CRITERIA-BASED OPINION MINING FROM PATIENT REVIEWS TO GENERAL PRACTICES AND CLINICS

Final Report

Information Technology Capstone Project

COMP5703/5707/5708

Group Members

1. Yuchen Zhang (470184899)
2. Danyu You (460445852)
3. Tiantui Guo (460398776)
4. Zhichao Zheng (470216400)
5. Dawun Han (460177643)

ABSTRACT

With the emergence of Web 2.0, patients are more likely to voice their reviews for general practices (GPs) and clinics in the social websites, and online free-text patient reviews have become significant sources for people who seeking for healthcare services to evaluate and select a GP or a clinic. A large number of researches have used sentiment analysis technique to detect patients' satisfaction with their experience in medical centres and many machine learning models were applied to measure the key features of customer reviews. This project aims to develop an advanced algorithm which provides a method to deep mine knowledge from filtered patient review contents. The algorithm has the ability to detect the aspects/criteria where commenters mainly focus on and identify their sentiment based on these criteria.

Keywords: criteria-based opinion mining, sentiment analysis, NLP, machine learning, SVM

TABLE OF CONTENTS

1. INTRODUCTION

With the emergence of Web 2.0, patients are more likely to voice their reviews for general practices (GPs) and clinics on social websites. Most of the people who seek medical services would like to use online reviews as the first step to evaluate the clinics and physicians. Furthermore, a large number of patients view, post comments and rate healthcare staff on social websites (Loria, 2017), such as RateMDs.com, HeathEngine, Twitter, etc. A research shows that there is a positive relation between review helpfulness and review depth (Mudambi, S. M., & Schuff, D., 2010), and such result was also proved by Chua, A. Y., & Banerjee, S., 2015. Therefore, an intelligent and efficient algorithm which can deep mine patient opinions from their reviews is extremely essential to assist medical service seekers to make better GP or clinic selections.

When studying existing online medical services rating and review websites, some observations were made:

1. There is a weak relative between ratings and reviews, so a healthcare seeker has to read text comments because ratings are so flawed (Wolff-Mann, 2016). The stars is only a part of patient reviews that does not contain too many useful information, while what patients mentioned in their comments – their health care experience – are the most valuable and helpful.
2. The overall rating of a GP or a clinic is calculated by the stars. In fact, such rating is just an average of stars in all patient reviews, while patients' written comments are never put into consideration.
3. The review ranking mechanism is too simple. Reviews are normally sorted by date, so that it is difficult for people to find helpful reviews in a sea of ratings. Useful reviews are submerged by a huge number of spam reviews and helpless comments with no valuable information (such as "Good.", "wow"). These reviews may seldom be seen, as guests have lost their patients and attention before they reach the bottom of the review list.
4. Most helpful patient reviews always focus on several aspects (Hopper, 2015). However, as the differences in professional medical training and

experience, there is a huge knowledge gap in medical service evaluation between patients and physicians (Rothenfluh, 2017). As a patient-oriented information service platform, aspects that patients are truly interested in should be understood and used to evaluate the helpfulness of patient satisfaction comment.

In this case, a computational treatment of web-based patient reviews could help healthcare managers and staff to improve their quality of medical services (Hopper, 2015). If a rating and review site contains a system that mines more information from reviewers' written comments, it would significantly help healthcare service seekers in the GP or clinic selection.

This project aims to develop an advanced algorithm which provides a method to deep mine knowledge from patient review contents. The algorithm has the ability to: 1) determine the evaluation criteria of GPs and clinics based on patient reviews; 2) analyse their sentiment towards these aspects. 26477 Data were collected from multiple trusted rating for algorithm training and testing. This algorithm could be applicated in medical rating and review websites, in order to fix the existing defects they have and significantly improve the quality of information services.

2. RELATED LITERATURE

As the development of social media, online text-based reviews have become significant resources for customers to make their purchasing decision (Chua, 2015). Such consumer reviews can reduce the uncertainty and potential risks when selecting from numerous products and services (Hong, 2017). Opinion mining model, as a popular technique, could discover the criteria of what patients concern about and make the analysis on each of them.

2.1 Literature Review

Customer opinion mining is a process of identifying the implications inside their reviews, which could classify the expressed reviews into review features and emotional polarity *(Raut, V. B., & Londhe, D. D., 2014).* According to *Singh, V., & Dubey, S. K. (2014)*, there are diverse study domains in opinion mining area, such as natural language processing, text mining, decision making, and linguistics. One of the

most common uses of these studies is to recognise the sentiment weight from customer reviews (Rouse, 2010).

Most current opinion mining methods focus on feature mining and sentiment analysis using different methods, such as machine learning algorithm and sentiment lexicon. Some current researches in this area were reviewed and listed in Table 1.

Table 1 - Literature Reviews

Articles	Initial Requirements	Feature-Based	Domain	Opinion mining approach
Penalver-Martinez, I. et al. (2014)	Sentiment words; N-grams	Yes	Movie	Domain Ontology; SentiWordNet
Jeyapriya,A.,& Selvi,C.K. (2015)	Sentiment words	Yes	Product	Naïve Bayes
Ahmad, M., & Aftab.S. (2017).	TF-IDF; N-gram Tokenizer	No	Twitter, Airline, Movie	Support Vector Machine (SVM)
Yu, B et al. (2017)	Sentiment Words	Yes	Restaurant	SVM; Bag-of-words; TF-idf
Rathan, M. et al (2018).	Labelled dataset	Yes	Product	SentiWordNet; SVM

Penalver-Martinez et al (2014) developed an ontology framework to improve opinion mining model. They employed domain ontologies and sentiment lexicon to extract features of movie reviews and identify the sentiment polarity of these reviews. For extracting the features, the domain ontology and a corpus of opinions were firstly input. By identifying the classes and object properties of these opinions, domain ontology could group the words based on their semantic distance. For identifying polarity, an innovative methodology were developed based on vector analysis. SentiWordNet was used to calculate the sentiment score of each word, and these scores would be used to determine the polarity. After that, these two methods' results would be associated to calculate the opinion scores of a review text, based on the Euclidean vector. The feature extracting accuracy could achieve at 69.72%, while the sentiment analysis accuracy was 67.2%. Although this innovative method clarified the basic structure of how to process reviews and developed basic principles of features

classification, there are two limitations: 1) The accuracy was too low to apply, 2) the ontology is static, which could not be modified towards changing requirements, such as adding new classes.

The reserch of *Jeyapriya, A.,& Selvi, C.K.(2015)* achieved a much higher accuracy by using a supervised learning algorithm – Naïve Bayes – to analysis the product reviews. In order to extract feature of reviews, a tool of frequent itemset mining was used to find frequent aspects of reviews. This tool is aiming to discover a set of properties of items from the dataset and represent the association relationships *(Djenouri, Y., Djenouri, D., Belhadi, A., Fournier-Viger, P., & Lin, J. C., 2018)*. For identifying sentence orientation, they used positive and negative reviews to train the Naïve Bayes model. The precision reached 75% for feature classification and 90% for sentiment analysis.

Ahmad, M., & Aftab, S. (2017) analyzed the performance of SVM by making comparison among three review domain datasets, including Twitter, Airline and movie reviews. They used SVM to do the sentiment analysis with different ratio of training and testing. The analysis accuracy of these three datasets was different, namely Twitter at 70.4%, Airline at 77.6%, and Movie at 78.8%. Thus, the performance of SVM varies from different domain as well as different ratio of training and testing dataset *(Ahmad, M., & Aftab, S., 2017)*. However, this research did not develope feature extraction to mine more information from reviews. Moreover, although they made a comparison of diversity ratio of training and testing dataset and identified a best ratio, there was no evidence to prove this ratio would be suitable in different domains (such as customer reviews to electronic products or medical care services).

Using the same algorithm, *Ahmad, M., & Aftab, S., Yu, B et al. (2017)* did a research on sentiment analysis to mine features from restaurant reviews. They used SVM to differentiate the negative and positive attitude towards reviews. In addition, for feature selection, they identified review features by 'Bag-of-Words' and TF-IDF respectively. 'Bag-of-words' and TF-IDF could calculate the frequency of words contained in the text and show how frequently it appears in the text (Calderon, 17). They removed irrelevant words, including stopping words, adjective, and adverb. Then, they assumed the rest of word were all related to the categories and extracted the important features. Also, they calculated the polarity score to reflect the sentiment

degree towards each of features. Overall, the accuracy for classification achieves an ideal result at 88.906%.

Rathan, M. et al (2018) classified Twitter reviews to smartphones into different aspects and identified each polarity of sentences towards these aspects as well as Emoji. The SVM classifier was developed to discover the attributes insides the product reviews. They use this classifier to analyse the sentiment of customer reviews in each aspects. The general SentiWordNet lexicon was used to detect the sentiment polarity. In order to improve accuracy, they modified a specific lexicon for each of review aspects and developed an Emoji detection. The advantage of this research is that it could obtain relatively high precision in discovering the sentiment based on aspects of product reviews. Furthermore, the spelling correction function and modifoed lexicon could help for higher accuracy.

Overall, based on these researches, the machine learning algorithms, such as Naïve Bayes and SVM, perfrom well on both review features identification and sentiment analysis. In addition, sentiment lexicon has been used frequently to implement sentiment analysis and calculating the weight of each word. Also, these researches provide the deep insight of the process flow of opinion mining, algorithm advantages and algorithm limitations separately. Thus, this project would focus on the development of machine learning algorithms and sentiment lexicon mine opinion from patient reviews to general practices and clinics.

3. RESEARCH/PROJECT PROBLEMS

In the view of the low intelligence and inefficiency of existing patient review websites, as well as the lack of algorithm research on opinion mining of healthcare services reviews, the project aims to develop an optimized algorithm for better opinion mining from patient reviews to GPs and clinics.

The development stage contains data collection, algorithms development, algorithm training and testing, followed by algorithm optimisation. the whole project lasts 13 weeks, ranging from August 30 to November 16, 2018.

3.1 Research/Project Aims & Objectives

This project aims to develop an advanced algorithm which provides a method to deep mine knowledge from patient review contents. The algorithm has the ability to identify the sentiment of patient reviews to a healthcare centre (GP or clinic), detect the

aspects where commenters mainly focus, and these aspects would be used as criteria to analysis reviewers' sentiment.

3.2 Research/Project Questions

The rating and review functions in existing patient rating and review websites are low-intelligent. They do not have the capability to analyse the opinion of a review content by neither identifying the patient's attitude nor the aspects that the comment writer cares about. These patient reviews are mostly sorted by date, rather than by helpfulness. Though some websites (such as Yelp) set a function for end-users to valid or flag reviews for an optimized and user-friendly reviews sorting, these patient comments are evaluated manually based on attitudes and preferences of individuals. Thus, an advanced algorithm to improve the intelligence of rating and review system is needed by people who seeking for medical services.

3.3 Research/Project Scope

The project could be divided into 4 stages: data collection & pre-processing, criteria identification, data labeling, algorithm testing & optimisation.

In the stage of data collection & pre-processing, patient reviews data would be crawled from multiple third-party (Google Maps, Yelp and RateMDs.com) and pre-processed for criteria identification. About 30,000 filtered data are expected to be collected, and this stage should be completed by Wednesday, September 5, 2018.

In the stage of criteria identification, pre-processed data would be used to determine the aspects/criteria that patients truly care about. This stage should be completed by Wednesday, September 19, 2018.

In the stage of data labelling, data would be manually labelled according to the criteria. The size of labelled data should be big enough (larger than 2,000) for algorithm training and testing. This stage should be completed by Wednesday, October 10, 2018.

In the stage of algorithm testing & optimisation, the algorithm would be trained using labelled data and evaluated according to testing results. Then it would be optimised by model selection and parameters tuning. The final accuracy should be not less than 85%, and this stage should be completed by Wednesday, November 7, 2018.

Table 2 - Project Scope

Stages	Expected Outcomes	Due Date
Data collection & pre-processing	About 30,000 processed data	05/09/2018
Criteria identication	Determine criteria	19/09/2018
Data labelling	More than 2000 labelled data	10/10/2018
Algorithm testing & optimisation	More than 85% accuracy	07/11/2018

4. RESOURCES

4.1 Software Tools

The following software would be used for development, project management and communication.

Table 3 - Development Tools & Materials

Development Tools & Materials	
Python3.7.1	A programming language that is easy and efficient to coding with machine learning.
Github Enterprise	A free code hosting platform for enterprise, integrating functions of version control, git management, source code, wiki, commit history and others
Natural Language Toolkit (NLTK)	A free tool to process the natural language, for example, stop word deletion, tense unification etc.
Word2Vec	A neural networks model that transfer words to text vectors and cluster those text vectors.
SentiWordNet	A dictionary that relates to lexical resource and emotion value.
Scikit-Learn	A machine learning library with algorithm packages and documents.

Table 4 - Project Management Tools

Project Management Tool	
Microsoft Project	A project management tool to draw the Gantt chart, helping team to do the dynamic progress management.

Table 5 - Communication & File Sharing Tools

Communiccation & File Sharing Tools	
Slack	A business communication tool.
Google Drive	A tool of storing file, sharing file, and synchronize files across devices.

4.2 Data Resources

The Data of patient review to GPs and clinics were collected from 3 popular online rating and review websites: 1) Yelp, 2) RateMDs.com and 3) Google Maps Reviews. These websites are three of most popular physician rating and review sites that contain millions of patient reviews of their doctor visiting experience (Houseman, 2017). Table 6 shows detailed descriptions of each data resource.

Table 6 - Data Resources Descriptions

Yelp	Internet rating and review website for small and medium business, providing highly reliable customer reviews with its advanced review filter.
RateMDs.com	A popular review website of doctors, group practices, care centers and hosptials.
Google Maps Reviews	A web mapping service, which allows users to get valuable information (including customer reviews) of business based on the location.

4.3 Opinion Mining Models

TF-IDF was used to process data before it was forwarded to classifications. Three machine learning algorithms were used in this project: Support Vector Machine (SVM), Naïve Bayes (NB) and a Logistic Regression (LR). A sentiment lexicon called SentiWordNet was also imported.

4.3.1 TF-IDF

Traditionally, word vectorization can be realized by one-hot-encoding, but such a method treats all words within a corpus equally and ignores the importance of a word or term. Therefore, we used TF-IDF (Term Frequency-Inverse Document Frequency)

which is a weight to measure how important a word is to a document in a collection or corpus to vectorize the data.

TF-IDF contains two parts: Term Frequency (TF) and Inverse Document Frequency (IDF).

Term Frequency: TF is calculated by the frequency of a term divided by the length of a document which indicated how frequently a term occurs. However, TF ignores how important a word is. For example, "a" and "the" will be more frequent than other words in a general context.

$$tf(w,d) = |\{w' \in d : w' = w\}| \; where \; w \; is \; a \; word \; and \; d = \{w_1, \dots, w_m\} \; is \; a \; document$$

As the length of each document is uneven, sublinear function should be implemented to reduce the influence brought from high frequency words.

$$\log tf(w,d) \; or \; \sqrt{tf(w,d)}$$

$$tf(w,d) = 1 + \log tf(w,d)$$

Inverse Document Frequency: IDF measures the importance of a term and complements the shortage of TF. And the importance of a term will be compromised on the frequency of this term in the corpus.

$$df(w,D) = |\{d \in D : w = d\}| \; where \; w \; is \; a \; word \; and \; D = \{d_1, \dots, d_N\} \; is \; the \; document \; corpus$$

$$idf(w,D) = \log \frac{|D| + 1}{df(w,D)}$$

The TfidfVectorizer method in Scikit-learn library is used to transform the word array to vector array. There are 21 parameters of the method, key parameters used in this project are listed as following:

- sublinear_tf: Sublinear TF can be implemented when a word is used too often in order to magnify the importance of the contextual relative words.
- norm: The length of the documents or corpus is uneven in this data collection and normalization need to be applied in the algorithm.
- ngram_range: Set the n-values for different n-grams to be extracted.

- min_df: The document frequency of a word that lower than min_df will be filtered.
- max_df: The document frequency of a word that higher than max_df will be filtered.
- stop_words: Eliminate the stop words with respect to different languages.

4.3.2 Support Vector Machine (SVM)

Introduction

Support Vector Machine (SVM), proposed by Cortes and Vapnik in 1995, is a supervised machine learning model that analyze data by classification and regression analysis. SVM is suitable to deal with the data with small sample size, nonlinear and high-dimensional pattern. In addition, SVM also shows outstanding performance when processing the sentiment classification (Bo Pang, 2002).

The mechanics of SVM is finding an optimal hyperplane to classify data into different sorts (Figure1). The basic formulations to calculate the hyperplane are followings:

- Assume hyperplane: $w^T x + b = 0$
- The boundary of hyperplane: $w^T x + b = 1$ and $w^T x + b = -1$.
- The distance between samples and hyperplane: $r = \frac{|w^T x+b|}{\|w\|}$
- The Margin between hyperplane boundary: $\gamma = \frac{2}{\|w\|}$

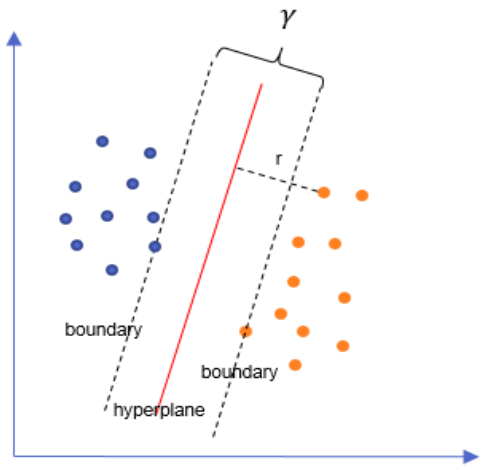


Figure 1 - Mechanics of SVM

Finally, the calculated hyperplane could be explained as following:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad ,s.t.\, y_i(w^T x_i + b) \geq 1, i = 1,2,\dots m.$$

However, considering the dual problem happened in the classify processing, the hyperplane could be operated by Lagrange function:

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j {x_i}^T x_j \quad ,s.t. \sum_{i=1}^{m} \alpha_i y_i = 0, \alpha_i \geq 0, i = 1,2,\dots,m.$$

In order to solve quadratic programming problem above, the Karush-Kuhn-Tucker condition (KKT) should be set as constraint to find the optimal α.

On the other hand, the above method is suitable for the linear divisible data. Once the samples are linear indivisible like Figure 2, it could be mapped to a higher dimensional space or infinite dimensional space and using kernel function to switch the high dimensional vector to lower dimensional inner product.

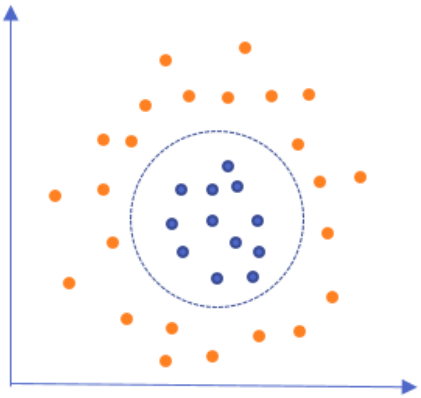


Figure 2 - Linear Separatable Dataset

According to the feature of classification and sentiment dataset, there are three types of kernel could be test during experiment:

Table 7 – Types of SVC Kernel

Kernel Function	Application Condition	Formulation
Linear Kernel	Linear divisible situation Large number of features Large sample size.	$\mathrm{K}(x_i,x_j) = x_i^T x_j$
Gaussian Radial Basis Function (RBF)	Linear indivisible situation Less feature dimensions Normal sample size	$\mathrm{K}(x_i,x_j) = \exp(-\gamma\left\|x_i - x_j\right\|^2)$
Polynomial Kernel	Linear divisible situation Image training Less limitation than Linear	$\mathrm{K}(x_i,x_j) = (\zeta+\gamma x_i^T x_j)^Q$

The Sequential Optimization (SMO)

SMO was proposed by Platt in 1998, it is an efficient optimization algorithm to solve the dual problem. The computational cost of traditional SVM algorithm is in direct proportion to the size of dataset, is not suitable to process the large-scale data. However, the SMO base on the KKT constraint conditions, randomly select α_i and α_j for each iteration, using the coordinate assent algorithm to optimize the target function. So that the computational efficiency could be raised when solve the convex quadratic programming problem.

SMO was proposed by Platt in 1998, it is a efficient optimization algorithm to solve the dual problem. The computational cost of traditional SVM algorithm is in direct proportion to the size of dataset, is not suitable to process the large-scale data. However, the SMO base on the KKT constraint conditions, randomly select α_i and α_j for each iteration, using the coordinate assent algorithm to optimize the target function. So that the computational efficiency could be raised when solve the convex quadratic programming problem.

4.3.3 Naïve Bayes (NB)

Naïve Bayes (NB) is a method to construct a sophisticated classifier based on Bayes' theorem. NB has demonstrated effective results as a classifier in practice such as text classification and automatic medical diagnosis (Rish, 2001). It requires less

training data and time so that it merges quicker than discriminative models. Thus, it is suitable for quick changes and new datasets to adapt.

NB can be implemented to learn classifiers by assuming features independently. It can have high accuracy of classification in that it assumes all attributes are independent which is called "Naive Bayes Assumption". It makes parameters can be implemented separately and high accuracy. Naïve Bayesian classifier can calculate the probabilities of the number of sentiments which are positive and negative. Firstly, reviews having positive and negative sentiments are separated in different text files. The review sentences are divided into a group of the first two words and a single word. And then the group of the first two words are checked to delete if they are matched. When the words are in positive opinion words, it increments the probabilities of positive sentiment. Otherwise, it counts get the probabilities of negative sentiment. Lastly, the review sentence is positive if the probabilities of positive sentiment is greater than that of negative sentiment. In this way, the review is negative when the probabilities of negative are predominant comparting that of positive sentiment. Therefore, reviews can be analysed into positive or negative opinion by NB.

A classifier has a class with discriminant functions that defines itself. Each class i can be collaborated with a discriminant function $f_i(x)$. Also, the deterministic function $h(x) = argmax_{i\in\{0,\dots,m-1\}} f_i(x)$.

Applying a feature vector $f_i^*(x) = P(C = i|X = x)$, Bayes classifier functions can be calculated as $h^*(x) = argmax\ P(X = x|C = i)P(C = i)$.

Finally, assumption that features are independent in the class is commonly used to avoid hard training. Through the assumption, the NB classifier NB(x) can be defined as follows:

$$f_i^{NB}(x) = \prod_{j=1}^{n} P(X_j = x_j|C = i)P(C = i)$$

According to Rish (2011), the naïve Bayer classifier is effective in that features are independent mutually and features are dependent functionally. Even though its estimation is incorrect, the NB has successful results of classification with understanding of dataset.

4.3.4 Logistic Regression (LR)

Logistic regression is a probability and statistic classification model. It could be used in machine learning to solve the problem of Binomial Logistic Regression and Multinomial Logistic Regression. LR could process the classification problem with low computing cost. This section would focus on Multinomial logistic regression to implement multi-class classification.

Based on the theory of linear regression, logistic regression should use Softmax classifier to implement Multinomial Logistic Regression. The basic principle is dividing multinomial tasks into multiple binomial tasks. Towards each of binomial task, the model would be trained for classifier. The final result of each binomial model would be integrated to get the classification. The method for dividing multi-tasks should use One-vs-All. The principle of One-vs-All is considering the sample of type i as positive examples, while the rest of the samples consider as negative examples. In this case, when inputting the samples to make the prediction of classification, multiple classifiers could generate multiple classification results.

As the function of Multinomial Logistic Regression, the formula of Softmax is:

$$P(y = i|\, x;\, \theta) = \frac{\exp(\theta_i^T x)}{\sum_{k=1}^{K} \exp(\theta_k^T x)}$$

In this function, k=1,2…K, $\theta_i^T x$ is multi-input to the model, θ^T is the best training threshold. When K=2, Multinomial Logistic Regression would be same with Binomial Logistic Regression.

In order to get the threshold θ for each of classifier, it is essential to use One-vs-All method to train the K training datasets. When inputting feature vectors x of testing samples, towards each of type k, Softmax function could calculate $P(y = i|x)$. The classification with highest estimation probability is the class of feature vector x.

4.3.5 Sentiment Lexicon – SentiWordNet

Sentiment lexicon is a commonly used approach to identify sentiment orientation by weighing words, with no need for data pre-processing nor classifier training (Dhaoui, C., Webster, C. M., & Tan, L. P., 2017). It relies on a lexical dictionary which contains multiple opinionated terms (Hamouda, A., & Rohaim, M., 2011). SentiWordNet is one of the sentiment lexicons to provide the opinion information for

each of English opinionated terms. In this project, the SentiWordNet 3.0 were chosen for sentiment analysis.

SentiWordNet 3.0 contains 28431 terms extracted from WordNet. Table 8 shows some samples of SentiWordNet polarity, with attributes of part-of-speech (POS) of the term, term ID, positive score, negative score, synonyms and description of synonyms. Each synset would consist of synonyms and a unique number.

Table 8 - SentiWordNet Polarity Smaples

POS	ID	PosS	NegS	Synonyms#rank	Gloss
a	00218440	0.75	0	sightly#1 fair#3	"young fair maidens"
a	00244054	0	0	fairish#2 fair#10	(used of hair or skin) pale or light-colored; "a fair complexion"

There are three types of objects in sentiment score, including positive score (PosScore), negative score (NegScore) and objective score (ObjScore). These three attributes have been assigned the scores by the following equation:

$$PosScore + NegScore + ObjScore = 1$$

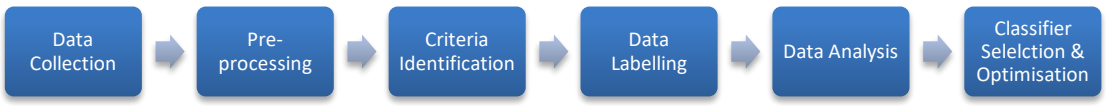
Most of time, a term has more than one senses. The term “fair”, for example, has 9 meanings as an adjective, has 8 meanings as a noun, it, and has 2 meanings as a verb. Although there is no need to determine what the actual meaning of this word in a specific sentence, it is essential to calculate the average score towards diverse senses as its sentiment score. The final score of each term could be calculated by the following formula:

$$Score = \sum_{r=1}^{n} (PosScore(r) - NegScore(r))/r \ , r = \text{synonyms's rank}$$

5. METHODOLOGIES

5.1 Methods

The methodologies contain 6 steps: 1) data collection, 2) data pre-processing, 3) criteria identification, 4) data labelling, 5) data analysis and 6) classifier selection & optimisation. Steps were progressed as components in a pipeline, the outcomes of each component would be the input of the next step.



5.2 Data Collection

The reviews of GPs and clinics within Australia would be crawled from RateMds, downloaded from Yelp Developers and collected from Google Maps.

A crawler was developed to crawl reviews text from Google Maps, and 2,878 patient reviews to 210 GPs and clinics in Sydney and Melbourne were collected. 1,053 reviews to 415 GPs' or clinics were also crawled from RateMds. Yelp open dataset with JSON format was downloaded from Yelp Developers, then reviews to GPs and clinics was filtered. Finally, we built a raw dataset with 26,477 patient reviews. All the collected data was merged into a CSV file to use in following steps.

Table 9 - Data Collection

Data Resources	Number of Reviews
Google Maps	2,878
RateMDs.com	1,053
Yelp Developers	22,546

5.3 Data Pre-processing

This step is one of the most important phases of the methodology pipeline, in which collected reviews are separated into sub-sentences and then are normalised. In this step, two processed datasets were built: 1) a separated reviews dataset for data labelling, 2) a normalised reviews dataset for criteria identification.

5.3.1 Reviews Separation

Normally, a helpful patient review contains more than one sentences, and these sub-sentences usually refer different aspects of the experience of their doctor visit, so determining categories of a whole review is a multi-label classification problem, and most machine learning models are not good at solving such problem. After dividing reviews into several review chips, each line of data would only be classified into one aspect. Thus, a multi-label classification problem was transferred into a single-label classification problem, and most machine learning models (such as SVM and NB) perform well on single-label classification. An example is shown in Figure 3.

Multi-label Classification Problem

Review	Aspects		
	Cleanliness	**Staff**	**Price**
What impressed me most is their clean room. Also, every staff there was very kind and friendly. Lastly, the price is affordable.	1	1	1

Single-label Classifcation Problem

Review	Aspects		
	Cleanliness	**Staff**	**Price**
What impressed me most is their clean room.	1	0	0
Also, every staff there was very kind and friendly.	0	1	0
Lastly, the price is affordable.	0	0	1

Figure 3 -Review Separation Transfers Multi-label Classification Problem into Single-label Classification Problem

After separating reviews, 26,477 data were divided into 200,980 review chips, and this separated dataset would be used in data normalisation and data labelling.

5.3.2 Normalisation

The separated reviews data was then normalised by stemming and stop-word removal.

Stemming is a necessary process in the area of Natural Language Processing (NLP), It translates words with same base/stem into a common form (Ahmad,2017). For example, the word “receptionists” would be stemmed into “receptionist” and the word “testing” would be stemmed into “test”. In our project, NLTK was used in the stemming phrase.

Then punctuations and stop-words were removed from stemmed data. Punctuations and stop-words are commonly used in natural language, they carry less information than keywords, and they sometimes could be distracting and non-informative (RxNLP, 2016). Table 10 shows an example:

Table 10 - Normalisation Process

Raw Review Chip	Dr. Kwiatkowski was cool and understanding, he took notes and he listened to my situation.
Stemmed Review Chip	dr. kwiatkowski was cool and understand, he took note and he listen to my situation.
Review Chip After Punctuations & Stop-words Removal	dr kwiatkowski cool understand took note listen situation

Review text are shortened from 15 words (and 3 punctuations) to 8 words, only about half of the content concludes keywords with valuable information. The normalisation phrase is very useful for following steps. It stems words for easier words counting and it reduces the gaps between keywords for better word vectors computing. After the normalisation process, 200,980 separated review chips were normalised into 200,980 narmalised data for criteria identification and algorithms training and testing.

5.4 Criteria Identification

As the reviews data had been pre-processed, the key features (or criteria) that patient truly care about should be figured out from the data, and these criteria are the core elements for data labelling and therefore have a significant impact on algorithm training and testing. Words in normalised data were first transformed into word vectors by a Word2Vec tool, then 56 keywords with high frequency were selected from word vectors and mapped into 2-dimension coordination, finally, these words were grouped into 5 clusters by K-means clustering. We summarised words in each cluster and finally identified 5 criteria that patients use to evaluate a GP or a clinic: 1) Staff, 2) Appointment & Waiting Time, 3) Value & Price, 4) Professionalism & Treatment, 5) Environment & Facility.

5.4.1 Word Embeddings

The Word2Vec model in gensim was used to transfer words into word vectors. The parameters setting is shown in Table 11.

Table 11 - Word2Vec Parameters Setting

size	window	min_count	workers
30	15	200	4

After these parameters were applied, 649 word vectors with a dimension of 30 were computed, the frequency of each word is higher than 200.

5.4.2 Keywords Selection

In the embedded words list, most of the words are commonly used words in all domain and have little relative to GPs and clinics (such as “day”, “love”, “person”, etc). We selected 56 keywords (Table 12) with the highest relative with the topic of medical care and forward their vectors to the K-means clustering.

Table 12 - Selected Keywords

staff	bill	charge	schedule	technology
helpful	insurance	fee	time	tv
assistant	pay	cost	room	professionalism
care	paid	money	food	physicians
polite	reschedule	credit	clean	professional
gentle	wait	dollar	environment	genuine
manner	copay	price	equipment	conversation
compassionate	refund	dirty	location	knowledge
book	nurse	communication	respect	
medicine	recovery	treatment	prescription	
acupuncture	diagnosis	appointment	medication	
appt	therapy	symptom	ultrasound	

5.4.3 K-means Clustering

56 selected keywords were clustered by K-means clustering model in sci-kit learn, and the best K value was measured by Silhouette analysis. Silhouette analysis is used to calculate the separation distance between the resulting clusters (Kou, 2014), and therefore to help measure the optimal number of clusters. The K value was set ranging from 2 to 10, and the calculation process was iterated 50 times for each K value. The analysis results are shown in Figure 4.

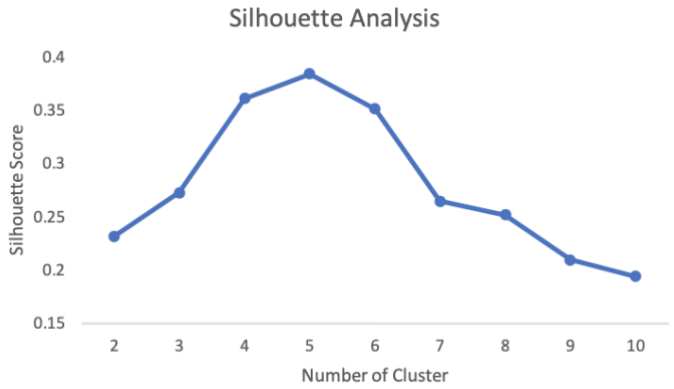


Figure 4 - Silhouette Analysis Result

By observing the Silhouette analysis result, the Silhouette score peaks when K=5. According to the definition of Silhouette Coefficient, the optimal number of cluster is 5. Then the 30-dimension vectors of 56 keywords were mapped into a 2-dimensional plane, and the clustering result was visualised in a coordinate system

(Figure 5). We summarised words in each cluster and finally identified the 5 criteria that the patient mainly mentioned in their reviews (Figure 6). The mapping of all word vectors is shown in **Appendix A**.

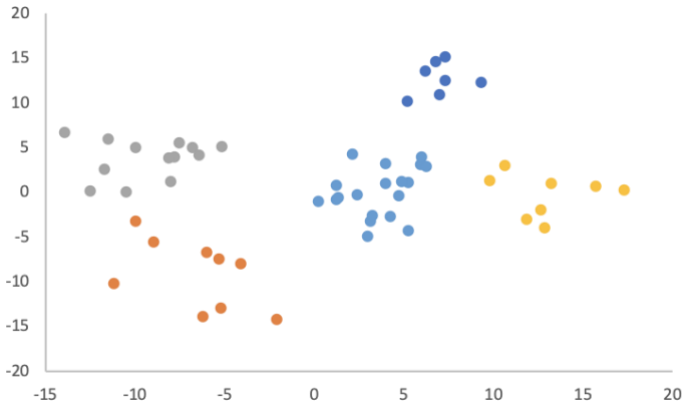


Figure 5 - Clustering of 56 Keywords

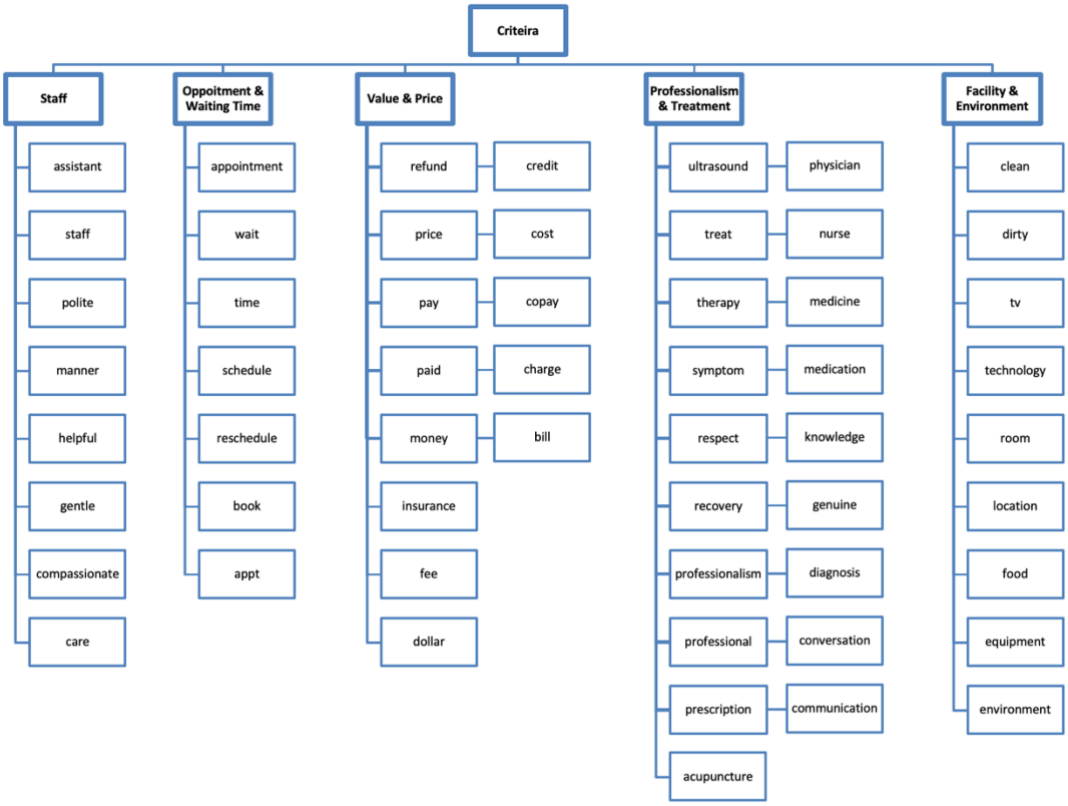


Figure 6 - Five Criteria

These five criteria would be the key labels in the next project phase – data labelling: separated review chips would be labelled according to the criteria.

5.5 Data Labelling

The separated patient review chips data was labelled manually based on five criteria (Staff, Appointment & Waiting Time, Value & Price, Professionalism &

Treatment, and Facility & Environment) as well as sentiment (Positive and negative). There are some sample labels shown in Table 13. “Clean room” was referred in the first review chip, so the criterion of “Facility & Environment” was labelled as “1” (True), and sentiment was labelled as “1” (Positive); In the third review, The reviewer mentioned “the price is too high”, so the criterion of “Value & Price” was labelled as “1” (True), and sentiment was labelled as “-1” (Negative).

Table 13 - Manual Labelling Samples

Review Chips	Staff	Appointment & Waiting Time	Value& Price	Professionalism & Treatment	Facility & Environment	sentiment
What impressed me most is their clean room.	0	0	0	0	1	1
Also, every staff there was very kind and friendly.	1	0	0	0	0	1
however, the price is too muh high.	0	0	1	0	0	-1

To build precise results of labelling, cross labelling is implemented among five project group members. This labelling method is an efficient way to improve the quality of labelling. Same data was labelled by five members, and their labelling results were compared to defect the differences, then different labels for the same data were corrected.

After cross labelling, 5,684 lines of review chips (including all-zeroes data) was labelled. Then these labels were further processed and restructured for data analysis. The labelled data was processed by following steps:

1. Labelled data was sliced into to sub-data: 1) criteria labels with fields of reviews and 5 criteria, 2) sentiment labels with fields of reviews and sentiment
2. Then, data with all-zeroes labels were removed from these two sub-datasets, as these data are normally story talking or conversations that have no relative to any criteria nor have no sentiment. For example: “I went upstairs and had a seat” or “could I have a look at your id card”.
3. Labels then were mapped to the narmalised data processed in the normalisation phase, as the review chips in labelled datasets are separated

reviews with punctuations and stop-words, such reviews could not be put into machine learning models for training and testing.

At the end of this phase, two labelled datasets were built for different data analysis purposes: 1) 2,049 criteria labels for criteria classification, 2) 2,395 sentiment labels for sentiment analysis. The data distribution of each dataset is shown in Figure 7.

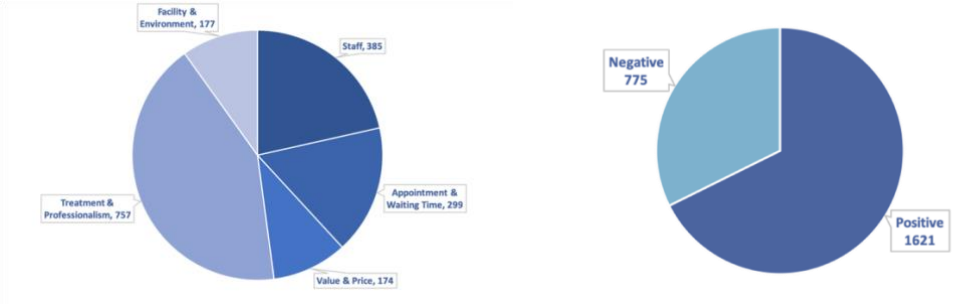


Figure 7 - Data Distribution in Criteria Labels and Sentiment Labels

5.6 Data Analysis

5.6.1 TF-IDF

The TfidfVectorizer was used at first to translate review data in text format into vectors. The default value of parameters will be set to transform the two labelled data. The output would be a matrix which row represented each document and column represent the tf-idf value of the corresponding term. The total number of rows equal to the number of documents fed in and the total number of columns equal to the length of the word set that only contained unique term.

5.6.2 Criteria Classification

Due to the data contains multi-class, the OneVsRestClassifier method of Scikit-learn library needs to be used to finish the multi-class classification task. This method loops over all the classes and the selected class is fitted against all the other which becomes a binary classification task for each class. Therefore, the output of each classifier contains 5 classification result: "staff", "Appointment & Waiting Time", "Value & Price", "Treatment & Professionalism" and "Facility & Environment". In order to compare the outcome of three classifiers which are SVC and Multinominal Naïve Bayes, parameters are set as default value. And the parameter, “Solver”, of Logistic Regression was set to “Sag” which is used for multi-class classification and remained default value for all the other parameters.

5.6.3 Sentiment Analysis

In sentiment analysis task, there are two methods has been used for classification. The first one is SVC and another one is SentiWordNet. In these two classifiers, parameters are set as default in order to compare the result.

5.7 Classifier Selection & Optimisation

For the parameters tuning, the GridSearchCV tool of Scikit-learn is used to tune the parameter mention in the above section. GridSearchCV is an exhaustive search looping over pre-defined values for an estimator. In this case, the pre-defined value of parameters for tuning are set in the following scope.

Table 14 - TfideVectorizer Parameters Tuning

Parameters	Valuses for Tuning
max_df	[0.1, 0.15, 0.2, …, 0.95]
min_df	[2, 3, 4, … , 10]
ngram_range	[(1, 1), (1, 2), (1, 3), (1, 4)]
stop_words	[None, "English"]
norm	['l1', 'l2']
sublinear_tf	[True, False]

Table 15 - SVC Parameters Tuning

Parameters	Valuses for Tuning
c	$[2^{-7}, 2^{-6}, \dots, 2^{6}, 2^{7}]$
kernel	["linear", "poly", "rbf"]
degree	[1, 2, 3, 4]

6. RESULTS

6.1 Analysis Approaches

This section mainly focuses on the results comparison and evaluation. Four evaluation methods were imported to measure the performance of the algorithms: 1)

Accuracy, 2) Precision & Recall & F-1 measurement 3) ROC Curve and 4) Confusion Matrix.

Accuracy illustrates the ratio of labels that are labelled correctly by a classifier in the total testing labels. Its formula is shown below:

$$accuracy(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{sample}-1} l(\hat{y}_i = y_i)$$

Precision and recall are calculated according to confusion matrix (Figure 8):

		Predicted class	
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Figure 8 - Confusion Matrix

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. It can be calculated by following formula:

$$Precision = \frac{TP}{TP + FP}$$

Recall is the ratio of correctly predicted positive observations to the all observations in actual class = yes. It can be calculated by following formula:

$$Recall = \frac{TP}{TP + FN}$$

F1 Score is the weighted average of Precision and Recall, and it can be calculated by following formula:

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

6.2 Criteria Classification

6.2.1 Algorithms Comparison

Criteria labelled dataset was used in criteria classifier models. It was divided into 2 sub-datasets with the ratio of 80:20 for training and testing respectively.

There were 3 classifier models evaluated in this section: SVM, Naïve Bayes and Logistic. All of these 3 classifiers were trained and tested with default parameters, and their prediction accuracy is shown in Table 16. SVM classification model conducts the best average accuracy (90.85%) and performs well on classification on each specific criterion, while the average accuracy of LR and NB is 88.59% and 85.57%.

As the best results of SVM classification, we did further evaluation and optimised it by tuning parameters for a better performance.

Table 16 - Accuracy Comparison

	SVC	**Naïve Bayes**	**Logistic Regression**
Staff	92.69%	85.23%	90.05 %
Appointment & Waiting Time	94.09%	87.56%	91.14%
Value & Price	95.33%	91.60%	91.91%
Treatment & Professionalism	78.23%	72.78%	78.54%
Facility & Environment	93.93%	90.67%	91.29%
Average Accuracy	90.85%	85.57%	88.59%

6.2.2 Algorithm Optimisation

Shown in Table 17, The weighted average of Precision, Recall and F1-Score of unoptimized SVC is 74%, 72% and 72% respectively. Furthermore, the Precision and Recall is not balance in criteria of “Value & Price”, “Treatment & Professionalism” and “Facility & Environment”, which results in low F1-Scores (54%, 67% and 59% respectively). The reason of it is because the training data is unbalance: The label size of “Treatment & Professionalism” takes about 40% of the dataset while labels of “Value & Price” and “Facility & Environment” takes the least parts. The result of this unbalance is that the there are higher possibility for SVC to incorrectly label a data as “Treatment & Professionalism” and to label data of “Value & Price” or “Facility & Environment” as another criterion. It causes high Precision with low Recall in “Treatment & Professionalism” and low Precision with high Recall in “Value & Price” and “Facility & Environment”.

To improve the performance of SVC, we tuned some key parameters of TfideVectorizer and SVC. Also, we balanced the training and testing dataset.

Table 17 - Results of Unoptimized SVC

	Precision	**Recall**	**F1-Score**
Staff	85%	78%	81%
Appointment & Waiting Time	85%	78%	81%
Value & Price	64%	47%	54%
Treatment & Professionalism	56%	82%	67%
Facility & Environment	71%	51%	59%
Micro Average	72%	72%	72%
Macro Average	72%	67%	68%
Weighted Average	74%	72%	72%

When $sublinear\_tf = True$, $noem = l2$, $max\_df = 0.25$, $min\_df = 2$, $ngram\_range = (1,1)$, $stop\_words = None$ in TfideVectorizer, $c = 0.4$, $kernel = "Linear"$ and $degree = 1$ in SVC, the model shews the best result (shown in Table 18). The weighted average of Precision, Recall and F1-Score of optimised SVC is 77%, 73% and 73% respectively. The ROC curve and confusion matrix are shown in Figure 9 and Figure 10.

Table 18 - Results of Optimised SVC

	Precision	Recall	F1-Score
Staff	76%	81%	78%
Appointment & Waiting Time	81%	75%	78%
Value & Price	87%	66%	75%
Treatment & Professionalism	60%	74%	66%
Facility & Environment	77%	55%	64%
Micro Average	73%	73%	73%
Macro Average	78%	69%	72%
Weighted Average	77%	73%	73%

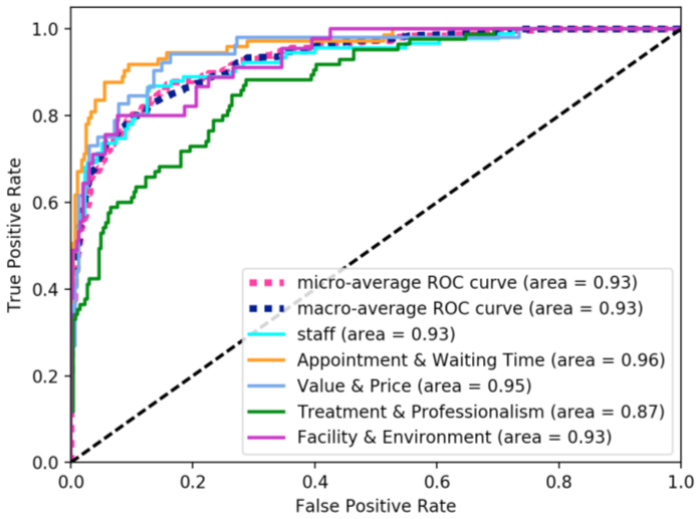


Figure 9 - ROC Curve of Optimised SVC

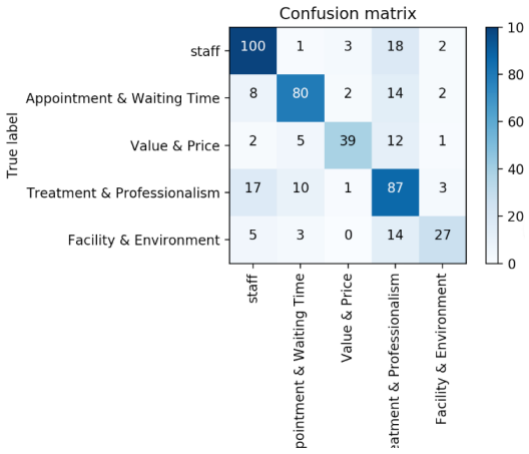


Figure 10 - Confusion Matrix of Optimised SVC

6.3 Sentiment Analysis

6.3.1 Algorithms Comparison

Sentiment labelled dataset was used in criteria classifier models. It was divided into 2 sub-datasets with the ratio of 80:20 for training and testing respectively.

SVC and SentiWordNet were used to analyse review sentiment, and both of them use the default parameters. The average accuracies are shown in Table 19. SVC model performed far better than SentimWordNet (81.04% compared with 77.22%), so we selected SVC as the tool of sentiment analysis and optimised it.

Table 19 - Accuracy Comparison

	SVC	**SentiWordNet**
Average Accuracy	81.04%	77.22%

6.3.2 Algorithm Optimisation

Shown in Table 20, the weighted average of Precision, Recall and F1-Score of unoptimized SVC is 79%, 79% and 79% respectively, and Precision and Recall of negative sentiment is too low and unbalance (only 73% and 64%). Thus, we tuned the parameters of TfideVectorizer and SVC for a ideal result.

Table 20 - Results of Unoptimised SVC

	Precision	Recall	F1-Score
Negative Sentiment	73%	64%	68%
Positive Sentiment	82%	87%	85%
Micro Average	79%	79%	79%
Macro Average	77%	76%	75%
Weighted Average	79%	79%	79%

When $sublinear\_tf = True$, $noem = l2$, $max\_df = 0.25$, $min\_df = 3$, $ngram\_range = (1,3)$, $stop\_words = None$ in TfideVectorizer, $c = 0.4$, $kernel = "Linear"$ and $degree = 1$ in SVC, the model shew the best result (shown in Table 21). The weighted average of Precision, Recall and F1-Score of optimized SVC is 85%, 85% and 85% respectively. The confution matrix is shown in Figure 11.

Table 21 - Results of Optimised SVC

	Precision	Recall	F1-Score
Negative Sentiment	80%	83%	82%
Positive Sentiment	88%	86%	87%
Micro Average	85%	85%	85%
Macro Average	84%	84%	84%
Weighted Average	85%	85%	85%

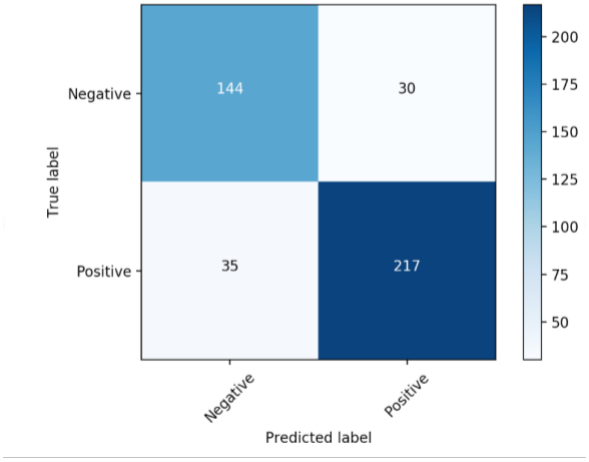


Figure 11 - Confusion Matrix of Optimised SVC

7. DISCUSSION

This project used different machine learning algorithms and sentiment lexicon to mine information from patient reviews to GPs and clinics. By comparing and evaluating the results of these algorithms, the SVC shew a remarkable performance on both criteria classification and sentiment analysis with the average accuracy of 90.85% and 81.04%. We then optimised the SVC model, and achieved ideal results: 1) Precision, Recall and F1-Score is 77%, 73% and 73% for criteria classification and 2) is 85%, 85%, 85% for sentiment analysis.

Furthermore, 26477 patient reviews were collected from 3 popular rating and review websites and two labelled datasets with the size of 2049 and 2395 were built. In this project, 5 criteria of the patient review were identified by a novel computing method.

The optimised opinion mining algorithm can be applied in existing rating and review websites to build a more intelligent review system for better healthcare information services. This advanced algorithm is expected to improve the customer experience of people who seek a good GP or a clinic.

8. LIMITATIONS AND FUTURE WORKS

There are some limitations in this project that could be studied in the future:

1. Reviews were only labelled as “Positive” or “Negative”, the label of “Neutral” was not taken into consideration

2. Due to the limited time and cost, only about 4500 data were labelled by group members. If the size of labelled dataset grows, some deep learning algorithms are expected to evaluated
3. The methodology was developed focusing on patient reviews to medical services. The adaption to other domains had not been assessed (such as movie reviews or product reviews)
4. The imbalance labelled data did influence the results of the SVC model, a larger, more balance training data are expected to improve the algorithm performance.

9. PROJECT MANAGEMENT

9.1 Roles & Responsibilities

There are 5 main roles in the project, the responsibilities of each role are showed below:

Project manager: The project manager has an overall perspective of the whole project, make the schedule of the progress and identify the goal of each sub-tasks. He or she is responsible for managing the design, development, implementation and maintenance of applications in support of audio and web meeting management. A manager should manage the design and maintenance of Internet and intranet websites and databases, including coordinating with the supervisor, tutors and other agencies.

Data scientist: The data scientist should utilize the crawler technology to collect the review information of clinic and general practice in Australia, conducting preliminary screening of data and clean the useless and inaccuracy data. Be responsible for the data processing and data analysis.

Issue tracker: The issue tracker is responsible for communication with every other group member, allocating weekly task, clarifying issues of the project, and launch them with proper tools (such as Github). Rank these issues to figure out the most urgent issues and track them to ensure they can be solved in time.

Algorithm engineer: The algorithm engineer is mainly responsible for studying the sentiment analysis algorithms, extracting the emotional factors of users from existing clinic and general practice comments, building sentiment analysis algorithms from emotional vector, emotional value and the reason of emotion three parts.

Tester: The web application tester is in charge of the unit testing, integration testing, data testing and algorithm testing, ensuring the well performance of each technical outcomes. Testers also need to closely communication with developer and support testing report to the project manager.

The roles and responsibilities are arranged as following:

Name	Roles
Yuchen Zhang	Project manager Algorithm engineer
Danyu You	Data scientist Tester
Tianrui Guo	Algorithm engineer Issues tracker
Zhichao Zheng	Data scientist Algorithm engineer
Dawun Han	Data scientist Tester

9.2 Milestones / Schedule

Table 22 shows the details of project milestones ans schedule, and the work breakdown structure (WBS) is attached in **Appendix B**.

Table 22 - Milestones and Schedule

Milestone	Tasks	Reporting	Date
Week-1	Initalisation Defining Project	Meet with the supervisor for the research direction	01-08-2018
Week-2	Defining Project	Meet with the supervisor for the research direction	08-08-2018
Week-3	Project Planning Defining Project: as the subject had not been determined, we spent the first 2 days of this week to work on it	Meet with the supervisor for the research direction WBS	15-08-2018

Week-4	Proposal Report Writing	Meet with the supervisor to modify the report Final edition of proposal report	22-08-2018
Week-5	Proposal Report Due Collect Data Develop Algorithm	Meet with the supervisor for data resources or advices	30-08-2018
Week-6	~~Execute Data~~ data would be excuted in various ways for different purpose. Data execution fot review aspects were moved to next week; taks of algorithm training were moved to week 9 and 10, after data labelling Pre-process Data: the raw data could not be executed directly, so the data pre-processing is needed before algorithm training ~~Web APP Development~~ as the search objectives switch to the algorithm development, the web development were unnecessary	Meet with the supervisor to dicuss the quality of data	07-09-2018
Week-7	Execute Data: Identify Review Aspects Manual Labelling: as there were not existing data for this project, we collected data from multiple review websites. So the data processing and labelling were needed	Meet with the superviser the discuss the result (5 aspects) and ask for some guidline of manual labelling	14-09-2018
Week-8	Manual Labelling Progress Report Writing	Meet with the supervisor to discuss the result of labelling and the group report	19-09-2018
Week-9	Progress Report Due Manual Labelling Execute Data: Algorithm Training and Testing	Meet with the supervisor to discuss the training result	05-10-2018
Week-10	Execute Data: Algorithm Training and Testing Optimise Algorithm: if the accuracy of the algorithm does not reach the expected feature, the optimization will be needed, or it would be optiomal	Meet with the supervisor to discuss the outcomes and adivices for optimisation	12-10-2018
Week-11	Final Report Writing Final Presentation Preparing Optimise Algorithm	Meet with the supervisor to discuss the optimisation results and advices of presentation draft	19-10-2018

	Final Presentation	20 minutes presentation	09-11-2018
	Final Report (thesis)	Final edition of the final report	16-11-2018

REFERENCES

Ahmad, M., & Aftab, S. (2017). Analyzing the performance of SVM for polarity detection with different datasets. International Journal of Modern Education and Computer Science, 9(10), 29-36. doi:10.5815/ijmecs.2017.10.04

Bakhsh, W., & Mesfin, A. (2014). Online ratings of orthopedic surgeons: analysis of 2185 reviews. Am J Orthop, 43(8), 359-63.

Bassig, M. (2015, February 24). 12 Reviews Sites Doctors, Hospitals, and Healthcare Marketers Need to Track. Retrieved from Review Trackers: https://www.reviewtrackers.com/doctor-review-sites/

Calderon, P. (17, 05 03). Bag of Words and Tf-idf Explained. Retrieved from Data meets Media: http://datameetsmedia.com/bag-of-words-tf-idf-explained/

Chua, A. Y., & Banerjee, S. (2015). Understanding review helpfulness as a function of reviewer reputation, review rating, and review depth. Journal of the Association for Information Science and Technology, 66(2), 354-362.

Dhaoui, C., Webster, C. M., & Tan, L. P. (2017). Social media sentiment analysis: Lexicon versus machine learning. Journal of Consumer Marketing, 34(6), 480-488. doi:10.1108/JCM-03-2017-2141

Djenouri, Y., Djenouri, D., Belhadi, A., Fournier-Viger, P., & Lin, J. C. (2018). A new framework for metaheuristic-based frequent itemset mining. Applied Intelligence, 48(12), 4775-4791. doi:10.1007/s10489-018-1245-8

Hamouda, A., & Rohaim, M. (2011, January). Reviews classification using sentiwordnet lexicon. In World congress on computer science and information technology. sn.

Hong, H., Xu, D., Wang, G. A., & Fan, W. (2017). Understanding the determinants of online review helpfulness: A meta-analytic investigation. Decision Support Systems, 102, 1-11.

Hopper, A. M., & Uriyo, M. (2015). Using sentiment analysis to review patient satisfaction data located on the internet. Journal of health organization and management, 29(2), 221-233.

Houseman, K. (2017, June 1). 10 Most Popular Physician Rating and Review Sites. Retrieved from GroupOne: http://www.grouponehealthsource.com/blog/10-most-popular-physician-rating-and-review-sites

Jeyapriya, A., & Selvi, C. K. (2015, February). Extracting aspects and mining opinions in product reviews using supervised learning algorithm. In Electronics and Communication Systems (ICECS), 2015 2nd International Conference on (pp. 548-552). IEEE.

Kou, G., Peng, Y., & Wang, G. (2014). Evaluation of clustering algorithms for financial risk analysis using MCDM methods. Information Sciences, 275, 1-12.

Loria, G. (2017). How Patients Use Online Reviews. Retrieved from software advice: https://www.softwareadvice.com/resources/how-patients-use-online-reviews/

McGlohon, M., Glance, N. S., & Reiter, Z. (2010, May). Star Quality: Aggregating Reviews to Rank Products and Merchants. In ICWSM.

McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, No. 1, pp. 41-48).

Mudambi, S. M., & Schuff, D. (2010). Research note: What makes a helpful online review? A study of customer reviews on Amazon. com. MIS quarterly, 185-200.

Penalver-Martinez, I., Garcia-Sanchez, F., Valencia-Garcia, R., Rodriguez-Garcia, M. A., Moreno, V., Fraga, A., & Sanchez-Cervantes, J. L. (2014). Feature-based opinion mining through ontologies. Expert Systems with Applications, 41(13), 5995-6008.

Raut, V. B., & Londhe, D. D. (2014, November). Opinion mining and summarization of hotel reviews. In Computational Intelligence and Communication Networks (CICN), 2014 International Conference on (pp. 556-559). IEEE.

Rathan, M., Hulipalled, V. R., Venugopal, K. R., & Patnaik, L. M. (2018). Consumer insight mining: aspect based Twitter opinion mining of mobile phone reviews. Applied Soft Computing, 68, 765-773.

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46). New York: IBM.

Rothenfluh, F., & Schulz, P. J. (2017). Physician rating websites: What aspects are important to identify a good doctor, and are patients capable of assessing them? A mixed-methods approach including physicians' and health care consumers' perspectives. Journal of medical Internet research, 19(5).

Rouse, M. (2010, 07). opinion mining (sentiment mining). Retrieved from searchbusinessanalytics.techtarget: https://searchbusinessanalytics.techtarget.com/definition/opinion-mining-sentiment-mining

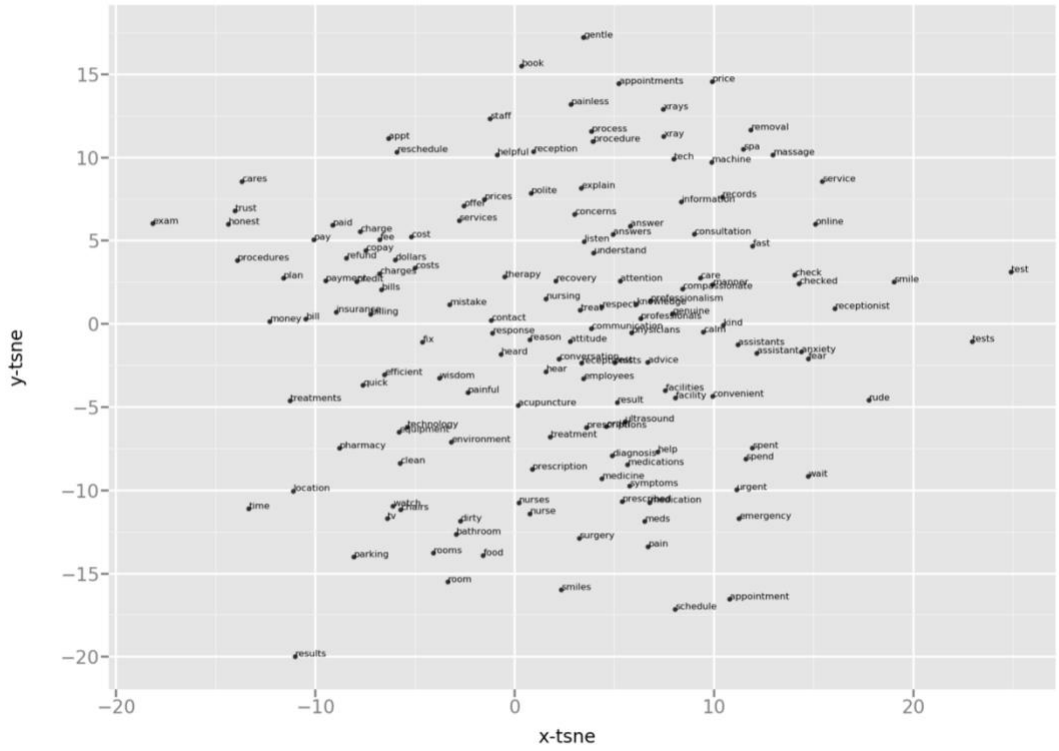
RxNLP. (2016, February 8). Why Use Stop Words For Text Mining? Retrieved from RxNLP: https://rxnlp.com/why-use-stop-words-for-text-mining/#.W-y4gnozZ24

Singh, V., & Dubey, S. K. (2014, September). Opinion mining and analysis: A literature review. In Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference- (pp. 232-239). IEEE.

Wolff-Mann, E. (2016, July 22). There's Everything Wrong with Online Reviews - and How to Fix Them. Retrieved from Money: http://time.com/money/page/online-reviews-trust-fix/

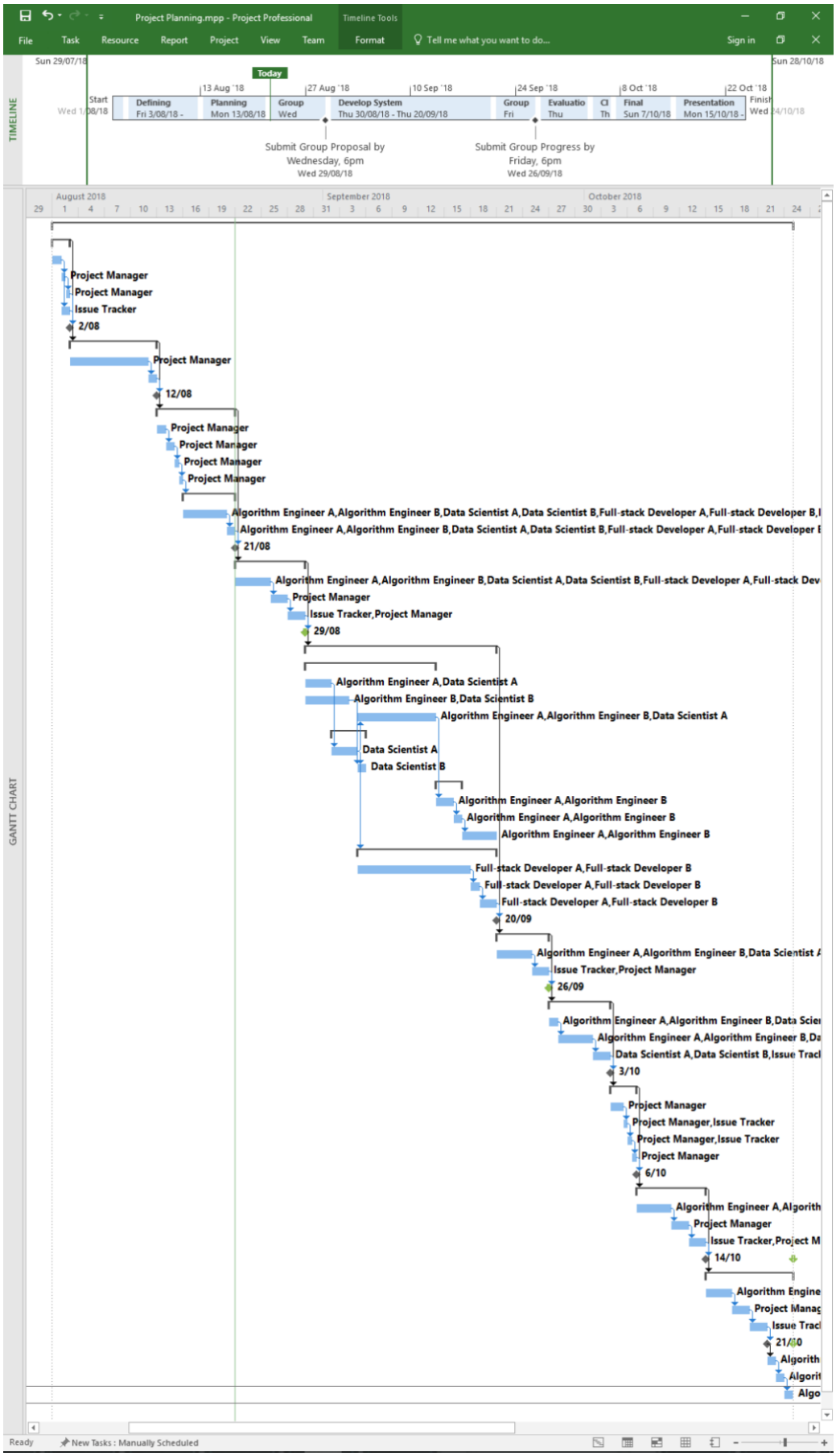
Yu, B., Zhou, J., Zhang, Y., & Cao, Y. (2017). Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews. arXiv preprint arXiv:1709.08698.

APPENDIX A – MAPPING OF WORD VECTORS



APPENDIX B – PROJECT WBS

	Task Mode	Task Name	Duration	Start	Finish	Predecessors	Resource Names
1		**Capstone Project**	**85 days**	**Wed 1/08/18**	**Wed 24/10/18**		
2		**Initiation**	**2 days**	**Wed 1/08/18**	**Thu 2/08/18**		
3		Group Forming	1 day	Wed 1/08/18	Wed 1/08/18		
4		Set up Slack workspace	0.5 days	Thu 2/08/18	Thu 2/08/18	3	Project Manager
5		Set up Github Organisation	0.5 days	Thu 2/08/18	Thu 2/08/18	4	Project Manager
6		Set up Google Drive	1 day	Thu 2/08/18	Thu 2/08/18	3	Issue Tracker
7		Initiation COMPLETE	0 days	Thu 2/08/18	Thu 2/08/18	5,6	
8		**Defining Project**	**10 days**	**Fri 3/08/18**	**Sun 12/08/18**	**2**	
9		Define Problem	9 days	Fri 3/08/18	Sat 11/08/18		Project Manager
10		Define Solution	1 day	Sun 12/08/18	Sun 12/08/18	9	
11		Defining Project COMPLETE	0 days	Sun 12/08/18	Sun 12/08/18	10	
12		**Planning Processes**	**9 days**	**Mon 13/08/18**	**Tue 21/08/18**	**8**	
13		Define Project Scope	1 day	Mon 13/08/18	Mon 13/08/18		Project Manager
14		Build WBS	1 day	Tue 14/08/18	Tue 14/08/18	13	Project Manager
15		Develop Resource Plans	0.5 days	Wed 15/08/18	Wed 15/08/18	14	Project Manager
16		Define Dependencies and Develop Project Schedule	0.5 days	Wed 15/08/18	Wed 15/08/18	15	Project Manager
17		**Develop Methodology**	**6 days**	**Thu 16/08/18**	**Tue 21/08/18**	**16**	
18		Article review & Case Study	5 days	Thu 16/08/18	Mon 20/08/18		Algorithm Engineer A,Algo
19		Define Tools, Techniques & Technologies	1 day	Tue 21/08/18	Tue 21/08/18	18	Algorithm Engineer A,Algo
20		Planning Processes COMPLETE	0 days	Tue 21/08/18	Tue 21/08/18	19	
21		**Group Proposal**	**8 days**	**Wed 22/08/18**	**Wed 29/08/18**	**12**	
22		Write Group Proposal	4 days	Wed 22/08/18	Sat 25/08/18		Algorithm Engineer A,Algo
23		Send Demo to Superviser & Tutor and Get Feedback	2 days	Sun 26/08/18	Mon 27/08/18	22	Project Manager
24		Modify Proposal	2 days	Tue 28/08/18	Wed 29/08/18	23	Issue Tracker,Project Mana
25		Submit Group Proposal by Wednesday, 6pm	0 days	Wed 29/08/18	Wed 29/08/18	24	Project Manager
26		**Develop System**	**22 days**	**Thu 30/08/18**	**Thu 20/09/18**	**21**	
27		**Develop Algorithm**	**15 days**	**Thu 30/08/18**	**Thu 13/09/18**		
28		Develop Data Collection Algorithm	3 days	Thu 30/08/18	Sat 1/09/18		Algorithm Engineer A,Data
29		Develop Data Filter Algorithm	5 days	Thu 30/08/18	Mon 3/09/18		Algorithm Engineer B,Data
30		Develop Review-based Clinic Rating Algorithm	9 days	Wed 5/09/18	Thu 13/09/18	32	Algorithm Engineer A,Algo
31		**Collect Data**	**4 days**	**Sun 2/09/18**	**Wed 5/09/18**		
32		Collect Data of Patient Reviews	3 days	Sun 2/09/18	Tue 4/09/18	28	Data Scientist A
33		Filter Cllected Data	1 day	Wed 5/09/18	Wed 5/09/18	29,32	Data Scientist B
34		**Execute Data**	**3 days**	**Fri 14/09/18**	**Sun 16/09/18**		
35		Train Clinic Rating Model	2 days	Fri 14/09/18	Sat 15/09/18	30	Algorithm Engineer A,Algo
36		Test Clinic Rating Model	1 day	Sun 16/09/18	Sun 16/09/18	35	Algorithm Engineer A,Algo
37		Evaluate Results & Optimise Algorithm	4 days	Mon 17/09/18	Thu 20/09/18	36	Algorithm Engineer A,Algo
38		**Web Application Development**	**16 days**	**Wed 5/09/18**	**Thu 20/09/18**	**32**	
39		Develop Web Application	13 days	Wed 5/09/18	Mon 17/09/18		Full-stack Developer A,Full
40		Test Web Application	1 day	Tue 18/09/18	Tue 18/09/18	39	Full-stack Developer A,Full
41		Deploy Web Application	2 days	Wed 19/09/18	Thu 20/09/18	40	Full-stack Developer A,Full
42		Web Application Development COMPLETE	0 days	Thu 20/09/18	Thu 20/09/18	41	
43		**Group Progress Report**	**6 days**	**Fri 21/09/18**	**Wed 26/09/18**	**26**	
44		Write Group Progress Report	4 days	Fri 21/09/18	Mon 24/09/18		Algorithm Engineer A,Algo
45		Modify Group Progress Report	2 days	Tue 25/09/18	Wed 26/09/18	44	Issue Tracker,Project Mana
46		Submit Group Progress by Friday, 6pm	0 days	Wed 26/09/18	Wed 26/09/18	45	Project Manager
47		**Evaluation**	**7 days**	**Thu 27/09/18**	**Wed 3/10/18**	**43**	
48		Plan Evaluation	1 day	Thu 27/09/18	Thu 27/09/18		Algorithm Engineer A,Algo
49		Implement Evaluation	4 days	Fri 28/09/18	Mon 1/10/18	48	Algorithm Engineer A,Algo
50		Analyse Feedback	2 days	Tue 2/10/18	Wed 3/10/18	49	Data Scientist A,Data Scien
51		Evaluation COMPLETE	0 days	Wed 3/10/18	Wed 3/10/18	50	
52		**Close Project**	**3 days**	**Thu 4/10/18**	**Sat 6/10/18**	**47**	
53		Review project performance	1.5 days	Thu 4/10/18	Fri 5/10/18		Project Manager
54		Check that the original objectives have been met	0.5 days	Fri 5/10/18	Fri 5/10/18	53	Project Manager,Issue Trac
55		Identify all unachieved objectives	0.5 days	Sat 6/10/18	Sat 6/10/18	54	Project Manager,Issue Trac
56		Identify furthur Study	0.5 days	Sat 6/10/18	Sat 6/10/18	55	Project Manager
57		Close Project COMPLETE	0 days	Sat 6/10/18	Sat 6/10/18	56	
58		**Final Report**	**8 days**	**Sun 7/10/18**	**Sun 14/10/18**	**52**	
59		Write Final Report	4 days	Sun 7/10/18	Wed 10/10/18		Algorithm Engineer A,Algo
60		Send Final Report to Supervisor & Tutor and Get Feedback	2 days	Thu 11/10/18	Fri 12/10/18	59	Project Manager
61		Modify Final Report	2 days	Sat 13/10/18	Sun 14/10/18	60	Issue Tracker,Project Mana
62		Submit Final Report by Friday, 6pm	0 days	Sun 14/10/18	Sun 14/10/18	61	Project Manager
63		**Presentation**	**10 days**	**Mon 15/10/18**	**Wed 24/10/18**	**58**	
64		Write Presentation Slides	3 days	Mon 15/10/18	Wed 17/10/18		Algorithm Engineer A,Algo
65		Send Slide Draft to Supervisor & Tutor and Get Feedback	2 days	Thu 18/10/18	Fri 19/10/18	64	Project Manager
66		Modify Slides	2 days	Sat 20/10/18	Sun 21/10/18	65	Issue Tracker,Project Mana
67		Submit Presentation Slides	0 days	Sun 21/10/18	Sun 21/10/18	66	Project Manager
68		First Rehersal	1 day	Mon 22/10/18	Mon 22/10/18	67	Algorithm Engineer A,Algo
69		Second Rehersal	1 day	Tue 23/10/18	Tue 23/10/18	68	Algorithm Engineer A,Algo
70		Third Rehersal	1 day	Wed 24/10/18	Wed 24/10/18	69	Algorithm Engineer A,Algo


Project Planning.mpp - Project Professional
Timeline Tools
File
Task
Resource
Report
Project
View
Team
Format
Tell me what you want to do...
Sign in
TIMELINE
Today
Defining
Planning
Group
Develop System
Group
Evaluatio
Final
Presentation
Submit Group Proposal by
Wednesday, 6pm
Wed 29/08/18
Submit Group Progress by
Friday, 6pm
Wed 26/09/18
August 2018
September 2018
October 2018
GANTT CHART
Project Manager
Issue Tracker
2/08
12/08
Algorithm Engineer A,Algorithm Engineer B,Data Scientist A,Data Scientist B,Full-stack Developer A,Full-stack Developer B
21/08
Issue Tracker,Project Manager
29/08
Algorithm Engineer A,Data Scientist A
Algorithm Engineer B,Data Scientist B
Data Scientist A
Data Scientist B
Full-stack Developer A,Full-stack Developer B
20/09
26/09
3/10
Project Manager,Issue Tracker
6/10
14/10
Ready
New Tasks : Manually Scheduled