Extracting Sentiment from Healthcare Survey Data: An Evaluation of Sentiment Analysis Tools

Despo Georgiou, Andrew MacFarlane
Department of Computer Science
City University London
London, U.K.

Tony Russell-Rose
UX Labs
Guildford, U.K.

***Abstract*—Sentiment analysis is an emerging discipline with many analytical tools available. This project aimed to examine a number of tools regarding their suitability for healthcare data. A comparison between commercial and non-commercial tools was made using responses from an online survey which evaluated design changes made to a clinical information service. The commercial tools were Semantria and TheySay and the non-commercial tools were WEKA and Google Prediction API. Different approaches were followed for each tool to determine the polarity of each response (i.e. positive, negative or neutral). Overall, the non-commercial tools outperformed their commercial counterparts. However, due to the different features offered by the tools, specific recommendations are made for each. In addition, single-sentence responses were tested in isolation to determine the extent to which they more clearly express a single polarity. Further work can be done to establish the relationship between single-sentence responses and the sentiment they express.**

Keywords—sentiment analysis; machine learning; tools; classification; healthcare

I. INTRODUCTION

Much can be gained by extracting sentiment from healthcare information, including views from beneficiaries such as patients. However, the inherent complexity of biomedical tests (e.g. due to the use of abbreviations and frequent misspellings) make the health information services one of the most difficult field for text analytics to be applied.

Consequently, there is a large volume of research aimed at providing improvements in this field, partly due to the benefits associated – better heath care for humans. Byrd et. al. [1], for example, focused on predicting heart failure by developing a NLP procedure to identify the signs and symptoms associated with the condition. The aim was to assist in the decision support for early detection of heart failure. They faced problems while transferring data from Electronic Heath Record (EHR) due to spelling errors - a common problem that creates confusion in the healthcare domain. Since part of our project is to ensure healthcare-related survey responses are analysed correctly, it is believed sentiment analysis could assist the patient-doctor communication and enhance patient's experience. It could also help with the analysis of EHR data.

In addition, systems like BioTeKS exist that *"support problem-solving in life science by analysing biomedical text"* [2]. In general, most researchers focus on creating a system or algorithm that is usually fully customisable depending on the goal of the project. Our focus, however, was to evaluate available sentiment analysis tools with healthcare survey data.

Our investigation was influenced by the study of [3]. Considering the popularity of the Social Web and its ability to *"capture the "pulse" of the society in real time"* [3], they focused on Technosocial Predictive Analytics (TPA) by examining different technologies and tools, such as 'Who Is Sick', a location-aware service and 'Google Flu Trends', which estimates flu activity. They concluded that TPA is a powerful method in the healthcare field and that *"as the technology develops it will provide more sophisticated and reliable tools for public health use"* [3]. Athough, our project is not focused on the Social Web per se, we wanted to explore whether the technology has developed as predicted, focusing on sentiment analysis and healthcare.

II. SENTIMENT ANALYSIS

Sentiment Analysis (SA) is concerned with the investigation of opinions, thoughts and feelings. It is used as a tool to understand Natural Language Processing (NLP). It aims to determine the thoughts of the speaker or writer regarding a specific subject or topic [4] or simply to 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. Hence, it can be found useful in these areas:

- Surveys: analyse open-ended questions [5].
- Business and Governments: ensure consistent and accurate information provision that can assist in the decision-making process [6, 7], as well as monitor the sources for increases in hostile or negative communication [8].
- Consumer feedback: analyse reviews to increase the consumer experience and satisfaction.
- Health: manage and analyse biomedical text that can help in treating diseases and speed up the drug-discovery process, in order to enhance the health of humans [2].

III. AIMS AND OBJECTIVES

The main aim of this project was to assess the difficulty of analysing sentiment in the healthcare domain. The complexity associated with this field was mentioned above, so the goal was

to investigate whether existing SA tools can respond to the needs of any healthcare-related matter. Technology has profoundly changed the way the healthcare system works by introducing, amongst others, EHR, advisory websites and computerised systems in the workplace. In this work, we were mostly concerned with data from patient feedback and evaluation of healthcare services. We believe that an accurate analysis of human language could contribute positively to both the patient's and doctor's experience.

The secondary aim was to find an appropriate tool on the market that could accurately extract the SA involved in the survey responses used in this project. There is a wide variety of tools on the market, but given the constraints we concentrated on two commercial and two non-commercial tools – how easy is it to configure the algorithms in the tools in order to assist users to make more informed decisions about their healthcare problems.

IV. Related Work

A basic concept related to SA is subjectivity – since by definition subjective texts are expected to explicitly express feelings and beliefs that construct an opinion [9]. Hence, there are various studies that focus on understanding and recognising subjective sentences [9, 10]. However, due to the clear link between subjectivity and opinions, researchers tend to ignore objectivity in the belief that there is no significant loss of information. However, Benamara et. al. [9] provide evidence to justify that sentiment can exist in both types of sentences by examining different combinations of subjective and objective sentences.

While subjectivity and objectivity are important in analysing sentiment there are other, more complicated and difficult concepts to consider when dealing with raw data, especially with survey data. Sarcasm is one and it is commonly found in reviews, feedbacks and survey data. People use sarcasm to say exactly the opposite of what they really mean. They are not lying, as they give their honest opinion for a specific product or service, but they do it in a sarcastic way. Identifying sarcasm in the text is difficult, especially when using artificial intelligence systems. Humans are generally better at recognising sarcasm, but due to its unclear nature, even humans may be confused as to whether a comment is sarcastic or not [11]. This is further explained in [12] which demonstrated that a lot of effort is required to understand and interpret sarcasm.

Tsur et. al. [11] focused on an approach to identify sarcasm. By processing their finalised dataset through pattern-based and punctuation-based features (high frequency words and punctuations were used), they built a semi-supervised model. Results showed an overall good precision and recall. Punctuation usage decreased precision but had a slightly better recall. However, Tepperman et. al. [12] showed that the use of punctuation may be actually beneficial in recognising sarcasm. Their study, however, focused on examining the different uses and forms of the expression "Yeah right"; a very commonly used phrase that can be found in sarcastic sentences.

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 utilized in SA, where many researchers use a list of keywords to process their data. Such keywords include: *'no', 'not', '-n't', 'never', 'without',* etc. [13]. Hogenboom et. al. [13] considered three different methods to determine negation and concluded that negating the sentiment of the first sentiment word following or around the negation keyword is the best approach as it significantly improves accuracy.

Furthermore, while studying conditional sentences by separating them into two clauses, Narayanan et. al. [14] found that using a single classifier to determine the polarity of the whole sentence has slightly better results than classifying the two clauses separately.

Taking this into a broader perspective, it was pointed out that the usual assumption of researchers that a sentence only contains a single opinion is rarely true [15]. Specifically, 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. as cited by [15] examined both document-level and sentence-level classification in an attempt to correct this by assigning labels to both sentences and the whole document. Training the classifier using both levels produced more accurate results for both levels of classification [15].

An approach to deal with sentences containing both positive and negative sentiment is to evaluate the dominant polarity of the sentence. This works; however, when faced with equal negative and positive opinions in the text, a different approach is needed. Despite the difficulty associated with the neutral polarity class, taking into consideration neutrality is a good solution to this problem. However, not all neutral statements indicate an equal amount of positive and negative opinions; they can also imply that no opinion is held. In general, many of the research papers ignore neutrality, making the problem easier [15]. In this report, neutrality expressed by people was considered as part of the analysis.

Apart from all the theories and concepts that help with recognising and categorising sentiment into different polarities (negative, neutral, positive), processing the data prior training is an important feature of SA. Since supervised learning methods are commonly used, a traditional approach of 'cleaning' the data is sometimes necessary to enhance the classifier's performance. This involves eliminating punctuation, examining frequency of words or even taking into consideration negation.

As the name indicates, supervised machine learning methods use pre-labelled data to train the model using a specific classifier and then predict the label of some unseen data (testing data) of the same domain as the training data. A more detailed explanation of this can be found in [16]. This

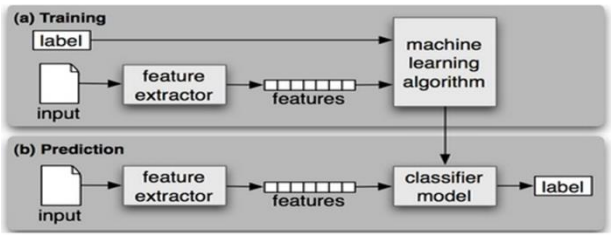

Fig. 1. Supervised classification [3]

approach is a way to assess the accuracy and precision of the classifier in order to *"establish the reliability of supervised machine learning methods"* [17]. Fig. 1 provides a visual interpretation of how supervised learning works in the domain of SA. Dobbin & Simon [16] explored different ratios between training and testing sets and found that the optimal splitting is 2/3 training and 1/3 testing. However, they also discovered that the proportion of the training set can range between 40% and 80% and still be optimal based on a wide range of conditions studies.

Choosing the appropriate classifier is equally important. In this work, Naïve Bayes was used to classify text as negative, neutral or positive since it *"tends to work well on text classifications"* [18] and trains quicker than Support Vector Machines (SVM). On the other hand, SVM can *"eliminate the need of feature selection"* [19]. Other classifiers include Maximum Entropy (MaxEnt), which has the ability to *"converge to a global optimum with respect to a training set"* [20].

Review data are often preferred for SA 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 [21]. Excessive use of exclamations or symbols that intend to show the emotions of the reviewer (emoticons) are examples of the informality. These can be found very useful when analysing sentiment [21]. Nevertheless, in order to build an accurate model, pre-processing is necessary. The aim is to minimise the noise in the text as much as possible [15]. 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 [22]. They are often eliminated in the pre-processing phase [23] to provide greater accuracy, but there are studies that chose to include them [24]

TABLE I. SEMANTRIA DATA CONVERSION

Original Range	**-2**	**-1**	**0**	**1**	**2**	***Conversion Method***
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*

V. METHODS

We compared commercial tools against non-commercial tools in order to assess their suitability for the healthcare domain. Two tools of each category were chosen with the main difference being that non-commercial tools, in this report, were based on supervised learning algorithms, while commercial tools used an unknown proprietary algorithm. Even though sentiment analysis is not the primary use of the non-commercial tools, they are appropriate tools for sentiment classification. Supervised learning is a common classification method, used in [25] and [26] to classify sentiment.

The commercial tools used were Semantria (a Microsoft Excel add-in) and TheySay (www.theysay.io, an online sentiment analysis tools). The non-commercial tools used were WEKA (version 3.7.10) and Google Prediction API.

The dataset used included the responses to an online survey that evaluated an online healthcare information service. The design of the service was changed and a survey was administered to evaluate the effect on the user experience. The survey consisted of eight questions (a combination of open and closed-ended), which provided 165 responses in total. However, none of the questions were mandatory and due to missing information the data consisted of the 137 responses to the question: "What is your feedback on the changes to the website?" In addition, numerical ratings for the navigation and design of the website and overall satisfaction were used in the analysis. In general, participants used a scale of 1-5, 1 representing the most positive feedback for the relevant question.

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 these were more likely to contain a single opinion. An assumption was made that single sentences would be more clearly categorised to a specific class and therefore easier for any tool to extract their polarity. It was also assumed that training a model using only single sentences might improve the accuracy and reliability of the results.

A. Commercial Tools

The sentiment polarity of the responses was found using both commercial tools. This was done with a 'click of a button' as each response was inserted and classified within seconds. These are commercial products and are meant to be easy to use. After individually evaluating each tool, a comparison between them was made. However, the format of the results differed and hence they were transformed into a uniform format for comparison. 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. Therefore, the outcome from Semantria was converted into a percentage (see Table I) in order to be compared with the dominant polarity number of each response from TheySay.

TABLE II. SENTIMENT LABELS/RANGES

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

TABLE III. NEUTRAL CLASSIFICATION GUIDELINES

Neutral Classification Guidelines	
1.	Any response that contained equal amount of positive and negative feedback. However, this was also dependent on the choice of words and how strong the negative or positive feedback was for the response.
2.	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.
3.	Any response that was considered irrelevant for the website.

TABLE IV. PROCEDURE FOLLOWED TO ANALYSE SENTIMENT BY GOOGLE PREDICTION API

Google Prediction API	
Procedure	1. PRE-PROCESSING THE DATA: Punctuation and capitals were removed. Strings were inserted into quotes and 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, where words like *'not'* were connected using an underscore with the following word [11]. 2. Import data into the tool. 3. Train the model using the training set. 4. Test the model using the testing set and comparing the predictions with the pre-labelled responses. 5. Perform a 4-fold cross validation.
Outcome	Each response was given three different values (adding-up to 1) to represent how negative, neutral and positive each response was.
Evaluation	The highest value among the three given to each response was used to the represent the polarity of each response.

TABLE V. PROCEDURE FOLLOWED TO ANALYSE SENTIMENT BY WEKA

WEKA	
Procedure	1. Choose graphical user interface: *"The Explorer"* was used since we believe it is much easier to comprehend and manipulate than the other two user interfaces, *"The Experimenter"* and *"The Knowledge Flow"*. 2. Import data into the tool. 3. Pre-processing of the data: Punctuation, capitals and stopwords were removed. Alphabetically tokenizing the data was also considered to compensate for any failure in punctuation removal. 4. Resample: Considered whether any resampling technique would positively influence the performance of the classifier, i.e. whether a balance dataset (equal number of responses in each class) was preferred to an unbalanced one. 5. Choose classifier: Naïve Bayes was used. 6. Classify the model using a 4-fold cross validation.
Outcome	The model was assessed by the percentage of correctly classified responses.
Evaluation	The percentage of correctly classified responses was used to evaluate the model's performance.

TABLE VI. THE PROCESS NEEDED TO ACHIEVE A COMMON OUTPUT FROM EACH TOOL FOR COMPARISON PURPOSES

Semantria	The single polarity suggested for each response was compared with the manually classified responses
TheySay	The dominant polarity of each response was compared with the manually classified responses.
Google Prediction API	The dominant polarity of each response, based on the highest value of classification (negative, neutral or positive) was compared with the manually classified responses.
WEKA	No action was needed. The output provided by WEKA is in the form of a single percentage representing the number of correct predictions.

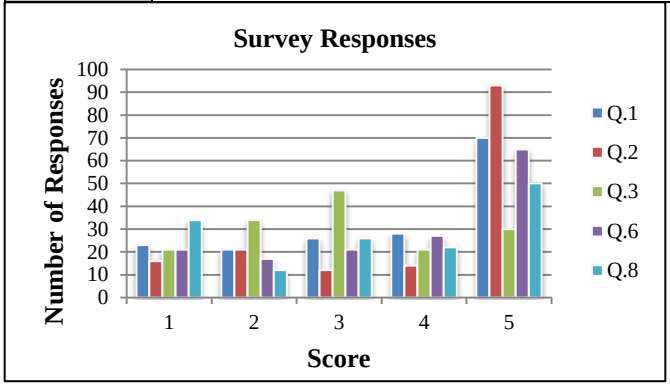

Fig. 2. Quantitative analysis of the data

B. Non-Commercial Tools

The use of supervised learning required labelling the data before training. Hence, using the quantitative (closed-ended) questions of the survey, the average score of each response was calculated. Then, using a pre-defined scale[1] shown in Table II, an initial label was assigned to each response. The labels were then manually checked to ensure they represented the correct final classification. If errors were found, the label was changed to the appropriate polarity class. For example, if the response 'I love the website' was classified as negative, it was changed to positive. The manual intervention was easy for positive and negative polarity responses, as these were more easily identified, than neutral responses. However, neutral classification guidelines were introduced to help with accurately evaluating the initial neutral labels. These guidelines can be found in Table III.

The labelled data were then randomly separated into training (75%) and testing (25%) sets. The training set was imported into the tool in order to create a model which was then tested using the testing set. The latter was done to check the accuracy of the predictions and the performance of the model. After that, cross validation followed to ensure the validity of the model. These were the major steps for both tools to analyse sentiment. However, the ordering and procedure was slightly different. Table IV and Table V show each tool's process in analysing the data given.

C. Evaluation Methods

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 compared with the baseline, i.e. the manually classified responses. The agreement between the predictions of the tool and the baseline was measured by the percentage of correctly classified responses, which was produced using (1). A statistical analysis was followed for a more advanced comparison and evaluation.

$$\frac{\textit{Number of correctly predicted responses}}{\textit{Total number of responses}} \times 100\% \quad (1)$$

Table VI shows the process needed to achieve a common output from each tool using the baseline.

VI. RESULTS

An overview of the survey responses is shown in Fig. 2.

As stated, pre-labelling the survey responses was mandatory since non-commercial tools were trained using supervised learning. The manually derived classification results (Fig. 3) were therefore used as a baseline to be compared with the results of the four tools. This ensured a consistent and accurate comparison between the four tools.

A. Commercial Tools

The classification results of the commercial tools (Fig. 4) point out the difference between each commercial 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, while TheySay favoured the negative class. Using the manually classified responses (Fig. 3) as a benchmark, we could say that TheySay performed better since the results produced are closer to the benchmark as far as the proportion of negative, neutral and positive statements are concerned. However, this could also be due to misclassification. Further statistical tests make this clearer.

Despite the classification differences, the two commercial tools provided different features to complement the results. For example, Semantria was only able to identify entities responsible for the polarity of 31 responses. However, this was not very accurate. For example, 'cancer' was recognised as a quote. This might be due to the nature of the survey and the fact that a lot of the entities were specific features of the website. Also the topic was predefined and the pronoun *'it'* was used to refer to the website. On the other hand, TheySay offers an in-depth analysis by considering the following features: POS recognition, comparison detection, humour detection, speculation analysis, risk analysis, intent analysis, named entity recognition, shallow chunk parsing and dependency parsing. In general, the features were accurate and

[1] For example, let's say a person gave the following scores to the five closed-ended questions: 2, 4, 3, 2, 5. Then we could say that this person gave an average score of $\frac{2+4+3+2+5}{5}=3.2$ to the website's features. The same person gave the following feedback: "The new look is nice but it is very frustrating that some information is not available". Combining the two, this response was classified as neutral based on Table II.

provided interesting insights on the data, particularly the POS and comparison detection features. However, this report emphasizes in classifying sentiment and hence, the features of

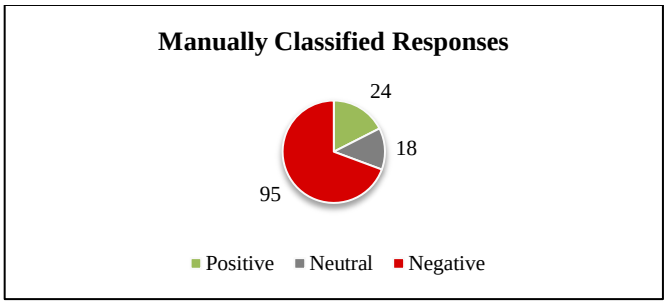

Fig. 3. Results of the manually classified survey responses

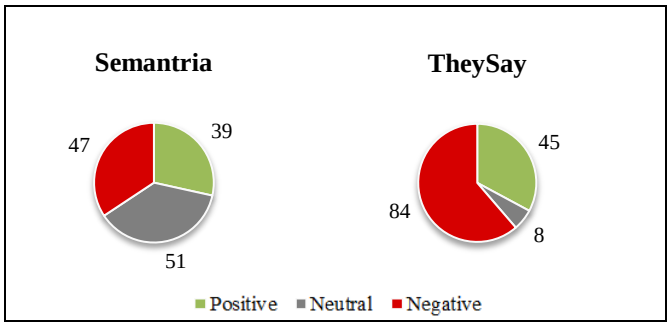

Fig. 4. Classification results of the commercial tools

each tool described above were not considered as part of our analysis. Furthermore, TheySay appeared to be case-sensitive since it was not able to identify the negativity associated with the word 'cant' while when adding the missing apostrophe, the overall classification changed.

B. Non-Commercial Tools

Importing the data into the two non-commercial tools required a different procedure. However, a consistency was kept between the number of negative, neutral and positive responses in the training sets used for the two tools. Furthermore, a 4-fold cross validation was used to measure the accuracy of the predictions of the non-commercial tools. This was done manually for Google Prediction API, where the average of the four folds was used as an indication of the tool's accuracy, as in (2), while cross validation was available within the WEKA tool.

$$\frac{Accuracy\ 1 + Accuracy\ 2 + \ldots + Accuracy\ k}{k} \quad (2)$$

where *Accuracy 1* is the accuracy of Fold 1, *Accuracy 2* is the accuracy of Fold 2, and so on.

Resampling also considered in the WEKA tool 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%. However, it was decided not to consider resampling at all due to the rather small size of the dataset. A 4-fold cross validation was used to verify the model using the Naïve Bayes classifier. MaxEnt and SVM, which are a more suitable fit for sentiment analysis according to Pang et. al. [24], are not available in WEKA. Nonetheless, Naïve Bayes performed well. It would be interesting to examine the difference between the three classifiers on data from the healthcare information service domain.

VII. Evaluation

Table VII shows the classification results of the four tools when compared with the baseline. With 82.35% accuracy, WEKA evidently produced the most precise predictions. It also proved more accurate when analysing single-sentence responses (70% compared with 54% of Google Prediction

TABLE VII. Classification Results for the Four Tools

Category	**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%

TABLE VIII. Statistical Comparison of the Four Tools Using All Responses

Tool	**Kappa Statistic**	**F-measure Weighted Average**
Semantria	0.269	0.550
TheySay	0.389	0.678
Google Prediction API	0.220	0.628
WEKA	**0.574**	**0.809**

API). Furthermore, it is interesting to see that TheySay exceeded WEKA in terms of accuracy of single- sentence responses.

We used the F-measure weighted average and the kappa statistic to account for the reliability and performance of the classifiers used among the four tools (Table VIII). Cohen's kappa statistic provides a numerical rating for the reliability of the agreement of two entities, whereby agreement due to chance is factored out. In our case, the agreement between the manual classified survey responses and each tool's classification results is measured [27]. Cohen's kappa takes values between -1 and 1, with 1 indicating perfect agreement, -1 indicating perfect disagreement and 0 indicating that any agreement is due to chance.

A kappa statistic of 0.574 indicated a high correlation between the manual classified responses and the prediction of the WEKA tool. In addition, the weighted average F-measure

(0.809) of WEKA suggested the high performance associated with the tool and hence proved that it is the most reliable tool when compared with the other three tools. It is also 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 (0.220) of Google Prediction API is the lowest among all the tools. This comes as a surprise since Google Prediction API is a non-commercial.

The weighted average F-measure and the kappa statistic were produced using each tool's confusion matrix, found in Tables IX – XII. It is observed that the majority of negative responses were predicted correctly by Google Prediction API, TheySay and WEKA, while an equal amount of negative responses was classified as negative and neutral by Semantria. Examining the wrongly classified responses, the performance of WEKA (18 wrong responses) seems relatively better compared with the other tools, following with Google Prediction API (39 wrong responses). Furthermore, neutral responses were misclassified as either negative or positive, which proves the difficulty in identifying neutral statements. Reasons why this may have happened is that Semantria, for instance, emphasises on words to extract sentiment and hence, any spelling mistakes can easily affect the polarity. Also, it was noticed that it lacks knowledge of common phrases that contribute the polarity of each response. In addition, large, explanatory responses create confusion for TheySay which sometimes leads to incorrect judgements.

Observing Fig. 5 and Fig. 6, the high precision[2] and recall[3] values (obtained using the table of confusion[4]) associated with

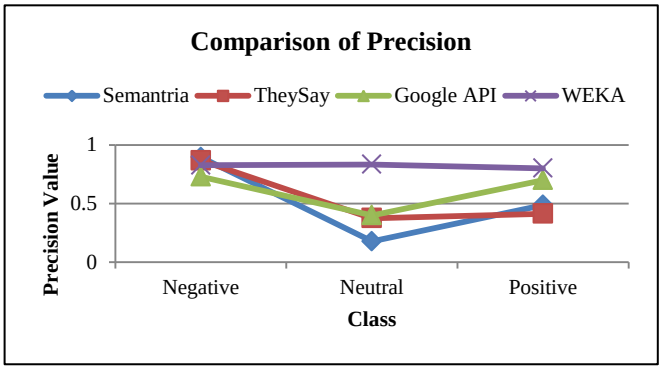

Fig. 5. Comparison of precision among the four tools

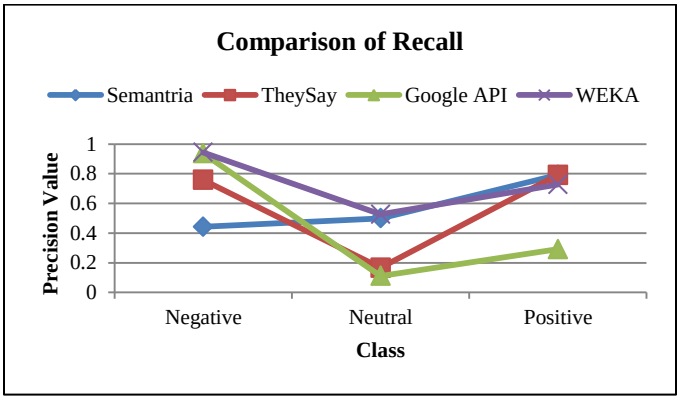

Fig. 6. Comparison of recall among the four tools

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.

TABLE IX. SEMANTRIA'S CONFUSION MATRIX USING ALL RESPONSES

Actual Class \ Predicted Class	Negative	Neutral	Positive
Negative	42	38	15
Neutral	4	9	5
Positive	1	4	19

Wrongly classified responses = 67

TABLE X. THEYSAY'S CONFUSION MATRIX USING ALL RESPONSES

Actual Class \ Predicted Class	Negative	Neutral	Positive
Negative	72	4	19
Neutral	7	3	8
Positive	4	1	19

Wrongly classified responses = 43

[2] Precision = $\frac{TP}{TP+FP}$

[3] Recall = $\frac{TP}{TP+FN}$

[4] The table of confusion is created by measuring the number of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) for each class independently. For example, the table of confusion for Semantria's positive class would consist of the following values:

TP = 19 FP = 5 TN = 93 FN = 20

TABLE XI. GOOGLE PREDICTION API'S CONFUSION MATRIX USING ALL RESPONSES

Actual Class \ Predicted Class	Negative	Neutral	Positive
Negative	89	3	3
Neutral	16	2	0
Positive	17	0	7

Wrongly classified responses = 39

TABLE XII. WEKA'S CONFUSION MATRIX USING ALL RESPONSES

Actual Class \ Predicted Class	Negative	Neutral	Positive
Negative	67	1	3
Neutral	8	5	0
Positive	6	0	12

Wrongly classified responses = 18

VIII. CONCLUSION

The aim of the project was to assess the difficulty of analysing sentiment in the healthcare domain. Commercial (non-configurable) and non-commercial (configurable) tools were evaluated on their ability to accurately extract 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 UXLabs) as positive, neutral or negative.

The results produced showed that the non-commercial tools, especially WEKA, performed best on our data. However, to generalise further from this we should ideally include a greater number of tools and use a larger dataset. It was also observed that punctuation usage, capital letters and stopwords do not impact the accuracy of the sentiment classification, as suggested by Aggarwal & Zhai [22]. Unlike most researchers, we included 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 tools when classifying neutral responses. Using a larger number of judges would provide more trustworthy results.

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 make recommendations for each tool based on the requirements of the potential user (see Table XIII).

Our initial hypothesis that commercial tools would outperform non-commercial tools was disproven. However, the modest number of tools examined allows for future work to challenge these findings. For example, Viralheat (a commercial tool that allows training a sentiment classified on user-supplied data on the social web) and TextBlob (a non-adaptive non-commercial sentiment analyser that uses a classification model that was trained on movie review corpus) would be good additions. The selection of tools was originally concluded based on popularity of use. Semantria and WEKA are widely used tools for sentiment analysis and classification [28, 29, 30, and 31]. In order to achieve a reliable comparison of the selection set, we were searching for tools that worked in a similar way as WEKA (machine learning) and Semantria ('click of a button' results). TheySay was preferred among other commercial tools due to its user-friendly interface. Similarly, Google Prediction API served a good second non-commercial tool by providing straightforward guidelines.

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. as cited by [15], might improve the accuracy. 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 improve the results produced by WEKA. Using different classifiers will of course have an effect on accuracy. In fact, even though tools have a variety of applications, the examination of different classifiers could be more beneficial in defining the accuracy of the predictions. Due to the tools' features, we could not account for this.

Basing our evaluation on the diversity of the tools used, we suggest that non-commercial tools are recommended for users that are willing to spare the extra time to achieve precision and accuracy. Although specific recommendations are provided for the use of each tool, WEKA was shown to be the best tool overall for our data.

TABLE XIII. RECOMMENDATIONS OF EACH TOOL BASED ON USER REQUIREMENTS

Tool	Recommendations
Semantria	• It is particularly recommended for business use since consumer insights can be discovered. • It can improve the decision-making process and ensure that better customer service will be achieved by tracking down feedback to the little details through the use of queries, especially negative feedback, where the origin of the problem can be found. *Benefits:* Sentiment analysis can be developed in several ways therefore, it allows the discovery of patterns and trends in the data.
TheySay	• It can aid businesses be better prepared for competition (comparative expressions are detected) and deal with sarcastic comments (humour detection) that harm reputation. *Benefits:* It takes into consideration a variety of features that improve the performance of the algorithm, like POS analysis.
Google Prediction API	• It involves a time-consuming procedure and hence it is recommended for people that are willing to spare the extra time to achieve precision and accuracy. *Benefits:* It is very straightforward to use since guidelines are available in the tool's website.
WEKA	• It could be considered the most suitable for sentiment analysis (extraction and classification) in a healthcare domain. *Benefits:* It provides an extensive analysis.

REFERENCES

[1] BYRD, R.J., STEINHUBL, S.R., SUN, J., EBADOLLAHI, S. & STEWART, W.F. (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: 28 October 2014].

[2] 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: 28 October 2014].

[3] KAMEL BOUROS, M.N., SANFILIPPO, A.P., CORLEY, C.D. & WHEELER, S. (2010) Computer methods and programs in Biomedicine. social web mining and exploitation for serious applications: Technosocial Predictive Analytics and related technologies for public health, environmental and national security surveillance. [Online] 100 (1). pp.16-23. Available from: http://www.sciencedirect.com/science/article/pii/S0169260710000386. [Accessed: 28 October 2014].

[4] 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: 28 October 2014].

[5] QUESTIONPRO. (2011) *Text analytics: visualizing and analysing open-ended text data.* [Online] Available from: http://www.questionpro.com/images/bookshelf/SurveyAnalytics-TextAnalytics.pdf. [Accessed: 28 October 2014].

[6] 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: 28 October 2014].

[7] 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: 28 October 2014].

[8] PANG, B. & LEE, L. (2008) Opinion mining and sentiment analysis. f*oundations and trends in information retrieval*. [Online] 2. p.1-135. Available from: http://www.cs.cornell.edu/home/llee/omsa/omsa-published.pdf. [Accessed: 28 October 2014].

[9] 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.

[10] 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: 28 October 2014].

[11] 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.

[12] 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: 28 October 2014].

[13] 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.

[14] 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.

[15] 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: 28 October 2014].

[16] DOBBIN, K.K. & SIMON, R.M. (2011) BMC medical genomics. *optimally splitting cases for training and testing high dimensional classifiers.* [Online] 4 (1) (31). pp.1-8. Available from: http://link.springer.com/article/10.1186%2F1755-8794-4-31. [Accessed: 28 October 2014]

[17] 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: 28 October 2014].

[18] NARAYANAN, V., ARORA, I. & BHATIA, A. (2013) Fast and accurate sentiment classification using an enhanced Naïve Bayes model. In *Intelligent Data Engineering and Automated Learning (IDEAL) 2013. pp. 194-201*. Springer Berlin Heidelberg.

[19] PATIL, G., GALANDE, M.V., KEKAN, M.V. & DANGE, M.K. (2014) International Journal of Innovative Research in Computer and Communication Engineering. Sentiment analysis using support vector machine.[Online] 2 (1) pp. 2607- 2612. Available from: http://www.ijircce.com/upload/2014/january/16K_Sentiment.pdf. [Accessed: 27 October 2014].

[20] LEE, H.Y. & RENGANATHAN, H. (2011) Chinese sentiment analysis using maximum entropy. *Sentiment analysis where AI meets Psychology*. (SAAIP) [Online] pp. 89-93. Available from: http://www.aclweb.org/anthology/W11-3713. [Accessed: 27 October 2014].

[21] DEY, L. & HAQUE, S. K. M. (2008) Opinion mining from noisy text data. In *Proceedings of the 2nd workshop on analytics for noisy unstructured text data*. Singapore, 24 July 2008. N.Y.: ACM. pp.83-90.

[22] 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: 28 October 2014].

[23] 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.

[24] 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.US: Association of Computational Linguistics.

[25] HAMOUDA, A., MAREI, M. & ROHAIM, M. (2011) Building machine learning based senti-word lexicon for sentiment analysis. *Journal of Advances in Information Technology*. [Online] 2 (4). pp.199-203. Available from: http://ojs.academypublisher.com/index.php/jait/article/viewFile/jait0204199203/3903. [Accessed 29 October 2014].

[26] MAAS, A.L. et. al. (2011) Learning word vectors for sentiment analysis. In *Proceedings of 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies.* Volume 1 Portland, Oregon, 19-24 June 2011. pp.142-150.

[27] VIERA, A.J & GARRETT, J.M. (2005) Family medicine. understanding interobserver agreement: the kappa statistic. [Online] 37 (5), pp. 360-363. Available from: http://virtualhost.cs.columbia.edu/~julia/courses/CS6998/Interrater_agreement.Kappa_statistic.pdf. [Accessed: 28 October 2014].

[28] CIEIEBAK, M., DURR, O. & UZDILLI, F. (2013) Potential and limitations of commercial sentiment detection tools. In *Proceedings of 1st International Workshop on Emotion and Sentiment in Social and Expressive Media (ESSEM): Approaches and Perspectives from AI.* Torino, Italy, 03 December 2013. pp. 47-58

[29] FRANK, E., HALL, M., HOLMES, G. & WITTEN, I.H. (2004) Data mining in bioinformatics using Weka. *Bioinformatics.* [Online] 20(15). pp. 2479-2481. Available from: http://bioinformatics.oxfordjournals.org/content/20/15/2479.short. [Accessed: 10 January 2015].

[30] TURDAKOV, D.Y. et. al. (2014). Texterra: a framework for text analysis. *Programming and Computer Software*. [Online] 40 (5). pp. 288-295. Available from: http://link.springer.com/article/10.1134/S0361768814050090. [Accessed: 10 January 2015].

[31] WHITELAW, C., GARG, N. & ARGAMON, S. (2005). Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management.* ACM. pp. 625-631.

[32] GOOGLE DEVELOPERS. (2012) *Google Prediction API.* [Online]. Google.

[33] SEMANTRIA, LLC. (2014) *Semantria.* [Excel add-in]. Semantria Inc. Available from: https://semantria.com/.

[34] HALL, M. et. al. (2009) *The WEKA Data Mining Software: An Update.* 11 (1). SIGKDD Explorations.

[35] GOOGLE DEVELOPERS. (2012) *Google Prediction API.* [Online]. Google.

[36] VIRALHEAT (2014). Viralheat Inc. Available from: https://www.viralheat.com.

[37] VassarStats: Website for statistical computation (2013). *Kappa as a measure of concordance in categorical sorting.* [Online] Available from: http://vassarstats.net/kappa.html. [Accessed: 28 October 2014].

[38] BIRD, S., KLEIN, E. & LOPER, E. (2009) *Natural Language Processing with Python.* O'reilly.

[39] BIRD, S., KLEIN, E. & LOPER, E. (2009). *Supervised classification.* O'reilly.

[40] MITCHELL, T. M. (1997) *Machine learning.* McGraw-Hill International Editions. N.Y.: McGraw-Hill, Inc

[41] NARAYANAN, V., ARORA, I. & BHATIA, A. (2013) Intelligent data engineering and automated learning. *Lecture notes in computer science.* Fast and accurate sentiment classification using enhanced Naïve Bayes model. [Online] 8206, pp. 194-201. Available from: http://link.springer.com/chapter/10.1007%2F978-3-642-41278-3_24# [Accessed: 27 October 2014].

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

[43] 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.

[44] 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: 28 October 2014].