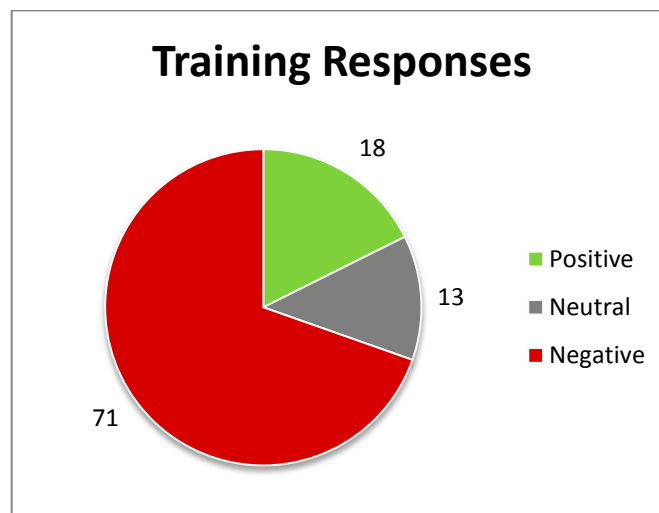


Figure 5.14: The three classes used for the training along with the number of responses contained in each of them. They represent the 75% of each class.

For analysis purposes, the data were separated into training (75%) and testing (25%) data. Each set of data was created by using the appropriate percentage of positive, neutral and negative responses.

Experimenting with the features provided by WEKA, an ideal combination was identified that would improve the classifier's overall performance and essentially increase the accuracy of prediction. First, the data were pre-processed and later on resampling was considered due to the anomalies that existed among the three different classes (negative responses were more in number than the other two types of responses). Resampling showed significant improvements for the classifier's performance; precision was increased by an average of 10% and accuracy by an average of 6%. A 4-fold cross validation was used to verify the model using the Naïve Bayes classifier. Even though MaxEnt and SVM are considered a more suitable fit for sentiment analysis, they are not available in the WEKA software. Due to time limitations, it was not possible to include new classifiers in the system. Nonetheless, Naïve Bayes performed rather well. It would be interesting to examine the differences between the three classifiers on data from the healthcare information service domain.

i. Pre-processing: String to word vector/Feature Selection

Overall, almost all the combinations of features resulted to an accuracy percentage above 70%, which is very good. Appendix H shows the different combinations of features we experimented with, along with a summary of their results upon classification. This was done as a means to test the literature and identify the most appropriate pattern for the healthcare information service domain. However, it is worth mentioning that the size of our dataset does not allow for generalization of the results, but rather it acts as a guideline to further research on this particular domain. Option 2 stands out as it is very close to 80% of accuracy. It was therefore concluded that the ideal combination for our healthcare dataset included the

removal of capital letters, punctuation and stopwords. Alphabetically tokenizing the data was also considered to compensate for any failure in punctuation removal.

*ii. **Experimentation: Resampling and Attribute Selection***

WEKA enables the selection of different algorithm combinations to assist the classifier in having a more accurate judgement. In this report, we experimented with two combinations:

- **CFS Subset (CFS) and Greedy Stepwise algorithms:** CFS subset evaluator assesses the worth of a subset of attributes, by considering individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low inter-correlation are preferred. Combining it with the greedy stepwise algorithm, we can distinguish the specific attributes which are more responsible for the classification result. Greedy stepwise performs a greedy forward or backward search through the attributes and stops only when the evaluation results experience a decrease.
- **Information Gain (InfoGain) Attribute and Ranker algorithms:** InfoGain determines the significance of each individual variable that contributed to the classification result. Combining it with the ranker algorithm, we can investigate the POS classification along with different patterns of sentiment classification.

As it may have been observed, the dataset used for this project report lacks in size. During the analysis, resampling the data was studied. The aim of WEKA's resample technique is to over-sample the minority class and under-sample the majority class, to produce a more balanced dataset. Therefore, a more accurate and efficient model would be created. In order to avoid the loss of information WEKA was instructed to use 100% of the data to create uniform classes.

Each feature selection combination (Options 1-7) was tested using the two different attribute selection algorithms to examine possible improvements on the model. Afterwards, resampling was considered to adjust the extreme differences between the class values. As it can be shown in Table 5.32, our model performs better when Option 1, Option D and resampling are present in the analysis; the accuracy increased by 11.77% compared to 77.45% of Option 2. However, it was decided not to consider resampling at all due to the rather small size of the dataset. Therefore, we concluded to the next appropriate model for our project based on the accuracy percentages, i.e. feature selection Option 2 along with the CFS Attribute Selection with Greedy algorithm. As expected attribute selection algorithms complemented the feature selection option.

Table 5.32: Attribute and resampling experimentation

String to Vector		Attribute Selection			Resample afterwards (uniform data)
	Result		Best Selection	Result	Result
Option1	71.5686%	Option A	CFS – Greedy	80.39%	73.53%
		Option B	INFOGAIN – Ranker	71.57%	89.22%
Option 2	77.451%	Option C	CFS – Greedy	82.35%	73.53%
		Option D	INFOGAIN – Ranker	77.45%	80.39%
Option 3	69.6078%	Option E	CFS – Greedy	75.49%	77.45%
		Option F	INFOGAIN – Ranker	69.61%	78.43%
Option 4	67.6471%	Option G	CFS – Greedy	72.55%	71.57%
		Option H	INFOGAIN – Ranker	64.65%	83.33%
Option 5	71.5686%	Option I	CFS – Greedy	70.59%	76.47%
		Option J	INFOGAIN – Ranker	71.57%	81.37%
Option 6	70.5882%	Option K	CFS – Greedy	71.57%	76.47%
		Option L	INFOGAIN – Ranker	70.59%	80.39%
Option 7	73.5294%	Option N	CFS – Greedy	78.43%	75.49%
		Option M	INFOGAIN – Ranker	73.53%	77.45%

iii. Analysis of final modelling

Further examination was undertaken to identify potential improvements and inefficiencies that may have occurred to the final model. A statistical analysis can be found in the following tables. These were provided by the WEKA tool as a result of the training.

Table 5.33: Assessment of final model's performance using statistical measures.

Statistical Evaluator	Result
Correctly Classified Instances	82.3529%
Incorrectly Classified Instances	17.6471%
Kappa statistic	0.5735
<i>Mean absolute error (MAE)</i>	<i>0.162</i>
<i>Root mean squared error (RMSE)</i>	<i>0.2981</i>
<i>Relative absolute error (RAE)</i>	<i>51.0693%</i>
<i>Root relative squared error (RRSE)</i>	<i>75.3984%</i>
<i>Coverage of cases (0.95 level)</i>	<i>98.0392%</i>
<i>Mean relative region size (0.95 level)</i>	<i>58.1699%</i>

The statistical measures above indicate the good accuracy and precision of the model and the classifier. The kappa statistic with a value of 0.6169 shows a substantial agreement between the classifier and the ground truth, i.e. the baseline set by the human labelling based on the quantitative results of the survey. Moreover, the low values of the MAE (0.162) and RMSE (0.2981) support the kappa statistic and indicate that the model is relatively accurate and precise. The small difference between them is insignificant and hence no variation in individual errors can be suggested. Also, the low values of RAE and RRSE represent the same. The narrow confidence interval (0.95 level) shows a stable model which indicates the reliability of the estimate values of the classifier.

Table 5.34: Confusion Matrix for WEKA.

		<i>Predicted Class</i>		
		Negative	Neutral	Positive
<i>Actual Class</i>	Negative	67	1	3
	Neutral	8	5	0
	Positive	6	0	12

The confusion matrix shown in Table 5.34 shows the performance of the algorithm used to train the classifier. It shows the actual value of each class and what was the predicted output of the model. For example, of the total of 71 actual negative responses consisted in the training set, the system predicted that one was neutral and three were positive. It is interesting to observe that neutral responses were never wrongly classified as positive and vice versa. The fact that more than half of the neutral responses were classified as negative by the system demonstrates the problem with neutral responses once again.

Table 5.35: Assessment of final model's reliability.

	Negative	Neutral	Positive	Weighted Average
True Positive (TP) Rate	0.944	0.385	0.667	0.824
False Positive (FP) Rate	0.452	0.011	0.036	0.322
Precision	0.827	0.833	0.800	0.823
Recall	0.944	0.385	0.667	0.824
F-Measure	0.882	0.526	0.727	0.809
MCC	0.560	0.529	0.679	0.577
ROC Area	0.894	0.883	0.930	0.899
PRC Area	0.950	0.551	0.792	0.871

Observing the weighted average of the evaluation indicators in Table 5.35, it is suggested that the model is precise and accurate with a high F-value (0.809). This can also be explained with the high value of the TP rate which indicated the number of correctly classified instances. In general, the results above indicate a good model. However, it can be seen that negative classification is more accurate and precise (0.827) than the other two classifications. This was expected since the majority of responses were negative and hence the classifier was trained on more negatives than positives or neutrals. Moreover, neutral classification is rather weak compared with the other two, especially with the positive classification which is also a minority class. As suggested in the literature of this report, classifying and distinguishing neutrality is a difficult concept. The dataset used for this project report did not help in overcoming this problem as, the neutral responses was not more than 18 out of the total of 137 responses, and only 13 were used for training.

5.3.2.2 Sentiment analysis using only single-sentences responses.

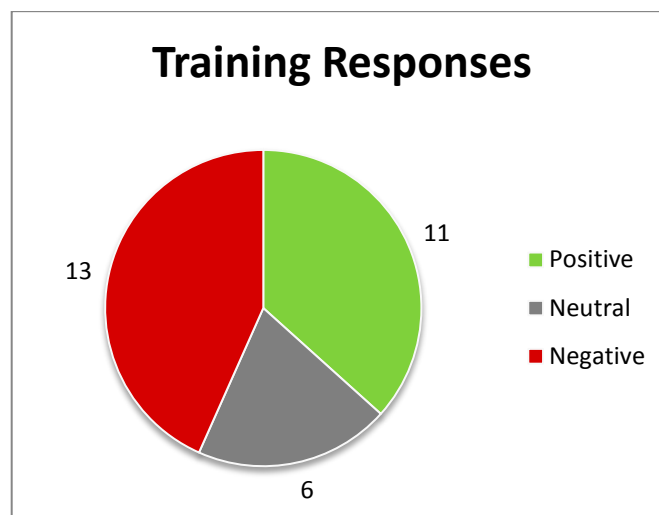


Figure 5.15: Number of single- sentence responses used for training purposes for WEKA.

The data were separated into training (75%) and testing (25%) sets by following the same procedure as before. Figure 5.15 represents the 75% training set.

After the extensive experimentation with all the data, an ideal combination of features and attributes was defined. For comparison purposes, we examined the performance of the single-sentence responses model using only that, defined combination. Results are shown in Table 5.36. However, for analysis purposes, we investigated some alternative combinations and whether they can improve the accuracy of the model (see Table 5.38).

Table 5.36: Correctly classified instances using Naïve Bayes classifier with 4-fold cross validation: attribute selection & resampling.

	Feature Selection	Attribute Selection	Resampling (uniform data)
Result	66.6667%	70%	80%

Table 5.37: Correctly classified instances using Naïve Bayes classifier with 4-fold cross validation: resampling & attribute selection.

	Feature Selection	Resampling (uniform data)	Attribute Selection
Result	66.6667%	73.3333%	86.6667%

As it can be seen from the two tables above, the model's performance has slightly decreased. Using all the available responses, the classifier reached 79.12% accuracy (Appendix H) before considering any attributes or resampling. Even though it was concluded that resampling methods are not suggested for our sample size, we experimented with the order of attributes. Interestingly enough, resampling before selecting attributes necessary for the analysis (Table 5.37) improves the model's accuracy by 5%. This is an interesting discovery that could be used with bigger set of data.

With an approximately 10% decrease in the accuracy of sentiment classification when training using only single-sentence responses, it can be safely said that examining single-sentence responses did not improve the learning of the WEKA tool. However, 70% accuracy is not bad and the results could be different if the sample was bigger in size.

Table 5.38: Experimentation with different attribute combinations along with resampling.

String to Vector		Attribute Selection			Resample afterwards (uniform data)
	Result		Best Selection	Result	Result
Option1	66.7%	Option A	CFS – Greedy	76.7%	73.3%
		Option B	INFOGAIN – Ranker	66.7%	66.7%
Option 2	66.7%	Option C	CFS – Greedy	70.0%	80.0%
		Option D	INFOGAIN – Ranker	66.7%	73.3%
Option 3	73.3%	Option E	CFS – Greedy	63.3%	70.0%
		Option F	INFOGAIN – Ranker	73.3%	80.0%
Option 4	73.3%	Option G	CFS – Greedy	63.3%	70.0%
		Option H	INFOGAIN – Ranker	73.3%	83.3%
Option 5	63.3%	Option I	CFS – Greedy	56.7%	63.3%
		Option J	INFOGAIN – Ranker	63.3%	66.7%
Option 6	60.0%	Option K	CFS – Greedy	56.7%	63.3%
		Option L	INFOGAIN – Ranker	60.0%	66.7%
Option 7	63.3%	Option N	CFS – Greedy	56.7%	63.3%
		Option M	INFOGAIN – Ranker	63.3%	63.3%

The positive effects of resampling methods are very clear when analysing single-sentence responses. It could be concluded that single-sentence responses can provide accuracy improvement if the sample is large enough. Resampling was used to create a uniform sample with the 43 single-sentence responses we had in place and even though the sample size was small, resampling showed significant improvement.

5.3.3 Comparison between Non-Commercial Tools

Overall, we tried to maintain a consistent format throughout training (Figures 5.10 and 5.14). Both non-commercial tools use machine learning algorithms and perform a supervised learning. They required pre-labelled data which had to be done manually. More details for each tool are given in the table below, where the difference between the two systems can be more easily identified.

Table 5.39: Comparison between non-commercial tools.

Google Prediction API	WEKA
<ul style="list-style-type: none">• Data must be converted to an acceptable format so Google Prediction API can read and process them. This is done manually.• No configuration is allowed.• Manual cross validation is required.• There is no choice in the classifier used for training. It is pre-defined by the tool.• It can account for negation and can be trained to learn phrases.	<ul style="list-style-type: none">• Data must be converted to an acceptable format so WEKA can read and process them. This is done by the tool itself.• Pre-processing the data is available. Feature and attribute selection mechanism exist.• Cross validation is included in the classification. The number of folds can be adjusted according to the needs of the project.• A large range of classifier to choose from. However, SVM and MaxEnt are not present.• We found no way to account for negation.• Graphical representation of the results can be produced through the tool.

A comparison between the accuracy of each tool is a good measure to assess sentiment classification in general and as far as healthcare information service domain is concerned. Table 5.40 shows the accuracy of predictions by each non-commercial tools.

Table 5.40: Accuracy of the classifier for non-commercial tools.

Responses Used	Google Prediction API	WEKA
All responses included	72%	82.35%
Only single-sentence responses included	54%	70%

Further statistical comparisons can be found in Table 5.41 for the classification results when all the responses were included in the analysis. Findings were based on the 4-fold cross validation.

Table 5.41: Comparison between non-commercial tools using statistical measures.

Statistical Measure	Google Prediction API	WEKA
Kappa Statistic	0.2199	0.5735
Precision Weighted Average	0.681	0.823
Recall Weighted Average	0.716	0.824

Despite the fact that WEKA had clearly performed better than Google API judging from the high precision and recall weighted averages, Google Prediction API's performance was not far behind. A relatively high recall (0.716) indicates that a large number of the responses were indeed categorised in the relevant class. However, the low kappa statistic suggests that this might be a result of chance and not due to tool's prediction performance. Also, this was followed by 0.681 average on precision, which pointed out the errors associated with training; there were responses that were given an incorrectly label.

6 EVALUATION OF SYSTEMS WITH RESPECT TO THE HEALTHCARE INFORMATION SERVICE DOMAIN

The aim was to identify the most appropriate of the four tools to execute sentiment analysis in the domain of healthcare information services. An online survey resulting to 137 responses was used as a starting point. We wanted to evaluate the difference between commercial and non-commercial products available on the market for sentiment analysis. Unfortunately, each tool had different analysis procedures which made the comparison unclear. A common measure for the four tools was created to deal with the inconsistencies that existed. All results were converted to a percentage representing the correct classification predictions to account for the accuracy of the algorithm used by each tool. This was considered an efficient measure throughout the report.

Moreover, we had to account for the difference between the commercial and non-commercial analysis. Due to the supervised learning associated with the non-commercial tools, a baseline (the human labelling) was defined in order to train the classifier as it was mentioned above. This was not done in the case of commercial tools. Taking into consideration that human labels were based on the quantitative results of the survey and assuming that they are relatively accurate, they were used as a baseline for the commercial tools as well, in order to maintain a consistent comparison. Therefore, the accuracy of the commercial tools in the table below was calculated using the quantitative survey responses as a ground truth. A summary for the accuracy of the four tools can be found in Table 6.1.

Table 6.1: Comparison among all four tools for sentiment classification.

	Tool	Accuracy based on correct classification	
		All responses	Single-sentence Responses
Commercial Tools	Semantria	51.09%	53.49%
	TheySay	68.61%	72.09%
Non-Commercial Tools	Google Prediction API	72.25%	54%
	WEKA	82.35%	70%

It proves that commercial tools are possible to compete with some more customisable tools.

Looking at Table 6.1, non-commercial tools are undoubtedly more accurate in sentiment classification as far as healthcare information service domain is concerned. Moreover, it is observed that single-sentence response analysis slightly increased the overall classification accuracy of the commercial tools, while it significantly decreased the accuracy of the non-commercial tools. The latter was expected possibly due to the very small dataset used. Also, it is worth mentioning that the accuracy of Google Prediction API's classifier decreased by approximately 18% when trained using only single sentences. With 82.35% accuracy, WEKA evidently produced the most precise predictions with a weighted average of 0.823 (Table 5.35). It also proved more accurate when analysing single-sentence responses (70% compared with 54% of Google Prediction API). This was expected since WEKA enables a lot of configuration that match the literature approaches defined in this report (Section 3). We were able to apply the techniques used by other researchers and examine their effect on the healthcare information service domain.

A statistical comparison of the four tools using the Cohen's kappa statistic showed a better indication on the reliability of each system (Table 6.2). Due to time limitations, statistical tests were not undertaken for the analysis of single sentences. WEKA was again proved to be the most reliable tool for sentiment classification in the healthcare domain. A kappa statistic of

0.5735 indicated a high correlation between the predictions and the actual labels. In addition the weighted average F-measure (0.809) of WEKA suggested the high performance associated with the tool.

Table 6.2: Statistical comparison of the four tools.

System	Kappa statistic	F-measure Weighted Average
Semantria	0.2692	0.550
TheySay	0.3886	0.678
Google Prediction API	0.2199	0.628
WEKA	0.5735	0.809

It is worth mentioning that the F-measure of Google Prediction API (0.628) is lower than the F-measure of TheySay (0.678). Furthermore, the kappa statistic (0.2199) of Google Prediction API is the lowest among all the tools. This comes as a surprise since Google Prediction API is a non-commercial, customisable tool.

Observing Figures 6.1 and 6.2, the high precision and recall values associated with WEKA indicate the reliability of the tool. The commercial tools have a slight advantage over WEKA for negative precision and positive recall. However, the difference is insignificant.

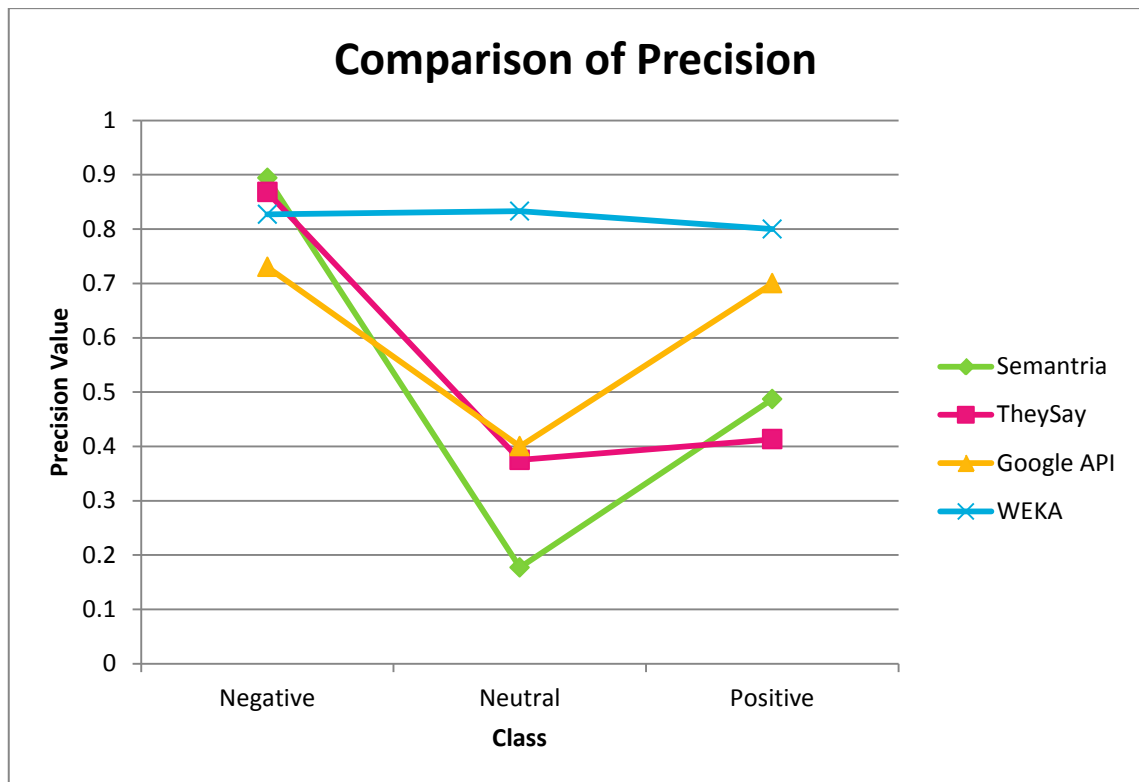


Figure 6.1: Comparison of precision among the four tools.

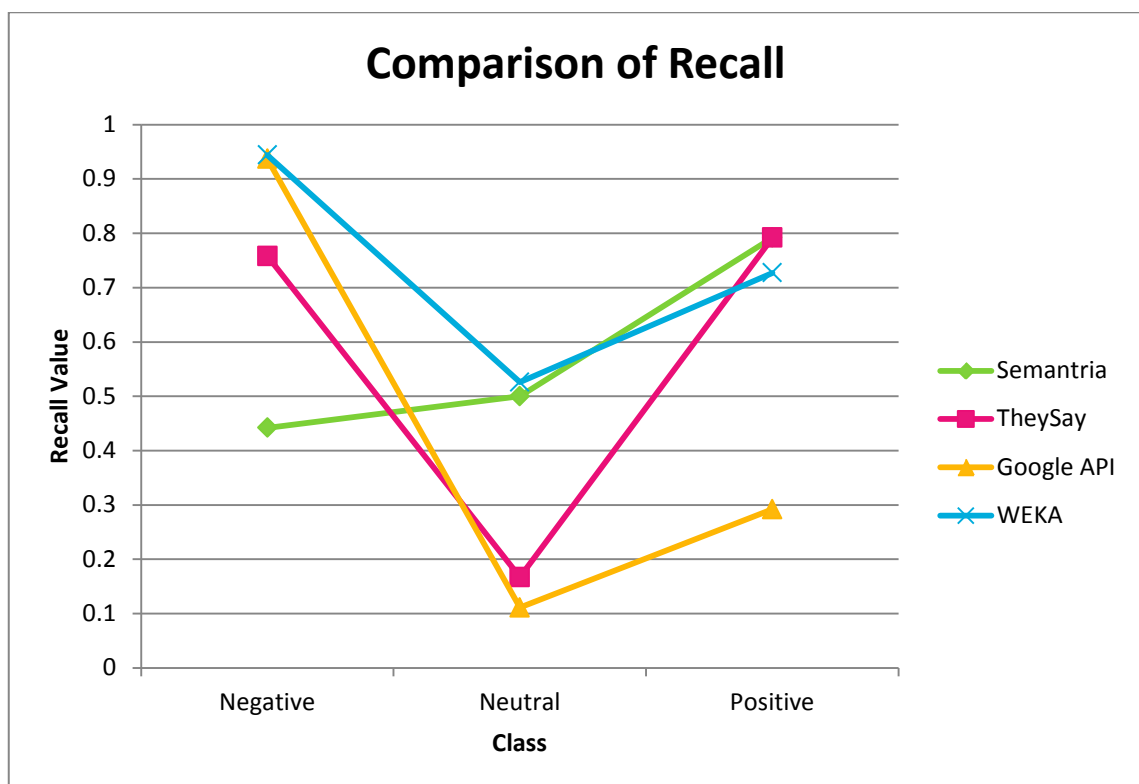


Figure 6.2: Comparison of recall among the four tools.

7 DISCUSSION

The aim of this project was to identify suitable tools that can accurately extract, predict and classify sentiment found in healthcare context. As suggested by Pang & Lee (2008), we combined the human intuitive knowledge with respect to sentiment classification, along with the machines. It was believed that the human touch would provide value to the analysis and it did; as in the human labelling procedure needed for supervised learning in the non-commercial tools (see Appendix C: final labels).

Four systems were chosen to be used in our project. Since a variety of commercial tools exist for sentiment analysis, we wanted to examine their accuracy. They are widely preferred since they are easy to use and provide results very quickly. However, such tools usually do not allow any configuration and a pre-defined algorithm is used for sentiment classification. Therefore, we wanted to compare them with some customisable tools (non-commercial). Most non-commercial tools are more time-consuming since they require pre-processing before training. This, however, helps in building an accurate and reliable model. In our case, we were not able to customise the algorithm used for one of the non-commercial tools.

The examination of just four systems already provided sufficient justification that favours the use of machines in the sentiment analysis area. Excluding Semantria, all systems achieved approximately 70% or above accuracy in the classification results (TheySay was the lowest with 68.61%). Full replication of the human knowledge might never be achieved, but research has shown some promising results. Examples for the healthcare domain can be found in Byrd et. al. (2013), Cameron, Bhagwan & Seth (2012), Mack et. al. (2004), and many other research papers. The high accuracy achieved by most of our tools is an initial investigation in the matter. However, observing the low values of kappa statistic, ranging from 0.2 – 0.4 (Table 6.2), might suggest otherwise. WEKA escaped from this range and achieved 0.5735 kappa statistic. A value

much closer to 1 that indicates a more reliable and stable model when compared with the other systems. This could also be supported by referring to Figure 6.1, where precision follows a stable pattern among the three classes (negative, neutral and positive). Conversely, a high recall is observed for the negative and positive classes, while the neutral class achieved 0.526 recall, indicating once again the difficulty in formulating neutrality.

The main difficulty faced in this project was that each tool had a unique way of performing sentiment analysis by using a variation of features and techniques. Semantria used categories and keywords. TheySay provided detailed analysis that included, among other things, POS analysis, comparative expressions and sarcasm recognition. Google Prediction API used supervised machine learning to learn the pre-labelled data in order to make accurate predictions. We were able to account for negation when preparing the data using one of the methods suggested by Hogenboom et. al. (2011). WEKA performed in a similar way as Google Prediction API, but it also enables configuration of the algorithm used for training. Even though results were converted into a uniform format (percentage of correctly classified responses) for comparison purposes, it was difficult to favour only one group; commercial or non-commercial tools. We were not able to choose which one performs better with a healthcare dataset in order to fully satisfy our objectives. Hence, we have provided recommendations for each tool according to its features by assuming the requirements of the user.

For example, with 82.35% of accuracy, WEKA could be considered the most suitable tool for sentiment analysis (extraction and classification) in a healthcare domain. On the other hand, the extensive analysis provided by WEKA was not found in any of the other three tools, something that compromised the validity of the comparison measures. To compensate for this limitation, statistical tests were conducted to verify the reliability of each tool.

Semantria is particularly recommended for business use. Consumer insights can be discovered to improve the products and services available. Sentiment analysis can be developed in several different ways by Semantria. Therefore, it allows the discovery of patterns and trends in the data. Feedback can be tracked down to the little details through the use of queries, especially negative feedback, where the origin of the problem can be found. This can improve the decision-making process (also observed by USA. IBM Corporation, 2012) and ensure that better customer service will be achieved. However, our results suggest that Semantria cannot be trusted when classifying sentiment in the healthcare domain. Based on the baseline created, only 51.09% of the reviews were correctly classified. This was followed with a weighted average F-measure of 0.550 and a kappa statistic of 0.2692, both indicating the unsuitability of Semantria for the healthcare field. We believe that Semantria's tendency towards neutrality is responsible for this low percentage. Semantria was the only tool that classified 51 responses as neutral (Figure 5.2). This was not expected since we personally examined the data beforehand. Even if we made a judgement mistake in our labelling, 51 neutral responses indicate a large difference in the result. Furthermore, the 40 default categories available might seem useful in a different domain and produce more accurate results.

TheySay had a better score. Almost 70% of the reviews were accurate (based on the baseline). This was expected due to the detailed analysis provided. TheySay takes into consideration a variety of features that improve the performance of the algorithm, like POS analysis; something that is not available in Semantria. It can aid businesses be better prepared for competition (comparative expressions are detected) and deal with sarcastic comments (humour detection) that harm reputation. While comparisons were found to be recognised easily by TheySay, sarcasm and irony detection was not very precise. Nevertheless, it still remains a powerful and accurate tool to use (68.61% accuracy). Of course, there is room for

improvement. The low kappa statistic (0.3886) provides strong evidence to suggest that any agreement between the predictions and the baseline are a result of chance. However, this can be justified by the fact that commercial tools were not trained using the human labels. The kappa statistic produced by Google Prediction API is more worrying (0.2199).

Furthermore, when examining single-sentence responses, TheySay showed a slight increase on the accuracy (72.09%), even though the sample size was small (43 responses). This supports our observation that TheySay seems to have difficulty of correctly identifying long responses, consisting of multiple sentences. TheySay can also contribute to the academic research of sentiment analysis, especially with POS, a widely studied concept (Singh, Mukherjee & Mehta, 2011; Nicholls & Song, 2009; Pang & Lee, 2008, Narayanan, Liu & Choudhary, 2009).

Both commercial tools attempted to classify the entities in the reviews, and results were not very good. Most of the entities were not identified. According to Ding & Liu (2010), knowing the entity responsible for the sentiment polarity is equally important as the polarity itself (in Liu 2012). The quintuple (Equation 1 – see Section 3.1) defined by Liu (2012), also states the importance of knowing the entity to determine the polarity of a statement (quintuple). Further examination of the relation between aspects and the final polarity is required, as it is considered a key part of the classification problem.

A comparison between commercial tools, definitely proved TheySay is superior to Semantria as far as healthcare is concerned. TheySay achieved higher accuracy (68.61% compared to 51.09%) and the second best F-measure (weighted average of 0.678). Semantria was not prepared for a healthcare classification. Consequently, all the features needed to be adjusted for the healthcare domain by the human. Errors were inevitable when it comes to human judgement (as shown by Bruce & Wiebe, 2000). All these negatively influenced Semantria's performance.

On the other hand, non-commercial tools investigated sentiment from a different perspective. Supervised machine learning methods were used that required data labelling before training. As explained in previous sections, in supervised learning the model is trained on pre-labelled data (usually done by a human) and it is tested on unseen data to assess its learning outcome. Different measures exist to evaluate the performance of this process, including cross validation (refer to Section 3.2.1 for more details). The overall process is time consuming and inefficient, allowing room for errors since it is based mostly on human judgement. Using a sufficient number of judges (between 10-20) for labelling could provide more trustworthy results. Furthermore, the tools used required data to be entered in a particular format. This was done manually for Google (extra effort), while in the case of WEKA, it was more straightforward; i.e. data had to be separated into negative, neutral and positive and a single line of code was run via SIMPLE CSI to achieve conversion.

The procedure described above is clearly quite different than the ones associated with the commercial tools. It is more detailed and it was created around healthcare data. Hence, it was expected to perform better than the commercial tools and it did. WEKA achieved 82.35% accuracy (based on cross validation), and Google Prediction API followed with a 72.25%. Comparing these percentages with the commercial output (Semantria ~51%, TheySay ~68%), the difference in accuracy is very clear. Considering the F-measure and kappa statistic, the non-commercial tools lose their advantage. TheySay seemed superior than Google Prediction API in both these statistical tests (F-measure: 0.678 compared to 0.628, Kappa statistic: 0.3886 compared to 0.2199) and forced re-considering the abilities of commercial tools. Considering additional systems for our evaluation would make things more clearer. Moreover, it has to be pointed out that no alterations were done to the algorithm used by Google Prediction API. Despite this, the supervised learning methods used, allowed the inclusion of the tool in the non-commercial group.

Basing our evaluation on the diversity of the tools used, we could suggest that non-commercial tools are recommended for users that are willing to spare the extra time to achieve precision and accuracy.

WEKA reasonably holds the higher position of the four tools. The option of pre-processing the data before training is a unique feature among the tools. It was found that removing punctuation, using lowercase letter and alphabetically tokenizing the data was the ideal combination. It resulted in higher accuracy and better performance of the classifier, proved by the high precision (0.823) and recall (0.824) values. One reason for removing punctuation was to maintain the format of the non-commercial tools as consistent as possible; since punctuation removal was mandatory when using Google Prediction API. Also, we based our decision on the literature of Aggarwal and Zhai (n.d.), who suggested that punctuation, capital letters and stopwords, do not offer extra information to the context. Our results support this. On the other hand, Tsur, Davidov & Rappoport (2010) proved that even though punctuation usage decreased precision, a slight increase in the recall was observed. Furthermore, Tepperman, Traum & Narayanan (2006) showed that punctuation may in fact be beneficial to certain studies. However, their research was based on the sarcastic expression “Yeah right” and hence any observations cannot be generalized, especially in the healthcare domain. Alphabetic tokenization was considered to be beneficial since symbols occurred in the text due to the informality of the survey. Moreover, through our analysis, it was found that considering the frequency of words and the inverse frequency did not improve the model’s performance. In fact, it did just the opposite. The lowest accuracy (67.65%) was found when taking into consideration these features (Appendix H). This contradicts the results of Aggarwal and Zhai (n.d.), who suggested they were useful features when comparing different documents. But, their field of research is very different from this project. In general, it was observed that some

of the findings through the literature were not valid for our project. This might be due to the complexity of the healthcare field or the small dataset used.

On the other hand, Google Prediction API is much easier to use since guidelines are available in the tool's website. However, it does not provide an efficient and productive analysis. The process to correctly analyse the data requires a great amount of effort and time. Cross validation has to be done manually in addition to the pre-processing of the context. During the processing it was suggested that punctuation had to be removed; so essentially Google Prediction API was trained with no punctuation. It would be interesting to investigate the impact punctuation may have had on the training of the model. As already said Tsur, Davidov & Rappoport (2010) found that punctuation usage decreased precision but produced slightly better recall results. With a sufficient amount of neutral and positive responses, the low recall values observed (Figure 6.2) may improve. At the moment, only 16 positive and 10 neutral responses were available for analysis. Furthermore, we were able to account for negation, something that we did not do for the other three tools.

It was interesting to observe the results of single-sentence analysis. Both non-commercial tools experienced a decrease in accuracy and precision, which was expected since fewer data were provided for learning. Google Prediction API experienced 18% decrease in accuracy, while WEKA only 10%. Pre-labelling each sentence as well as the whole response as suggested by McDonald et. al. (2007), might improve the accuracy (in Liu 2012). Further work can be done to investigate this claim from a healthcare perspective.

Through the analysis of the data, we observed that negativity might be associated with long reviews. This is an interesting theory that could be studied in the future. Also, using different classifiers, such as SVM and MaxEnt, could optimize the results produced by WEKA.

The amount of data (137 responses) used is not sufficient to establish the generalisation of the results found. Nevertheless, our findings provide a starting point into the investigation of system's accuracy in the healthcare field.

Based on our findings, we would recommend WEKA for sentiment analysis in the domain of healthcare information services. The features available enable full customisation and help in constructing a reliable model. Even though it was not part of this project, building a customisable algorithm would probably improve the accuracy of the classification, since it would target specific features. Knowledge of programming language is essential for this, which may prove to be an obstacle when time comes. In general, if time permits and the necessary programming knowledge exist, non-commercial tools are preferred since they can be trained to respond to the necessary field; in our case, healthcare. As far as commercial tools are concerned, Semantria proved relatively insufficient for the healthcare domain (for example: no relevant categories existed), while TheySay performed better (F-measure 0.678 and 68.61% accuracy). However, due to the scope of the project we were not able to examine more commercial tools. This is an important limitation of the project and it is intended to extend our research in the future to compensate. Conducting a sentiment analysis with a larger amount of data and additional systems is essential to establish generalisation of our results. Furthermore, different techniques can be examined to conclude to the most suitable one for the healthcare domain. Labelling each sentence individually in addition to the whole document, as well as conducting a more thorough analysis for single sentences polarity can be the starting points for continuing this project.

8 EVALUATION AND CONCLUSIONS

The aim of the project was to investigate whether machines can understand the human language. Commercial (no configuration is allowed) and non-commercial (configuration is allowed) tools were assessed on their abilities of accurately extracting sentiment from a healthcare context. The domain of healthcare information services was reviewed since it uses one of the most complex languages mostly due to the abbreviations associated. The four systems selected were used to classify 137 responses from an online survey (provided by UX Labs) as positive, neutral or negative.

The results produced showed that non-commercial tools, especially WEKA, are preferred for the healthcare field. Due to the timeframe of an MSc project, we were not able to examine more than four tools. This, along with the small dataset provided, does not allow us to generalise our results. Non-commercial tools, particularly WEKA, supported some of the literature examined. It was observed that punctuation usage and capital letters do not impact the accuracy of the sentiment classification. Unlike most researchers, we include a neutral class in our analysis. It was found that neutrality is indeed a difficult concept to examine, which is why most researchers tend to ignore this category. Problems were faced by both human and machines when classifying neutral responses.

The end result of the four tools was a polarity label (positive, neutral or negative) for each response. Therefore, we were able to conduct a comparison based on a baseline created from the quantitative part of the survey. However, the process and algorithm used to reach the final result differ from tool to tool. This allowed us to recommend each tool based on the requirements of the user. Comparing the four systems, it was observed that our expectation that commercial tools would overpower non-commercial tools was not met. However, the low

number of tools examined allows for future work to contradict our findings in this report and examine the reliability of commercial tools even further.

Overall, the project examined various aspects of sentiment analysis within the domain of healthcare information services. It represents a text classification example from a technological perspective. As far as we are aware of, existing systems, which are widely used every day by many organisations, have not been evaluated. An initial investigation was undertaken using four tools. Further work can be done for artificial language replications using more systems.

9 REFLECTIONS

This project report is the result of a four month internship at UX Labs. The goal was to investigate methods for extracting sentiment from surveys using software. During this time, different tools were explored and feature combinations were identified based on their suitability for the healthcare field. In general, all the available features of each tool were examined.

However, looking back, some things could have been done differently in order to achieve better results. It is believed that labelling all the sentences as well as providing a label for the whole responses would have improved the accuracy, as suggested by McDonald et. al. (2007) (in Liu 2012). This was omitted from this project due to time limitations. Even though the labels were mainly based on the quantitative data provided, it is suggested that more than one human was needed to verify the labels assigned, especially for the neutral responses. This would have provided a more reliable baseline and might have helped with the identification of neutrality. While the timeframe did not allow it, it would have been interesting to investigate the performance of different tools not used in this project, especially NLTK; a widely used tool for sentiment analysis ran on the Python platform. However, consistency among the number of tools in each group (commercial and non-commercial) was preferred for comparison purposes. It can be observed that there was a change in the choice of non-commercial tools (Appendix A). The intended software were NLTK which uses the Python platform and GATE. These were replaced with Google Prediction API and WEKA since difficulties were faced in the programming associated with the NLTK tool. Hence, a machine learning approach was chosen. Furthermore, the remaining two ways for examining negation suggested by Hogenboom et. al. (2011), could have been considered.

Overall, knowledge of different methods for analysing data, both qualitatively and quantitatively, was acquired. An accurate labelling of the data was attempted by introducing guidelines to be followed by the human. Time management skills were enhanced since the work plan had a pre-defined deadline. In fact, problems were faced with the work plan suggested at the beginning. However, the complications and obstacles were tackled with success. This resulted in better knowledge for time allocation. Moreover, it is believed that this project report contributed to the sentiment analysis area and the healthcare field. My knowledge in the field was definitely enhanced in the last months. However, text analytics is a very broad field and there are still many things to learn. There is still a lot of work to be done and a lot of existing software to be evaluated.

GLOSSARY

arff – Attribute Relation File Format

Conditionals – conditional sentences

FN – False Negative

FP – False Positive

JJ – adjective

MaxEnt – Maximum Entropy

NB – Naïve Bayes

NLP – Natural Language Processing

NLTK – Natural Language ToolKit

NN – noun

NNS

O – objective with no opinion

OO – objective with opinion

PMI – Pointwise Mutual Information

POS – Part of Speech

RB – adverb

RBR – comparative adverb

RBS – superlative adverb

S – subjective and evaluative

SN – subjective and non-evaluative

SO – Sentiment Orientation

SVM – Support Vector Machine

TN – True Negative

TP – True Positive

VB

VBD

VBG

VCN

WEKA – Waikato Environment for Knowledge Analysis

REFERENCES

- AGGARWAL, C.C. & ZHAI, C. (n.d.) *Mining Text Data*. [Online] Boston, Dordrecht, London. Kluwer Academic Publishers. Available from: <http://charuaggarwal.net/text-content.pdf>. [Accessed: 05 May 2013].
- BENAMARA, F., CHARDON, B., MATHIEU, Y. & POPESCU, V. (2011) Towards Context-Based Subjectivity Analysis. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand, 08-13 November 2011. Asian Federation of Natural Language Processing (AFNLP). pp. 1180-1188.
- BIRD, S., KLEIN, E. & LOPER, E. (2009) *Natural Language Processing with Python*. O'reilly.
- BIRD, S., KLEIN, E. & LOPER, E. (2009). *Supervised classification*. O'reilly
- BOLLEGALA, D., WEIR, D. & CARROLL, J. (2013). Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus. *Knowledge and Data Engineering, IEEE Transactions on*. [Online] 25 (8) pp.1719-1731. Available from: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6203505&contentType=Early+Access+Articles&matchBoolean%3Dtrue%26searchField%3DSearch_All_Text%26queryText%3D%28%28sentiment+analysis%29+AND+method%29 [Accessed: 10 November 2013].
- BREW, A., GREENE, D. & CUNNINGHAM, P. (2010) Using Crowdsourcing and Active Learning to Track Sentiment in Online Media. In *ECAI 2010: 19th European Conference on Artificial Intelligence*. Lisbon, Portugal, 16-20 August 2010. Amsterdam: IOS Press. pp.145-150.

- BRUCE, R.F. & WEIBE, J.M. (2000) Recognizing subjectivity: a case study in manual tagging. *Natural Language Engineering*. [Online] 5 (2) (6) pp.187-205. Available from: <http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=48503> [Accessed: 09 November 2013].
- BYRD, R.J. et. al. (2013) International Journal of Medical Informatics. *Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records*. [Online] pp.1-10. Available from: <http://www.sciencedirect.com/science/article/pii/S1386505612002468>. [Accessed: 06 May 2013].
- CAMERON, D.; BHAGWAN, V. & SHETH, A.P. (2012) Towards comprehensive longitudinal healthcare data capture. In *Bioinformatics and Biomedicine Workshops (BIBMW) 2012*. Philadelphia, PA, 04-07 October 2012. Philadelphia, PA: IEEE International Conference. pp. 240-247.
- DEY, L. & HAQUE, S. K. M. (2008) Opinion mining from noisy text data. In *Proceedings of the second workshop on Analytics for noisy unstructured text data*. Singapore, 24 July 2008. N.Y.: ACM. pp.83-90.
- GOOGLE DEVELOPERS. (2012) *Google Prediction API*. [Online]. Google.
- GRIMMER, J. & STEWART, B.M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*. [Online] pp.1-31. Available from: <http://www.stanford.edu/~jgrimmer/tad2.pdf> [Accessed: 06 November 2013].

- HALL, M. et. al. (2009) *The WEKA Data Mining Software: An Update*. 11 (1). SIGKDD Explorations.
- HOGENBOOM, A., VAN ITERSON, P., HEERSCHOP, B., FRASINCAR, F. & KAYMAK, U. (2011) Determining negation scope and strength in sentiment analysis. In *Systems, Man, and Cybernetics (SMC), 2011*. Anchorage, AK, 09-12 October 2011. IEEE International Conference. pp.2589-2594.
- JINDAL, N. & LIU, B. (2006a) Identifying comparative sentences in text documents. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Seattle, Washington, USA, 6 August 2006. New York, USA: ACM. pp.244-251.
- JINDAL, N. & LIU, B. (2006b) Mining Comparative Sentences and Relations. In *Proceedings of AAAI-06, the 21st National Conference on Artificial Intelligence*. Boston, US. US: AAAI Press. pp.1331-1336.
- JINDAL, N. & LIU, B. (2008) Opinion Spam and Analysis. In *Proceedings of the International Conference on Web Search and Web Data Mining*. Palo Alto, California, USA, 11-12 February 2008. N.Y., USA: ACM. pp.219-230.
- KEIM, D. A., OELKE, D. (2007) Literature Fingerprinting: A New Method for Visual Literary Analysis. In *Visual Science and Technology, 2007. VAST 2007. IEEE Symposium on*. Sacramento, CA, 30 October – 01 November 2007. IEEE. pp.115-122.
- KEKE, C. et. al. (2008) Leveraging Sentiment Analysis for Topic Detection. In *Web Intelligence and Intelligent Agent Technology*. Volume 1 Sydney, NSW, 09-12 December 2008. Sydney, NSW: IEEE/WIC/ACM International Conference. pp. 265-271.

- KIM, J., LI, J. J. & LEE, J. H. (2009) Discovering the Discriminative Views: Measuring Term Weights for Sentiment Analysis. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapore, 02-07 August 2009. USA: Association for Computational Linguistics and AFNLP. pp.253-261.
- LI, N. & DASH WU, D. (2010) Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*. [Online] 48 (2). pp. 354 – 368. Available from: <http://www.sciencedirect.com/science/article/pii/S0167923609002097>. [Accessed: 13 November 2013].
- LIU, B. (2012) *Sentiment Analysis and Opinion Mining*. [Online] 22nd April. Available from: <http://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf>. [Accessed: 13 May 2013].
- MACK, R. et. al. (2004) IBM Systems Journal. *Text analytics for life science using the Unstructured Information Management Architecture*. [Online] 43 (3). pp.490-515. Available from: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5386750>. [Accessed: 03 May 2013].
- MAKS, I. & VOSSEN, P. (2012). A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems*. [Online] 53 (4) pp.680-688. Available from: <http://www.sciencedirect.com/science/article/pii/S0167923612001364?np=y> [Accessed: 07 November 2013].

- MCDONALD, S. (1999). Exploring the process of inference generation in sarcasm: A review of normal and clinical studies. *Brain and Language*. [Online] 68 (3) pp.486-506. Available from: <http://www.sciencedirect.com/science/article/pii/S0093934X99921247> [Accessed: 14 November 2013].
- MITCHELL, T. M. (1997) *Machine Learning*. McGraw-Hill International Editions. N.Y.: McGraw-Hill, Inc.
- NARAYANAN, R., LIU, B. & CHOUDHARY, A. (2009) Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Volume 1 Singapore. Stroudsburg, PA, USA: Association for Computational Linguistics. pp.180-189.
- NICHOLLS, C. & SONG, F. (2009) Improving Sentiment Analysis with Part-of-Speech Weighting. In *Proceedings of 8th International Conference on Machine Learning and Cybernetics*. Volume 3 Baoding, 12-15 July 2009. IEEE. pp.1592-1597.
- PANG, B, LEE, L. & VAITHYANATHAN, S. (2002) Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Conference of Empirical Methods in Natural Language Processing*. Philadelphia, July 2002. US: Association of Computational Linguistics. pp.79-86.
- PANG, B. & LEE, L. (2008) Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*. [Online] 2. p.1-135. Available from: <http://www.cs.cornell.edu/home/llee/omsa/omsa-published.pdf>. [Accessed: 13 May 2013].
- PANG, B., LEE, L. & VAITHYANATHAN, S. (2002). *Average three-fold cross-validation accuracies, in percent. Boldface: best performance for a given setting (row). Recall that our baseline results ranged from 50% to 69%*. US: Association of Computational Linguistics.

PULMAN, S. & MOILANEN, K. (2013) *TheySay*. [Online]. TheySay Limited.

QUESTIONPRO. (2011) *Text Analytics: Visualizing and analysing open-ended text data*. [Online]

Available from: <http://www.questionpro.com/images/bookshelf/SurveyAnalytics-TextAnalytics.pdf>. [Accessed: 03 May 2013].

ROGERS, Y., SHARP, H. & PREECE, J. (2011) *Interaction Design: Beyond human-computer interaction*. 3rd Edition. UK: John Wiley & Sons Ltd.

SEMANTRIA, LLC. (2014) *Semantria*. [Excel add-in]. Semantria Inc.

SINGH, V. K., MUKHERJEE, M. & MEHTA, G. K. (2011) Sentiment and Mood Analysis of Weblogs Using POS Tagging Based Approach. *Communications in Computer and Information Science*. [Online] 168 pp.313-324. Available from: http://link.springer.com/chapter/10.1007%2F978-3-642-22606-9_33 [Accessed: 16 November 2013].

SPANGLER, S. et. al. (2010) SIMPLE: Interactive Analytics on Patent Data. In *Data Mining Workshops (ICDMW)*. Sydney, NSW, 13 December 2010. Sydney, NSW: IEEE International Conference. pp. 426-433.

TEPPERMAN, J., TRAUM, D.R. & NARAYANAN, S. (2006) "Yeah Right": Sarcasm recognition for spoken dialogue systems. *INTERSPEECH 2006*. Pittsburgh, Pennsylvania, 17-21 September 2006. pp.1838-1841.

TSUR, O., DAVIDOV, D. & RAPPOPORT, A. (2010) ICWSM – A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence (AAAI). pp.162-169.

- USA. IBM CORPORATION. (2012) *IBM SPSS Modeler Text Analytics 15 User's Guide*, USA: IBM Corporation. [Online] Available from:
ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/en/Users_Guide_For_Text_Analytics.pdf. [Accessed: 04 May 2013].
- USA. IBM CORPORATION. (2012) *IBM SPSS Text Analytics for Surveys 4.0.1 User's Guide*, USA: IBM Corporation. [Online] Available from:
ftp://public.dhe.ibm.com/software/analytics/spss/documentation/tafs/4.0.1/en/Users_Guide.pdf. [Accessed: 04 May 2013].
- USA. SAS INSTITUTE INC. (2010) *Text Analytics for Social Media: Evolving Tools for an Evolving Environment*, SAS Institute Inc. [Online] Available from:
http://www.sas.com/resources/whitepaper/wp_24091.pdf. [Accessed: 04 May 2013].
- VassarStats: Website for Statistical Computation. (2013) *Kappa as a Measure of Concordance in Categorical Sorting*. [Online] Available from: <http://vassarstats.net/kappa.html> [Accessed: 20 December 2013].

APPENDICES

A Project Proposal

Name: Despo Georgiou

E-mail address: despo.georgiou.1@city.ac.uk

Contact Phone number: +44 (0)7599295661

Project Title: Text Analytics

Supervisor: Dr Andrew MacFarlane

Introduction

Analysing text can be more useful than one might think. It is well known that knowledge comes with information and more knowledge is always an advantage (Mack et. al., 2004). While the importance of discrete data was never underestimated, it must be pointed out that the combination of structured and unstructured data can provide new insights in finding patterns and trends (USA. IBM Corporation. 2012).

Sentiment analysis concerns with the investigation of opinions and thoughts. It is only reasonable that such information can be vital to the decision-making process (Pang & Lee, 2008). It can offer enormous opportunities to knowledge which will in turn influence productivity, efficiency and so on (QuestionPro. 2011; USA. IBM Corporation. 2012; Cameron, Bhagwan & Sheth, 2012). A number of methods for analysing sentiment exist, that involve both artificial intelligence and the human element, and they are widely used in a variety of fields. Some examples and their applications can be seen below:

- **Surveys:**

Explore and build models (USA. IBM Corporation. 2012) and analyse surveys that contain open-ended questions (QuestionPro. 2011).

- **Consumer feedback:**

Analyse reviews with the aim to understand and satisfy customers. Use these reviews to evaluate a product (Pang & Lee, 2008). Moreover, effort to understand consumer opinions about products and services is made, that will enable the creation of new innovation opportunities and competitive advantages (Keke et. al., 2008).

- **Social Media:**

Detect sensitive content inappropriate for ads (Pang & Lee, 2008).

- **Business & Government:**

Provide clear, immediate and actionable insights into current performance and the ability to predict future activities. Also, it can provide consistent and accurate information that can be trusted to make more complete decisions in order to improve business performance (USA. IBM Corporation. 2012). Identifying key strategies or market shifts using a trend analysis can also be useful, as well as fraud detection capabilities. Governments can monitor the sources for increases in hostile or negative communication (Pang & Lee, 2008).

- **Politics:**

Understand voters' way of thinking and in return provide a more clear view to voters about a politician's position as well as enhance the quality of information they have access to (Pang & Lee, 2008).

- **Health (Mack et. al., 2004):**

Support problem-solving in life science by managing and analysing biomedical text. The aim is to treat diseases and enhance the health of humans. Analysing treatment effects as reported in various forums is one way to go. It has also been found that drug-discovery can become faster.

Project objectives

In this project the importance of text analytics is discussed and the different methods used for analysing text, focusing on sentiment analysis, are outlined. The aim is to investigate the extent to which sentiment can be reliably extracted from qualitative survey data, focusing on verbatim responses from the domain of healthcare information services. Ways used for validating these methods, using quantitative measures taken from the same data, are also investigated. A

comparison between commercial and non-commercial products for sentiment analysis will reveal the most appropriate method to evaluate the data.

Academic Context

Each individual is different in a unique way. Exposure to different environmental influences, which help in developing unique characteristics, defines individuality. Subsequently, the way each human mind perceives information changes. This involves opinions, sentiments, evaluations, attitudes and emotions that can reasonably affect the decision-making process (Liu, 2012). Individuality can only be expressed in an unstructured way and with the help of text mining, hidden knowledge can be uncovered.

One of the most common text mining tasks is text clustering. In a few words, clustering uses a single function to group together similar objects in the data. It has numerous applications. Key methods based on document organization and classification, were identified by Aggarwal and Zhai. These include feature selection methods such as normalising the text by using the inverse document frequency, to reduce the importance of the most frequent words in the document. This ensures that the matching of documents is more influenced by that of more discriminative words. Stop words, such as “a”, “the” or “of”, which are common in every document, are also removed to provide greater accuracy. Likewise, infrequent words, which are usually a result of misspelling or typographical errors, *“do not add any extra value to the similarity”* (Aggarwal & Zhai) and are also removed. All these can be *“leveraged in order to improve the quality of results”* (Aggarwal & Zhai).

While Aggarwal and Zhai struggled with choosing the appropriate function, Keim and Oelke (2007) introduced a new solution to the problem – the creation of characteristic fingerprints by calculating features for different hierarchy levels (words, sentences, paragraphs, etc.). They believed that using *“a single feature vector to characterise the whole text, disregards important information, and*

interesting patterns and traits/trends are lost". In their paper, it is claimed that the combination of a fingerprint and visualization can offer a deeper understanding, and they were able to analyse the average word length to identify the document genre.

The above methods give a clear idea of the importance of text analytics and what they have to offer. However, the challenge in this field is to analyse people's beliefs, thoughts and opinions. Societies are based around people, and individuality is a key aspect to be studied. The most common way to retrieve public opinions is questionnaires. It is up to researchers to decide, but it has been found that a combination of closed-ended and open-ended questions give a more accurate idea of the interviewee's replies (Rogers, Sharp & Preece, 2011). Closed-ended questions are usually used as an introduction to the field questioned, and later open-ended questions are introduced to give the research a 'chunk' of useful unstructured data to analyse (QuestionPro. 2011).

There are various techniques to extract information from such unstructured data:

- **Eyeball browsing.** The most obvious – reading the comments. Having the analytical (closed questions) *in the back of your head* can aid to have a more complete picture of how to interpret the overall survey.
- **Grouping.** Create groups for closed-ended questions and use them to filter open-ended questions. This is efficient and ensures more accurate results.
- **Keywords.** Specify a particular topic and use appropriate keywords from the wording respondents were using.
- **Hierarchical categories.** *Subcategories can be used to group items such as different concepts or topic areas more accurately.* (USA. IBM Corporation. 2012)

Eyeball browsing is a purely human technique. Even though it is impractical, humans are more keen in understanding another human compared to an artificial intelligence system. On the other hand, a system is cost and time efficient, and can increase the reliability and accuracy of the results, since the extraction and categorisation is performed in a consistent and repeatable manner (USA. IBM Corporation. 2012).

Sentiment analysis can also aid business and governments to understand their public and be better prepared for the future. Nowadays, where social media are ruling the world, advanced linguistic technologies and Natural Language Processing (NLP) can *“rapidly process a large variety of unstructured data and extract and organise the key concepts”*. They are a speed and cost effective means of finding the meaning in text with a higher degree of accuracy (USA. IBM Corporation. 2012).

Systems aid in trend analysis and predictive modelling by plotting promotional activities against historic map of on-going customer responses, for example (USA. SAS Institute Inc. 2010). This is done by converting source data to a standard format and then identifying candidate terms necessary for the creation of classes. Experimentation with different techniques is often inevitable to see which procedure best suits each company (USA. IBM Corporation. 2012). If used wisely, technology can provide a competitive advantage (Spangler et. al., 2010) to all business in *“attracting, retaining and growing customers, while reducing fraud and mitigating risks”* by trying to understand the customer’s behaviour (USA. IBM Corporation. 2012). However, IBM Corporation (2012) sets an important question: *“Can machines understand human communication?”*

“Humans often have trouble understanding each other, even when speaking face to face”. The answer lies in the combination of the two; humans and machines working closely together. Computers can calculate whether a given phrase is positive or negative with a certain degree of

confidence and humans can review the low-confidence results and advice the machine how to grade them. Pang et al. (2008) provided evidence that could support the collaboration between machines and humans. Their pilot study examined the difficulty of solving the sentiment-polarity classification problem by having two humans and a corpus' statistic to classify keywords that represent negative and positive sentiment polarity. The results are shown below.

	Proposed word lists	Accuracy (%)	Ties (%)
Human 1	positive: <i>dazzling, brilliant, phenomenal, excellent, fantastic</i> negative: <i>suck, terrible, awful, unwatchable, hideous</i>	58	75
Human 2	positive: <i>gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting</i> negative: <i>bad, cliched, sucks, boring, stupid, slow</i>	64	39
Statistics-based	positive: <i>love, wonderful, best, great, superb, still, beautiful</i> negative: <i>bad, worst, stupid, waste, boring, ?, !</i>	69	16

“Indeed, applying machine learning techniques based on unigram models can achieve over 80% in accuracy, which is better than the performance based on hand-picked keywords reported above” (Pang & Lee, 2008). With time, computers will absorb more and more and results will be more accurate and useful.

Due to the complexity of biomedical texts, all the abbreviations and the frequent misspellings, health information services is one of the most difficult and challenging fields for text analytics to be applied. However, it is also one of the most rewarding fields to work in. Text analytics can provide *“new insights that could impact diagnosis, treatment and overall patient care”* (Cameron, Bhagwan & Seth, 2012). There are systems that combine a patient’s symptoms, medication and history record, along with their family history, and a confident diagnosis (USA. IBM Corporation. 2011).

To be more specific, Byrd et. al. (2013) developed a system that *“accurately identifies and labels affirmations and denials of Framingham diagnostic criteria (a symptom of heart failure) in primary care clinical notes and may help in the attempt to improve the early detection of a heart failure”*.

Sentiment Analysis and Opinion Mining

According to Liu (2012), an opinion is defined by a quintuple,

$$(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$$

where e_i is the name of an entity, a_{ij} is an aspect of e_i , s_{ijkl} is the sentiment on aspect a_{ij} of entity e_i , h_k is the opinion holder, and t_l is the time when the opinion is expressed by h_k . The sentiment s_{ijkl} is positive, negative, or neutral, or expressed with different strength levels.

The quintuple is used to transform unstructured text to structured data. It is important to note that all five components are essential to avoid any errors, since the definition above represents a hierarchy of parts that form an opinion. It has been found that simplification can result in information loss using the quintuple representation. However, the definition is considered sufficient for most applications. *“It provides a good source of information and also a framework for generating both qualitative and quantitative summaries”*.

A sentiment classification problem can be solved by either a supervised or unsupervised learning. Supervised machine learning applications use features such as frequency of terms, the part of speech of words, negation words, etc. to help in the classification procedure. It has been found that adjectives and adverbs are more likely to express sentiment. This, however, does not mean that nouns (e.g., *rubbish and junk*) and verbs (e.g., *hate and love*) cannot be used to express sentiment. Turney’s unsupervised technique *“performed classification based on some*

fixed syntactic patterns that are likely to be used to express opinions". His algorithm extracted two consecutive words that their Part Of Speech (POS) was any of the patterns identified in the table below and estimated their sentiment orientation (SO) using the *pointwise mutual information* (PMI) measure:

$$PMI(term_1, term_2) = \log_2 \left(\frac{\Pr(term_1 \wedge term_2)}{\Pr(term_1) \Pr(term_2)} \right)$$

PMI measures the degree of statistical dependence between two terms.

	First word	Second word	Third word (not extracted)
1	JJ	NN or NNS	anything
2	RB, RBR, or RBS	JJ	not NN nor NNS
3	JJ	JJ	not NN nor NNS
4	NN or NNS	JJ	not NN nor NNS
5	RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything

Afterwards, the algorithm calculates the average *SO* of all extracted information and classifies the document as positive or negative, depending on the polarity of the average *SO*.

Methodology

Data

An online survey was conducted to evaluate a healthcare information service website that has recently been updated. A questionnaire of eight questions resulted in an overall of 210 responses. The questions were divided to six closed-ended questions, with a scale of 1-5, 1 being the highest, and two open-ended questions to give the respondents the freedom to express their personal opinion on the matter.

Procedure

A quantitative analysis of the data was made first, to identify possible patterns and trends. However, these are not absolute; bias should be avoided. Afterwards, eyeballing browsing was executed to better understand the material – mostly qualitative data were involved here. This helped in forming an initial awareness of the concept hidden between the words.

Using relevant articles, journals, research papers etc., that further enhanced knowledge and skills as far as text analytics are concerned, possible categories and groups were identified with the aid of keywords.

All of the above formed a general idea of the data from a human perspective. Two commercial products (free version/trial/demo) were chosen to be compared with two non-commercial products (programming may be required) in order to identify the most efficient, accurate and consistent sentiment extraction technique.

Commercial products:

- Semantria – add-in Microsoft Excel tool
- TheySay

Non-commercial products:

- GATE – General Architecture for Text Engineering
- NLTK – uses Python language

The two non-commercial products are closely connected.

A comparison between the methods will eventually result in identifying the most suitable method for analysing health information service related text. The comparison will include the following measures:

- Creating groups to identify different concepts and ideas.
- Using systems to analyse data and predict possible trends and patterns.
- Keywords will be identified at the beginning to help with the analysis. These keywords will be defined according to the text provided to the student.

Other measures will be added if necessary - more appropriate methods may be discovered on the way.

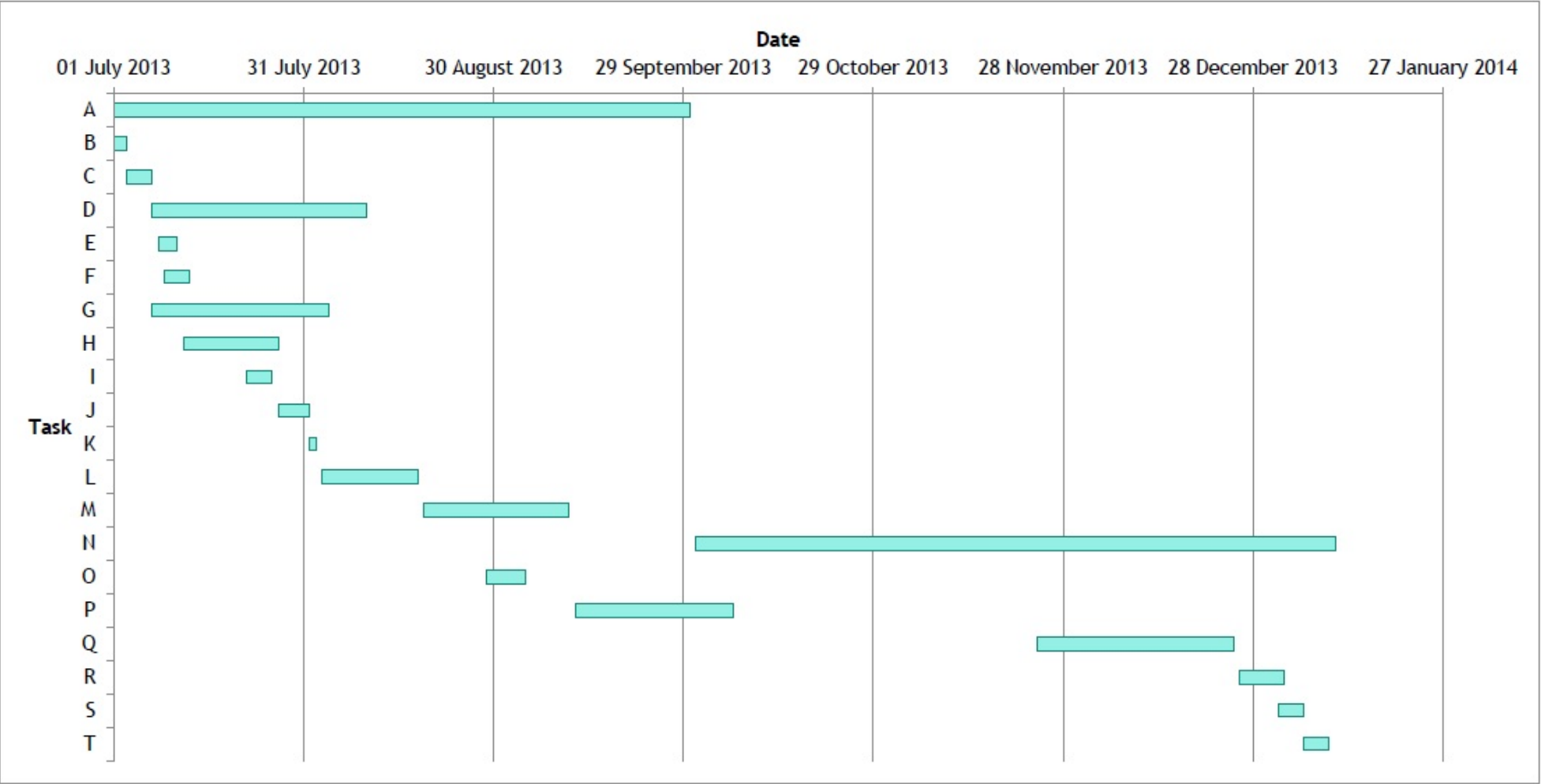
In order to compare and evaluate the performance of each product effectively, a ground truth must be established using the quantitative data from the responses.

Visualisation will be used to identify possible similarities between respondents and hence create a separate segment or group for them.

Work Plan

	TASK	DURATION (in days)	START DATE	END DATE
*	<i>Internship meeting: introduction</i>	1	21 June 2013	21 June 2013
A	Internship	91	01 July 2013	30 Sept 2013
B	First examination of data	2	01 July 2013	02 July 2013
C	Search for products available	4	03 July 2013	06 July 2013
D	Internship-based Project Proposal	34	07 July 2013	02 Aug 2013
E	Eyeball browsing	3	08 July 2013	10 July 2013
F	Identify possible keywords, groups and categories	4	09 July 2013	12 July 2013
*	<i>Internship meeting</i>	1	12 July 2013	12 July 2013
G	More research into sentiment analysis and opinion mining	28	07 July 2013	03 Aug 2013
H	Experiment with different products for sentiment analysis	15	12 July 2013	26 July 2013
I	Choose appropriate products	4	22 July 2013	25 July 2013
J	Internship-based Project Proposal: Write-Up	5	27 July 2013	31 July 2013
*	<i>Internship meeting</i>	1	30 July 2013	30 July 2013
K	Internship-based Project Proposal: Proof-read report	1	01 Aug 2013	01 Aug 2013
◇	<i>Milestone 1: Internship-based Project Proposal Submission</i>			02 Aug 2013
L	Commercial products experimentation	15	03 Aug 2013	17 Aug 2013
*	<i>Internship meeting</i>	1	20 Aug 2013	20 Aug 2013
M	Non-commercial products experimentation	23	19 Aug 2013	10 Sept 2013
N	Project Report	101	01 Oct 2013	10 Jan 2014
O	Project Report: Write-up draft of methodologies and products used	6	29 Aug 2013	03 Sept 2013
*	<i>Internship meeting</i>	1	15 Sept 2013	15 Sept 2013
P	Analyse results	25	12 Sept 2013	30 Sept 2013
*	<i>Supervisor meeting</i>	1	28 Oct 2013	28 Oct 2013
Q	Project Report: Write-up	31	24 Nov 2013	24 Dec 2013
*	<i>Supervisor meeting</i>	1	03 Dec 2013	03 Dec 2013
R	Project Report: Proof-read	7	26 Dec 2013	31 Dec 2013
S	Project Report: Proof-read by external parties	4	01 Jan 2014	04 Jan 2014
T	Project Report: Final touches	4	05 Jan 2014	08 Jan 2014
◇	<i>Milestone 2: Project Report Submission</i>			10 Jan 2014

Gantt Chart



Risks

Below there are possible risks that may occur during the project, their impact and likelihood as well as potential recovery from them should they occur. This aims to assess the project feasibility.

PERSONAL RISKS – CONTROLLABLE RISKS

Risk	Explanation	Impact	Likelihood	Recovery
Lack of Knowledge and Skills	The student must have substantial knowledge on the text analytics field to be able to cope with the project requirements.	Severe consequences: the project will not be feasible since the researcher will not be familiar with the knowledge required to fulfil the project aim.	45%	<u>Mitigation:</u> ✓ If area is unknown to the student in question, they must prepare themselves for the purpose of the project by doing appropriate research on the matter.
Time Limitation	Risk of missing the project deadline for any reason.	Pressure will be added to finish on time which may result to a rushing work being done. The project's quality will be at risk.	15%	<u>Mitigation:</u> ✓ Ensure that the work plan is well-planned beforehand and stick to it.
Out of Scope	Moving away from the scope of the project.	The project will not be complete since it will not satisfy what it was meant to satisfy.	20%	<u>Mitigation:</u> ✓ Have regular meetings with the supervisor to ensure that the project is not getting out of scope.
Plagiarism	Academics research involves referencing articles, journals, etc. Subsequently, plagiarism becomes a risk.	Months of work will be wasted and the consequences are severe, one of which is failing the MSc course.	10%	<u>Mitigation:</u> ✓ Include references while writing the report and do not leave them at the end. ✓ Use citation, italic

PROJECT RISKS – UNCONTROLLABLE RISKS

RISK	EXPLANATION	IMPACT	LIKELIHOOD	RECOVERY
Low Quality of Data	There might be a lot of data to work with but only a small part of that data might be useful for the aim of the project.	Data gathered are not sufficient to draw valid conclusions for the project.	35%	✓ Extend research if possible to gather more data relevant to the
Low Performance	Neither of the chosen methods gives good results. This is may be due to: <ul style="list-style-type: none"> • Inappropriate methods was chosen • Quality of data could influence the performance 	The project aim will not be completed covered. However, no severe consequences are associated since this is part of the research.	40%	✓ Extend research if possible to examine other methods and their suitability for sentiment analysis for health information services.
Unanswered Questions	Required comments of the survey might be left clear.	Individual's opinions and thoughts will be ignored due to lack of availability. This will impact the quality of data received and important points may be missing that could offer insights.	30%	<i>Mitigation:</i> ✓ Ensure that the aim of the survey is clear to the respondents. This will give them an incentive to help with their comments.

Ethics and Confidentiality

The methodology of the project was carefully planned; taking under consideration any legal or ethical issues that might arise. Surveys are the main part of this research. Responses will act as the data and will be analysed using the methods described previously in this report.

Like in every research, the consent of the respondents is necessary to use the information they provide. The briefing of the survey must be concise and precise in order to specify to the respondents their involvement to the project.

Respondents will be informed that the aim of the survey is to gather sentiment data regarding health information services. Subsequently, their opinions and thoughts on the subject will be the main source of this project and that is why their participation is important. Anyone involved in this research will be clearly informed about the project procedure. It will be made clear that participants' personal data, i.e. name, address, telephone, etc., will not be mentioned anywhere in the report. Instead, they will be treated anonymously and confidentially. However, their age and gender may be used for statistical analysis. Furthermore, it will be clarified that any data gathered during the survey will only be kept until analysed.

A Research Ethics Checklist can be found in Appendix 1.

References

AGGARWAL, C.C. & ZHAI, C. *Mining Text Data*. [Online] Boston, Dordrecht, London. Kluwer Academic Publishers. Available from: <http://charuaggarwal.net/text-content.pdf>. [Accessed: 05 May 2013].

BYRD, R.J. et. al. (2013) International Journal of Medical Informatics. *Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records*. [Online] p.1-10. Available from: http://0-www.sciencedirect.com.wam.city.ac.uk/science?_ob=MiamilImageURL&_cid=271161&_user=910131&_pii=S1386505612002468&_check=y&_origin=article&_zone=toolbar&_coverDate=2013-Jan-11&view=c&originContentFamily=serial&wchp=dGLbVlt-zSkzS&md5=3b35f2a43e7305cdd068d43315c5446c&pid=1-s2.0-S1386505612002468-main.pdf. [Accessed: 06 May 2013].

CAMERON, D.; BHAGWAN, V. & SHETH, A.P. (2012) Towards comprehensive longitudinal healthcare data capture. In *Bioinformatics and Biomedicine Workshops*. Philadelphia, PA, 4th October to 7th October. Philadelphia, PA: IEEE International Conference. pp. 240-247.

KEIM, D.A. & OELKE, D. (2007) Literature Fingerprinting: A New Method for Visual Literary Analysis. In *Visual Analytics Science and Technology*. Sacramento, CA, 30th October to 1st November. Sacramento, CA: IEEE International Conference. pp. 115-122.

KEKE, C. et. al. (2008) Leveraging Sentiment Analysis for Topic Detection. In *Web Intelligence and Intelligent Agent Technology*. Volume 1 Sydney, NSW, 9th December to 12th December. Sydney, NSW: IEEE International Conference. pp. 265-271.

LIU, B. (2012) *Sentiment Analysis and Opinion Mining*. [Online] 22nd April. Available from:

<http://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf>. [Accessed: 13 May 2013].

MACK, R. et. al. (2004) IBM Systems Journal. *Text analytics for life science using the Unstructured Information Management Architecture*. [Online] 43 (3). p.490-515. Available from: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5386750>. [Accessed: 03 May 2013].

PANG, B. & LEE, L. (2008) Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*. [Online] 2. p.1-135. Available from: <http://www.cs.cornell.edu/home/llee/omsa/omsa-published.pdf>. [Accessed: 13 May 2013].

QUESTIONPRO. (2011) *Text Analytics: Visualizing and analysing open-ended text data*. [Online] Available from: <http://www.questionpro.com/images/bookshelf/SurveyAnalytics-TextAnalytics.pdf>. [Accessed: 03 May 2013].

ROGERS, Y., SHARP, H. & PREECE, J. (2011) *Interaction Design: Beyond human-computer interaction*. 3rd Edition. UK: John Wiley & Sons Ltd.

SPANGLER, S. et. al. (2010) SIMPLE: Interactive Analytics on Patent Data. *In Data Mining Workshops*. Sydney, NSW, 13rd December. Sydney, NSW: IEEE International Conference. pp. 426-433.

USA. IBM CORPORATION. (2011) *IBM Watson and Medical Record Text Analytics: HIMSS Presentation*, USA: IBM Corporation.

USA. IBM CORPORATION. (2012) *IBM SPSS Modeler Text Analytics 15 User's Guide*, USA: IBM Corporation. [Online] Available from: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/en/Users_Guide_For_Text_Analytics.pdf. [Accessed: 04 May 2013].

USA. IBM CORPORATION. (2012) *IBM SPSS Text Analytics for Surveys 4.0.1 User's Guide*, USA: IBM Corporation. [Online] Available from: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/tafs/4.0.1/en/Users_Guide.pdf. [Accessed: 04 May 2013].

USA. SAS INSTITUTE INC. (2010) *Text Analytics for Social Media: Evolving Tools for an Evolving Environment*, SAS Institute Inc. [Online] Available from: http://www.sas.com/resources/whitepaper/wp_24091.pdf. [Accessed: 04 May 2013].

Appendix

1. Research Ethics Checklist

NUMBER	QUESTION	YES/NO
--------	----------	--------

If the answer to the following questions is NO, then the project plan needs to be modified, because the project should not continue as currently planned. Seek advice very early about it.

1.	Does the planned project pose only minimal and predictable risk to the student?	Yes
2.	Does it pose only minimal and predictable risk to other people affected by the project?	Yes
3.	Are arrangements for the supervision of the project appropriate?	Yes
4.	Is the project carried out or supervised by competent researchers?	Yes
5.	Do the foreseeable benefits of the project outweigh the foreseeable risks?	Yes

If the answer to any of the following questions is YES, then authorisation from the Senate's Ethics Committee is required. Seek advice very early about it.

6.	Does the project involve interaction with, or collection personal information about, people who are vulnerable because of their social, psychological or medical circumstances?	No
7.	Does the project involve animals?	No
8.	Does the project involve research on pregnant women or women in labour?	No
9.	Does the project involve research on persons under the age of 18?	No
10.	Does the project involve research on human tissue?	No
11.	Does the project involve research on vulnerable categories of people who may include minority groups?	No

The following questions must be answered YES, i.e. the student must commit to satisfy these conditions and have a plan to ensure they are satisfied.

Will the student ensure that any people subject to observation or data collection are:

12.	fully informed about the procedures affecting them and affecting any information collected about them (how the data will be used, to whom they will be disclosed, how long they will be kept)?	Yes
13.	fully informed about the purpose of the research?	Yes
14.	Will the consent of these people be obtained?	Yes
15.	When these people can be classified as research subjects, will it be clear to them that they may withdraw at any time?	Yes
16.	Will the student make arrangements that material or private information obtained from or about these people remains confidential?	Yes

B Semantria: Conversion Results

We wanted to create a range with no negative values to maintain a consistent format throughout the tools. The new range was created by adding 2 to each value range. The new ranges were then converted into percentages by dividing the total 100% into equal groups.

Table B.I: Conversion ranges.

Original Range	-2	-1	0	1	2	Conversion Method/Procedure
New Range	0	1	2	3	4	add 2 to each group
Percentage Range	0%	25%	50%	75%	100%	divide the total 100% in equal group ranges

The percentage range was used to convert the original Semantria polarity values into percentages for comparison purposes.

Table B.II: Converted polarity values.

Response Number	Original Polarity Value	New Polarity Value (percentage)
Response 1	0.26392987370491	56.60%
Response 2	0.62749999761581	65.69%
Response 3	-0.19492141902447	45.13%
Response 4	0.00000000000000	50.00%
Response 5	-0.09999999403954	47.50%
Response 6	0.01646505296230	50.41%

C Data Labelling

The average score of each response was calculated based on the values of Questions 1, 2, 3, 6 and 8. Using the guideline showed in Table C.I, an initial label was assigned to each response.

These/The labels were then checked by the human to ensure they represent the correct classification. If errors were found, the label was changed to the appropriate polarity class. The final classification labels can be found in Table C.II. The survey responses are omitted from this table to ensure client confidentiality. Due to the complexity associated with the neutral classification, guidelines were created to ensure as consistent labelling as possible. These can be seen below.

Table C.I: Scores assigned to each class.

Class	Score Range
Positive	1 – 2.7
Neutral	2.8 – 4.2
Negative	4.3 - 5

Neutral Classification Guidelines:

- Any response that contained equal amount of positive and negative feedback was considered neutral. However, this was also dependent on the choice of words and how strong the negative or positive feedback was for the response.
- Any response that stated facts. For example: the response “I am doing it for a friend with Anklosis Anklosing”, does not express any kind of positive or negative opinion or thought. It simply states that the respondent is doing it for a friend with the particular disease. Similar responses were categorised as neutral.
- Any response that was considered irrelevant for the website.

Table C.II: The human labels.

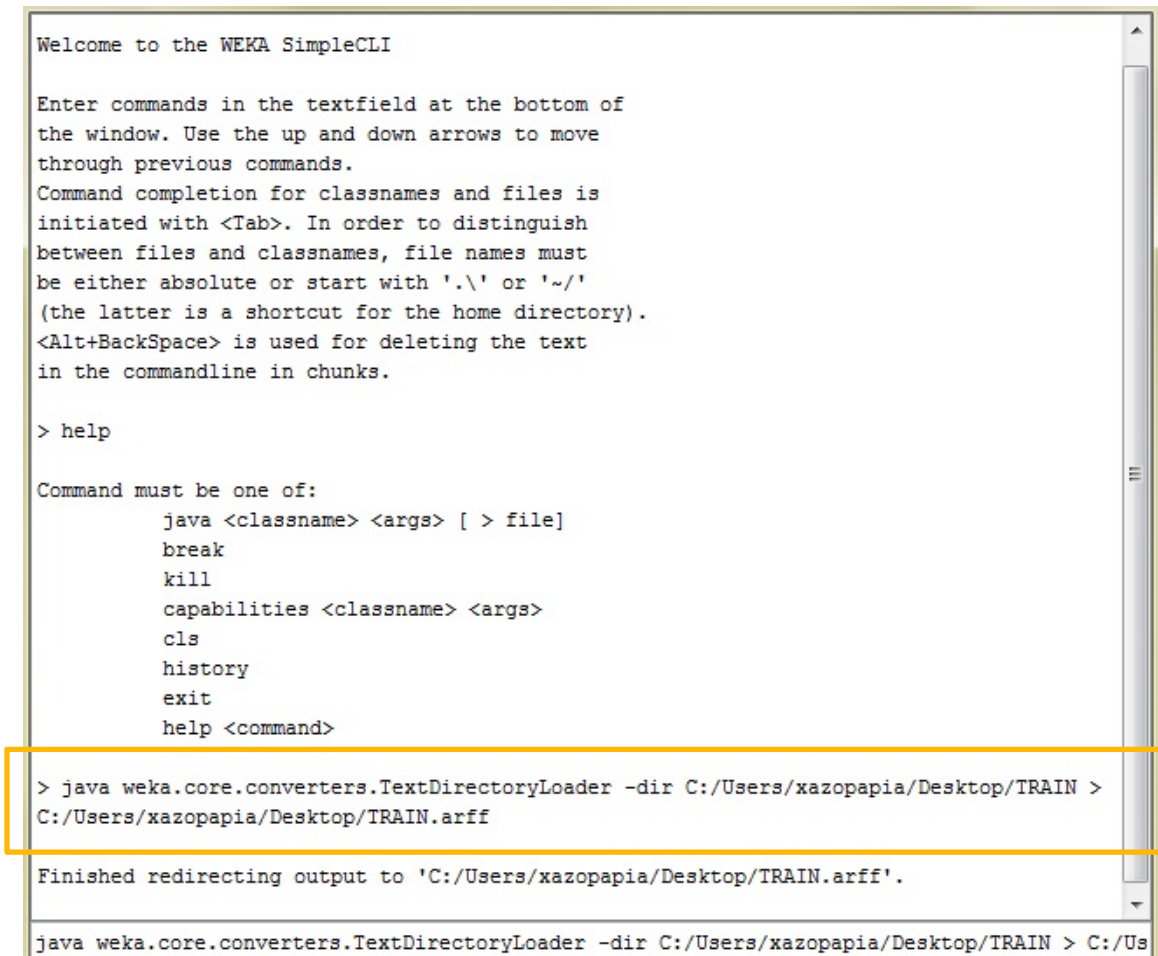
Responses	Q.1	Q.2	Q.3	Q.6	Q.8	Average Score	Initial Label	Final Label
Response 1	2	2	2	3	3	2.4	Positive	Negative
Response 2	5	5	1	2	1	2.8	Positive	Negative
Response 3	3	4	2	4	3	3.2	Neutral	Negative
Response 4	3	3	4	3	3	3.2	Neutral	Negative
Response 5	5	5	2	3	1	3.2	Neutral	Negative
Response 6	3	4	4	3	3	3.4	Neutral	Negative
Response 7	4	5	3	2	3	3.4	Neutral	Negative
Response 8	3	5	3	5	1	3.4	Neutral	Negative
Response 9	4	5	2	5	1	3.4	Neutral	Negative
Response 10	5	5	2	4	1	3.4	Neutral	Negative
Response 11	4	4	3	3	3	3.4	Neutral	Negative
Response 12	4	4	2	4		3.5	Neutral	Negative
Response 13	4	5	2	4	3	3.6	Neutral	Negative
Response 14	5	5	3	3	2	3.6	Neutral	Negative
Response 15	4	4	4	4	2	3.6	Neutral	Negative
Response 16	4	4	3	4	3	3.6	Neutral	Negative
Response 17	2	5	4	4	3	3.6	Neutral	Negative
Response 18	3	5	2	5		3.75	Neutral	Negative
Response 19	1	5	3	5	5	3.8	Neutral	Negative
Response 20	4	4	3	4	4	3.8	Neutral	Negative
Response 21	5	5	1	5	3	3.8	Neutral	Negative
Response 22	3	5	3	4	4	3.8	Neutral	Negative
Response 23	5	5	1	5	4	4	Neutral	Negative
Response 24	5	5	3	4	3	4	Neutral	Negative
Response 25	4	5	4	4	3	4	Neutral	Negative
Response 26	3	5	3	4	5	4	Neutral	Negative
Response 27	4	5	2	5	5	4.2	Neutral	Negative
Response 28	5	5	5	3	3	4.2	Neutral	Negative
Response 29	5	5	1	5	5	4.2	Neutral	Negative
Response 30	4	5	4	4	4	4.2	Neutral	Negative
Response 31	4	5	3	5	4	4.2	Neutral	Negative
Response 32	5	5	3		4	4.25	Neutral	Negative
Response 33	5	5	2	5	5	4.4	Negative	Negative
Response 34	5	5	3	5	4	4.4	Negative	Negative
Response 35	5	5	4	4	4	4.4	Negative	Negative
Response 36	5	5	3	5	4	4.4	Negative	Negative

Responses	Q.1	Q.2	Q.3	Q.6	Q.8	Average Score	Initial Label	Final Label
Response 37	5	5	4	4	4	4.4	Negative	Negative
Response 38	5	5	3	5	4	4.4	Negative	Negative
Response 39	5	5	2	5	5	4.4	Negative	Negative
Response 40	5	5	3	5		4.5	Negative	Negative
Response 41	4	5	5	4	5	4.6	Negative	Negative
Response 42	5	5	3	5	5	4.6	Negative	Negative
Response 43	5	5	3	5	5	4.6	Negative	Negative
Response 44	5	5	3	5	5	4.6	Negative	Negative
Response 45	5	5	3	5	5	4.6	Negative	Negative
Response 46	5	5	3	5	5	4.6	Negative	Negative
Response 47	5	5	3	5	5	4.6	Negative	Negative
Response 48	5	5	3	5	5	4.6	Negative	Negative
Response 49	5	5	3	5	5	4.6	Negative	Negative
Response 50	5	5	3	5	5	4.6	Negative	Negative
Response 51	5	5	3	5	5	4.6	Negative	Negative
Response 52	5	5	3	5	5	4.6	Negative	Negative
Response 53	5	5	5	4	4	4.6	Negative	Negative
Response 54	5	5	3	5	5	4.6	Negative	Negative
Response 55	5	5	5	5	3	4.6	Negative	Negative
Response 56	5	5	3	5	5	4.6	Negative	Negative
Response 57	5	5	5	5	4	4.8	Negative	Negative
Response 58	5	5	4	5	5	4.8	Negative	Negative
Response 59	5	5	5	5	4	4.8	Negative	Negative
Response 60	5	5	4	5	5	4.8	Negative	Negative
Response 61	5	5	4	5	5	4.8	Negative	Negative
Response 62	5	5	5	5	5	5	Negative	Negative
Response 63	5	5	5	5		5	Negative	Negative
Response 64	5	5	5	5		5	Negative	Negative
Response 65	5	5	5	5	5	5	Negative	Negative
Response 66	5	5	5	5	5	5	Negative	Negative
Response 67	5	5	5	5	5	5	Negative	Negative
Response 68	5	5	5	5	5	5	Negative	Negative
Response 69	5	5	5	5	5	5	Negative	Negative
Response 70	5	5	5	5	5	5	Negative	Negative
Response 71	5	5	5	5	5	5	Negative	Negative
Response 72	5	5	5	5	5	5	Negative	Negative
Response 73	5	5	5	5	5	5	Negative	Negative
Response 74	5	5	5	5		5	Negative	Negative

Responses	Q.1	Q.2	Q.3	Q.6	Q.8	Average Score	Initial Label	Final Label
Response 75	5	5	5	5	5	5	Negative	Negative
Response 76	5	5	5	5		5	Negative	Negative
Response 77	5	5	5	5	5	5	Negative	Negative
Response 78	5	5	5	5	5	5	Negative	Negative
Response 79	5	5	5	5	5	5	Negative	Negative
Response 80	5	5	5	5	5	5	Negative	Negative
Response 81								Negative
Response 82								Negative
Response 83								Negative
Response 84								Negative
Response 85	4	5	3	5	5	4.4	Negative	Negative
Response 86	5	5	5	5	5	5	Negative	Negative
Response 87	3	4	2	4	3	3.2	Neutral	Negative
Response 88	1	5		5	3	3.5	Neutral	Negative
Response 89	1	5	2	5	5	3.6	Neutral	Negative
Response 90	5	5	3	3	3	3.8	Neutral	Negative
Response 91	4	4				4	Neutral	Negative
Response 92	4	5	5	5	4	4	Neutral	Negative
Response 93	5	5	4	4	5	4.2	Neutral	Negative
Response 94	5	5	5	5	5	5	Negative	Negative
Response 95	5	5	4	4	4	4	Neutral	Negative
Response 96	3	3	3	3	3	3	Neutral	Neutral
Response 97								Neutral
Response 98	4	1			3	3.25	Neutral	Neutral
Response 99	5	5				5	Negative	Neutral
Response 100	1	1	1	1	1	1	Positive	Neutral
Response 101	4	4	4	4	4	3.6	Neutral	Neutral
Response 102	4	5	4	4	3	3.8	Neutral	Neutral
Response 103	2	3	2	2	2	2.2	Positive	Neutral
Response 104	4	3	3	3	1	2.8	Neutral	Neutral
Response 105	3	3	3	3	4	3.2	Neutral	Neutral
Response 106	4	4	4	4	3	3.6	Neutral	Neutral
Response 107	4	4	4	4	2	3.6	Neutral	Neutral
Response 108	5	5	4	4	3	3.8	Neutral	Neutral
Response 109	3	5	4	3	4	3.8	Neutral	Neutral
Response 110	5	5	5	1	5	4.2	Neutral	Neutral
Response 111	3	5	4	5	5	4.4	Negative	Neutral
Response 112	4	5	3	5	5	4.4	Negative	Neutral

Responses	Q.1	Q.2	Q.3	Q.6	Q.8	Average Score	Initial Label	Final Label
Response 113	2	2	2	2	2	2	Positive	Neutral
Response 114	1	1	1	1	1	1	Positive	Positive
Response 115	1	1	1	1	1	1	Positive	Positive
Response 116	1	1		1	1	1	Positive	Positive
Response 117	1	1	1	1	1	1	Positive	Positive
Response 118	1	1	1	1	1	1	Positive	Positive
Response 119	2	1	2	1	1	1.2	Positive	Positive
Response 120	1	1	2	1	1	1.2	Positive	Positive
Response 121	1	1	3	1	1	1.4	Positive	Positive
Response 122	1	2	2	1	1	1.4	Positive	Positive
Response 123	2	1	1	1	2	1.4	Positive	Positive
Response 124	1	2	2	2	1	1.6	Positive	Positive
Response 125	2	3	1	2	1	1.8	Positive	Positive
Response 126	1	2	2	2	2	1.8	Positive	Positive
Response 127	2	2	1	2	2	1.8	Positive	Positive
Response 128	2	2	2	2	1	1.8	Positive	Positive
Response 129	2	2	3		1	2	Positive	Positive
Response 130	2	2	2	2	2	2	Positive	Positive
Response 131	2	2	2	2	2	2	Positive	Positive
Response 132	1	2	2	3		2	Positive	Positive
Response 133	2	3	2	2	1	2	Positive	Positive
Response 134	2	2	2	3	1	2	Positive	Positive
Response 135	2	2	3	2	2	2.2	Positive	Positive
Response 136	4	3	4	3	4	3.6	Neutral	Positive
Response 137	4	3	4	4	4	3.8	Neutral	Positive

D WEKA Simple CLI



```
Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.
Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with './' or '~/ '
(the latter is a shortcut for the home directory).
<Alt+BackSpace> is used for deleting the text
in the commandline in chunks.

> help

Command must be one of:
    java <classname> <args> [ > file]
    break
    kill
    capabilities <classname> <args>
    cls
    history
    exit
    help <command>

> java weka.core.converters.TextDirectoryLoader -dir C:/Users/xazopapia/Desktop/TRAIN >
C:/Users/xazopapia/Desktop/TRAIN.arff

Finished redirecting output to 'C:/Users/xazopapia/Desktop/TRAIN.arff'.

java weka.core.converters.TextDirectoryLoader -dir C:/Users/xazopapia/Desktop/TRAIN > C:/Us
```

Figure D.I: WEKA Simple CLI.

The single line of code was used to convert the data to an arff file in order to match the format suitable for WEKA. The format of an arff file is shown in Figure D.II.

```
@relation C__Users_xazopapia_Desktop_TRAIN

@attribute text string
@attribute @@class@@ {negative,neutral,positive}

@data

'i»¿There are so many spelling mistakes and typos on this page
'i»¿Please could the XXX advice for schools summary be updated
'i»¿I am delighted that the website has had an upgrade and it i
'i»¿Very frustrating as I cannot find the guidance that I am lc
'i»¿It\'s a more modern look .',neutral
```

Figure D.II: Arff file created.

E Semantria: Collection Analysis

Facets	Count	Negative	Positive	Neutral	Attributes	Attributes count
guidance	19	1	0	18	full	2
website	15	1	0	14	new	6
website	15	1	0	14	old	2
site	9	1	0	8	old	2
time	8	0	0	8	guideline	2
time	8	0	0	8	several	2
version	8	0	0	8	full	2
version	8	0	0	8	quick	2
guidelines	7	0	0	7	clinic	2
link	7	1	1	5	drug	2
page	7	0	0	7	same	2
one	6	0	0	6		
search	5	0	0	5	result	2
information	5	0	1	4		
use	5	1	0	4		
screen	4	0	0	4	colorectal	2
access	4	0	0	4		
people	4	0	0	4		

Figure E.1: The collection analysis results. Semantria identified the facets of the text, i.e. the most frequent words.

F Commercial Tools: Precision, Recall and F-measure

Using the Equations 4 – 6 defined in Section 3.2, the precision, recall and F-measure were calculated.

F.1 Semantria

i. Positive Class

Table F.I: Table of confusion for positive responses.

	Positives	Negatives
True	19	93
False	5	20

$$\text{True Positive (TP)Rate} = \frac{TP}{TP + FN} = \frac{19}{19 + 20} = 0.48718$$

$$\text{False Positive (FP)Rate} = \frac{FP}{FP + TN} = \frac{5}{5 + 93} = 0.05102$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{19}{19 + 5} = 0.79167$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{19}{19 + 20} = 0.48718$$

$$F - \text{measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.79167 \times 0.48718}{0.79167 + 0.48718} = 0.60318$$

ii. **Neutral Class**

Table F.II: Table of confusion for neutral responses.

	Positives	Negatives
True	9	78
False	9	41

$$\text{True Positive (TP)Rate} = \frac{TP}{TP + FN} = \frac{9}{9 + 41} = 0.18$$

$$\text{False Positive (FP)Rate} = \frac{FP}{FP + TN} = \frac{9}{9 + 78} = 0.10345$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{9}{9 + 9} = 0.5$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{9}{9 + 41} = 0.18$$

$$F - \text{measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.5 \times 0.18}{0.5 + 0.18} = 0.26471$$

iii. **Negative Class**

Table F.III: Table of confusion for negative responses.

	Positives	Negatives
True	43	37
False	52	5

$$\text{True Positive (TP)Rate} = \frac{TP}{TP + FN} = \frac{43}{43 + 5} = 0.89583$$

$$\text{False Positive (FP)Rate} = \frac{FP}{FP + TN} = \frac{52}{52 + 37} = 0.58427$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{43}{43 + 52} = 0.45263$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{43}{43 + 5} = 0.89583$$

$$F - \text{measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.45263 \times 0.89583}{0.45263 + 0.89583} = 0.60140$$

F.2 TheySay

i. Positive Class

Table F.IV: Table of confusion for positive responses.

	Positives	Negatives
True	19	86
False	5	27

$$\text{True Positive (TP)Rate} = \frac{TP}{TP + FN} = \frac{19}{19 + 27} = 0.41304$$

$$\text{False Positive (FP)Rate} = \frac{FP}{FP + TN} = \frac{5}{5 + 86} = 0.05495$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{19}{19 + 5} = 0.79167$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{19}{19 + 27} = 0.41304$$

$$F - \text{measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.79167 \times 0.41304}{0.79167 + 0.41304} = 0.49145$$

ii. **Neutral Class**

Table F.V: Table of confusion for neutral responses.

	Positives	Negatives
True	3	114
False	15	5

$$\text{True Positive (TP)Rate} = \frac{TP}{TP + FN} = \frac{3}{3 + 5} = 0.375$$

$$\text{False Positive (FP)Rate} = \frac{FP}{FP + TN} = \frac{15}{15 + 114} = 0.11628$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{3}{3 + 15} = 0.16667$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{3}{3 + 5} = 0.375$$

$$F - \text{measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.16667 \times 0.375}{0.16667 + 0.375} = 0.23077$$

iii. **Negative Class**

Table F.VI: Table of confusion for negative responses.

	Positives	Negatives
True	72	31
False	23	11

$$\text{True Positive (TP)Rate} = \frac{TP}{TP + FN} = \frac{72}{72 + 11} = 0.86747$$

$$\text{False Positive (FP)Rate} = \frac{FP}{FP + TN} = \frac{23}{23 + 31} = 0.42593$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{72}{72 + 23} = 0.75790$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{72}{72 + 11} = 0.86747$$

$$F - \text{measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.75790 \times 0.86747}{0.75790 + 0.86747} = 0.80899$$

G Google Prediction API: Precision, Recall and F-measure

Using the Equations 4 – 6 defined in Section 3.2, the precision, recall and F-measure were calculated.

i. Positive Class

Table G.I: Table of confusion for positive responses.

	Positives	Negatives
True	7	110
False	3	17

$$\text{True Positive (TP)Rate} = \frac{TP}{TP + FN} = \frac{7}{7 + 17} = 0.29167$$

$$\text{False Positive (FP)Rate} = \frac{FP}{FP + TN} = \frac{3}{3 + 110} = 0.02655$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{7}{7 + 3} = 0.7$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{7}{7 + 17} = 0.29167$$

$$F - \text{measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.7 \times 0.29167}{0.7 + 0.29167} = 0.20588$$

ii. **Neutral Class**

Table G.II: Table of confusion for neutral responses.

	Positives	Negatives
True	2	116
False	3	16

$$\text{True Positive (TP)Rate} = \frac{TP}{TP + FN} = \frac{2}{2 + 16} = 0.11111$$

$$\text{False Positive (FP)Rate} = \frac{FP}{FP + TN} = \frac{3}{3 + 116} = 0.02521$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{2}{2 + 3} = 0.4$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{2}{2 + 16} = 0.11111$$

$$F - \text{measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.4 \times 0.11111}{0.4 + 0.11111} = 0.17391$$

iii. **Negative Class**

Table G.III: Table of confusion for negative responses.

	Positives	Negatives
True	89	9
False	33	6

$$\text{True Positive (TP)Rate} = \frac{TP}{TP + FN} = \frac{89}{89 + 6} = 0.93684$$

$$\text{False Positive (FP)Rate} = \frac{FP}{FP + TN} = \frac{33}{33 + 9} = 0.78571$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{89}{89 + 33} = 0.72951$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{89}{89 + 6} = 0.93684$$

$$F - \text{measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.72951 \times 0.93684}{0.72951 + 0.93684} = 0.82028$$

H Summary of WEKA results

Table H.I: Experimentation of WEKA's features.

Features	Description and Comments	Option 1	Option 2	Option 3	Option 4	Option 5	Option 6	Option 7
IDF Transform	<p>Document Frequency. The lower, the better. DF value is flipped in the formula ($= TF \times IDF$)</p> <p><i>Low IDF: words that rarely appear in the document collection.</i></p>	False	False	False	True	True	False	False
TF Transform	<p>Term Frequency. This feature finds the words and documents that are strongly related.</p> <p>The higher, the better. 'OutputWordCount' needs to be turned ON since TF needs to know how often a word appears in the document and not if it is present.</p> <p><i>High TF: words that frequently appear in particular documents.</i></p>	False	False	False	True	True	False	False

Features	Description and Comments	Option 1	Option 2	Option 3	Option 4	Option 5	Option 6	Option 7
Attribute Indices	Select the attributes that should be considered when WEKA analyses the data.	First-last	First-last	First-last	First-last	First-last	First-last	First-last
Attribute Name Prefix	If more dataset are present, this feature provides an easy way to distinguish them.	-	-	-	-	-	-	-
Do not operate on per class basis	It is partly related with 'WordsToKeep' . Essentially, this feature defines if the classes present should strictly obey the limitation numbering of words defined in the 'WordsToKeep' feature.	False	False	False	False	False	False	False
Invert Selection		False	False	False	False	False	False	False
Lower case Tokens	Removal of capitals and transforming everything to lowercase letters.	True	True	True	True	True	True	True
Min Term Frequency	The number of times a word has to appear in the document in order to be considered as attribute.	1	1	1	1	1	1	1
Normalize Document Length	Data measured in different scale and re-measuring them in a common scale. According to the document length, the words are given a different value.	Normalize all data	Null	Null	Null	Null	Null	Null

Features	Description and Comments	Option 1	Option 2	Option 3	Option 4	Option 5	Option 6	Option 7
Output Word Counts	Whether WEKA should count the frequency of words or not. Setting this to FALSE implies that WEKA counts the word presences and not the frequency of the word.	False	False	False	True	True	True	False
Periodic Pruning	This changes when count pruning, i.e. words to keep, is done. Here prune simply removes low frequency words.	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
Stemmer	Stemming words. Tries to use words better by breaking them down to a smaller form called stem.	Null	Null	Lovins	Lovins	Null	Null	Null
Stopwords	Available stopwords lists in many languages.	English	English	English	English	English	English	English
Tokenizer	Different ways of splitting up the text into tokens.	Alphabetic	Alphabetic	Alphabetic	Alphabetic	Alphabetic	Alphabetic	Alphabetic
Use stoplists	Enable stoplist usage.	False	False	True	True	True	True	True
Words to keep	Usually used in a large dataset to minimise the material processed. <i>Default: 1000 words.</i>	1000	1000	1000	1000	1000	1000	1000
Correctly Classified Instances	Naïve Bayes (4-fold cross validation)	71.5686%	77.451%	69.6078%	67.6471%	71.5686%	70.5882%	73.5294%