

# **SENTIMENT ANALYSIS ON PRODUCT TWEETS**

By

Ira Iryani bt. Mohd Ariff

15187

Dissertation submitted in partial fulfilment of

the requirements for the

Bachelor of Technology (Hons)

(Business Information System)

SEPTEMBER 2013

Universiti Teknologi PETRONAS

Bandar Seri Iskandar

31750 Tronoh

Perak Darul Ridzuan

# **CERTIFICATION OF APPROVAL**

**Sentiment Analysis on Product Tweets**

by

Ira Iryani bt Mohd Ariff

A dissertation submitted to the  
Business Information System Programme  
Universiti Teknologi PETRONAS  
in partial fulfilment of the requirement for the  
BACHELOR OF Technology (Hons)  
(Business Information System)

Approved by,

---

(Dr Yong Suet Peng @ Vivian)

UNIVERSITI TEKNOLOGI PETRONAS

TRONOH, PERAK

September 2013

## **CERTIFICATION OF ORIGINALITY**

This is to certify that I am responsible for the work submitted in this project, that the original work is my own except as specified in the references and acknowledgements, and that the original work contained herein has not been undertaken or done by unspecified sources or persons.

---

**IRA IRYANI MOHD ARIFF**

## CONTENTS

CONTENTS .....	4
LIST OF TABLES .....	6
LIST OF FIGURES .....	6
ABSTRACT .....	7
CHAPTER 1 .....	8
INTRODUCTION .....	8
1.1 Background .....	8
1.2 Problem Statement .....	11
1.3 Objectives .....	13
1.4 Scope of Study.....	14
CHAPTER 2 .....	16
Literature Review .....	16
CHAPTER 3 .....	21
Methodology .....	21
3.1 Research Method .....	21
3.2 Data Collection .....	23
3.3 Gantt-Chart .....	24
3.4 Key Milestones .....	25
3.5 Design and Implementation.....	26
3.5.1 Requirement Analysis.....	26
3.5.2 System Architecture.....	27
3.5.3 System Development .....	29
3.5.4 System Operations .....	32
CHAPTER 4 .....	38
Result and Discussion .....	38
4.1 Case Base Pattern Testing: .....	38
4.2 Human Evaluation Testing .....	40
CHAPTER 5 .....	43

Conclusion .....	43
5.1 Limitation of the Project.....	43
5.2 Recommendation.....	44
REFERENCES.....	45
APPENDIX.....	50

## LIST OF TABLES

Table 1: Informal Language.....	12
Table 2 : Emoticons .....	12
Table 3 : Short-form.....	13
Table 4 : Gantt-Chart .....	24
Table 5 : Key Milestones .....	25
Table 6 : Positive and Negative Lexicons.....	32
Table 7 : Feature Lexicon .....	33
Table 8 : Positive Pattern Case Base.....	35
Table 9 : Negative Pattern Case Base .....	36

## LIST OF FIGURES

Figure 1 : People Relationship .....	10
Figure 2 : Compose New Tweet.....	14
Figure 3 : Tweets.....	15
Figure 4 : Sanders Analytics .....	20
Figure 5 : Prototyping Model.....	21
Figure 6 : System Requirements .....	26
Figure 7 : Flow for System Architecture.....	28
Figure 8 : System Interface .....	29
Figure 9 : Result Based on Pattern.....	38
Figure 10 : Result Based on Frequent Words .....	39
Figure 11 : Human Evaluation Testing (Positive outcome).....	42

## **ABSTRACT**

With the advancement of Web 2.0, the user could do more than just retrieve information from a static website. Better user-interface, software and storage facilities all in one place which is called the web browser. One of the main features of this Web 2.0 includes the social media. The hype of social media such as Twitter and Facebook have made people express their opinions and feelings more easily publicly. Everyone interprets the information they got differently. They have their own understanding and interpretation on how the information is. With the technology that is rapidly growing, we can use the information that the user is displaying on social media and make this as opportunity thus identifying the problems as soon as it occurs. Sentiment analysis is about finding subjective information and grouped it into polarity classification (positive, negative or neutral). One of the objectives of this project is, to automatically categorize data into either positive sentiment, negative sentiment or neutral sentiment based on the subjective data that is obtained from the social media. This system can be useful for companies who are interested to get the fastest way to obtain juicy and latest information from the social network. Another target user could be the institutions that are reputation conscious. Case-Base Reasoning (CBR) will be used in this project. CBR is done by looking at past situations to solve the possible same current issue. Large amount of data is hard to comprehend thus, machine learning techniques could be used to automate the tasks and also provide the predictions over that matter.

# CHAPTER 1

## INTRODUCTION

### 1.1 Background

The booming of mobile Internet and social media are growing fast in the technology industry worldwide and it is predicted that mobile Internet will also increase. (Meeker, 2013).

With the availability of accessing the Internet at our fingertips makes everything easier and faster. Stories are shared, feelings are being expressed, everything in social media.

Artificial Intelligence (AI) is a technology when machines such as computers are made able to make intelligent decisions as humans.

- As stated in Oxford Dictionaries (2013), artificial is defined as made or produced by human rather than naturally occurrence.
- Intelligence is stated as the ability to think and understand instead of doing things by instinct or automatically.

AI technology has been into discussion since the late 1940's. After the research done by Alan Turing on 'Computing Machinery and Intelligence', his approach has become universal. Turing defined the intelligence behaviour of a computer as the skill to achieve human-level performance is an intellectual task. (Negnevitsky, 2005).

Sentiment analysis or can be known as opinion mining is an important factor nowadays. Interpreting one's simple statement could lead to resolving issues and preventing it from occur. Humans may be interpreting one statement differently. This is mainly because we are thinking differently, our intelligence gives a huge impact on what we are saying and experiences that we have are varies from one another.



The aim for sentiment analysis is to conclude the attitude of the writer of the statement. The writer attitude could be based from these 3 factors:

*i. Appraisal Theory*

This is to say that what the writer went through and how he perceived the situation.

*ii. Affective State*

The emotional state when the writer expressed his feelings.

*iii. Intended Emotional Communication*

This has the meaning of what the writer wants other readers to feel too.

There can be either one the ternary scales. The result will be called as polarity which can be either positive sentiment, negative sentiment or neutral sentiment.

*i. Positive Sentiment*

Texts that will be grouped into this section will be positive texts with positive words such as “like” and “love”.

*ii. Negative Sentiment*

Texts that will be grouped into this section will be negative texts with direct negative words such as “hate” and “loathe”.

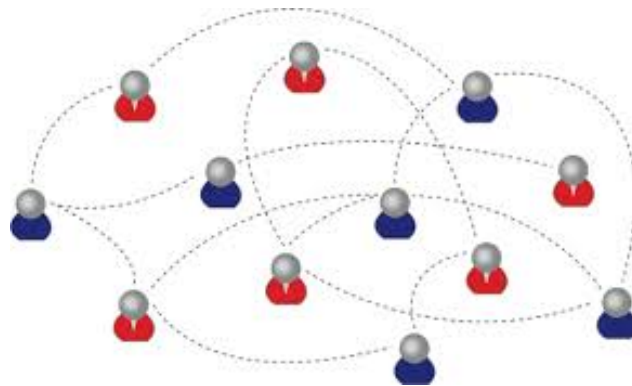
*iii. Neutral Sentiment*

Informative texts that do not fall neither in positive or negative sentiment will be grouped in neutral sentiment. This is to make easier to distinguish between those 2 sentiments above.

Text based micro blogging such as Twitter has become a challenge for the user as it only allow 140 characters to be entered. Short texts such as Tweets will be extracted and analysed in this project.

There are some challenges in this project. Those are tweets that contain informal language, emoticons and short-forms.

This project will be focusing on one particular object. Iphone is chosen to be the object in this research. Every phone has different features. For a quick example, Iphone has features of larger display of 4 inch display, 18% lighter and an 8MP iSight camera. All of these features will be taken into consideration to indicate whether the writer is producing a positive, negative or neutral sentiment.



**Figure 1 : People Relationship**

Figure 3 is showing the people relationship on the Internet. For an example, if a person is spreading negative comment over one particular product, it will spread to his friends and his friends will tell others. Reputable and huge companies and also institutions would not want this kind of bad comments to spread. Reputation will be corrupted and thus it will bring down their image.

With that being said, the target user for this system can be useful for companies who are interested to get the fastest way to obtain juicy information from the social media. Another target user could be the institutions that are reputation conscious. These entities will need the fastest way to interpret what people are writing on the social media. By doing so, they could acquire what the people are saying about their products.

## **1.2 Problem Statement**

The problem statement of this project is how to determine positive, negative or neutral Tweets by using Case Base method?

Previous work in the research titled A Case-Based Approach to Cross Domain done by Bruno Ohana, Sarah Jane Delany, and Brendan Tierney (2012), they focused on many domains. They have looked into various case base scopes which include books, electronics, film, music, hotels and apparel. Furthermore, in this research the case base population is done on huge amount of documents which differs from Sentiment Analysis on Product Tweets.

On the other hand, for Sentiment Analysis on Product Tweets, Case Base method is adopted and Iphone is to be the focus object on this project.

There are some difficulties encountered in this project. The usage of improper English and slangs that are used on social network such as Twitter has made everyday decisions tougher.

### **(i) Informal Language**

Informal language is described as unplanned speech in situations that may be described as natural or "real-life". This can be seen as the usage of the words as shown below:

<b>Formal</b>	<b>Informal</b>
We cordially invite you to the Year 12 formal.	Hey buddy! <i>Wanna</i> go to the dance?

**Table 1: Informal Language**

**(ii) Emoticons**

Emoticons were normally used in the daily communication. For example, it can be used on the social media. By using emoticons or known as smiley, the writer is including his mood towards the written sentence. Below is a table showing the translation of the emoticons.

<b>Emoticons</b>	<b>Meaning</b>
: -)	Happy
: -(	Sad
: -')	Tears of happiness

**Table 2 : Emoticons**

**(iii) Short-form**

Short-form is widely used even with short message service (SMS). The usage of short-form will be used more frequently on Twitter as to help to minimize the characters used. This is because Twitter has limited characters. Below is shown examples of short-form.

<b>Short-form</b>	<b>Meaning</b>
Tba	To be announced
Fyi	For your information
Tbc	To be continued

**Table 3 : Short-form**

### **1.3 Objectives**

Objectives for this project can be stated as below.

- To automatically categorize data into either positive sentiment, negative sentiment or neutral sentiment based on the subjective data that is obtained from Twitter.
- To build a prototype of sentiment analysis system that will be able to generate conclusion.
- To generate a statistical report over processed tweets.

## 1.4 Scope of Study

In Sentiment Analysis on Product Tweets, the most leading social networking which is Twitter will be the platform to gather information. Twitter is a text micro blogging which allows the user to post text messages.

These text messages are known as tweets. Below is an example (taken from desktop view) of tweet box where tweets can be written.



**Figure 2 : Compose New Tweet**

These tweets can be sent through their mobile phones or computers and could be written up to 140 characters.

Tweets will be used as the dataset for this system testing. The tweets that will be taken are all made public (not private tweets).

This project will focus tweets on Iphone. For example, any tweet that has the word “iphone” will be taken. Research papers on sentiment analysis on product reviews, comments and tweets that will be used in this project will be cited.

Below is the example of tweets that are taken from Twitter.



**Figure 3 : Tweets**

As Figure 2 above has shown, all the tweets contain the “#” symbol. This symbol is called a “hashtag”. This is used to mark keywords or topics in a Tweet. It is used to categorize messages. For an example, #iphone will have all the tweets relating to Iphone. It was created organically by Twitter users. (Twitter, 2013).

Social media is spreading like wildfire. Every company would want to spread their products and achievements around the world with the easiest way. With many people connecting with each other in the social network, it makes some companies and institutions feel the need to get involve too. By going through social network like Twitter, they could capture larger market. Word of mouth is the key to either successful business or otherwise.

## **CHAPTER 2**

### **LITERATURE REVIEW**

Twitter provides real-time information that could be valuable to certain people. 140-character messages are called Tweets. With the existence of Twitter, one could discover the latest news with the easiest way possible. (Twitter, 2013).

Traditional ways such as collecting surveys incur high cost and people are not interested on answering the questionnaires. Sentiment analysis is important as it will detect early warning system of possible disruption in a timely manner by detecting early feedbacks from the citizens and the unknown problems. (Osimo and Mureddu, 2010). In Research Challenge on Opinion Mining and Sentiment Analysis stated that opinion mining is helpful to identify by listening rather than by asking as this will produce rather more accurate information.

A research paper written by Prabowo and Thelwall (2009) said that, with the advancement of technology, sentiment analysis could help companies to estimate the acceptance of its product. Through that way, they could come up with strategies and improve their quality. This could also be used to facilitate the policy makers or politicians by analysing public sentiments without breaching the policies, public services and political issues. Taking advantage over the current technology such the booming of social media could help us to interpret data even better.

Case-Base Reasoning (CBR) is one of the methods available to implement Sentiment Analysis. CBR is known by recalling the past successfully solved problems and use the same solutions to solve the current closely related problems. (Pantic, n.d.) Some of the advantages of using CBR are mentioned below:

- CBR does not require an explicit domain model and so elicitation becomes a task of gathering case histories.



- Implementation is reduced to identify significant features that describe a case, an easier task than creating an explicit model.
- CBR systems can learn by acquiring new knowledge as cases. This and the application of database techniques make the maintenance of large volumes of information easier.

To conclude CBR, it can be described in 4 simple steps. Those are:

- 1) Retrieve – retrieve familiar cases against the current cases.
- 2) Reuse – adapt the previous cases to map to new case.
- 3) Revise – if the solution cannot be found in the case base, proposed a new solution.
- 4) Retain - if the new solution can be used, include the solution in the case base.

Research done by Ohana, Delany and Tierney (2012) said that when the domains used for training and evaluation have little in common, poor result will be obtained. By using out-of domain data to build classifier ensembles and also extending training data with in-domain unlabelled documents could overcome the poor result. They select only term that is tagged as adjectives and verbs during the document scoring. They add that by using tagging pre-processing step could improve the accuracy of lexicon queries.

The sentiment lexicons are normally used in unsupervised approaches with an algorithm that scans the document and extract the sentiment score based on the clues.

For an example, “I like you” and “I do not like you” both have positive term “like”. NegEx algorithm is used in research titled A Case-Based Approach to Cross Domain Sentiment Classification (Ohana et al., 2012) to identify opinion in negated sentences. They scan the document to negate n-grams and inverting sentiment orientation of nearby terms in the same sentence. Their result is that extension of

more lexicons during the case base population stage, and this can help reducing the discard ratio by recording more cases for later reuse.

One research paper stated that the result by using unsupervised method outperforms machine learning solutions in majority cases. It is a reliable solution for sentiment analysis of informal communication on the Web. The advantage of the approach used is that it requires no training and can be applied into wide selection of environment. (Paltoglou and Thelwall, 2012).

Now in 2013, the user can do more than just retrieving data from the web such as writing their opinions and expressing their thoughts on the web. This could open up the opportunity for the institutions and companies to take advantage of what is written online and make use of it.

Many researches focused on product reviews. With that little information on the web, institution and companies could predict whether a reviewer recommends their product or otherwise. (Dave et al. 2003; Turney 2002).

“Naïve Bayes polarity classifier, the subjectivity extracts are shown to be more effective input than the originating document, which suggests that they are not only shorter, but also “cleaner” representations of the intended polarity.” (Pang and Lee, 2004)

Existing supervised learning methods can be readily applied to sentiment classification such as Naïve Bayesian, and Support Vector Machines (SVM). Further important enhancement is on identification of the features indicating whether sentences are on-topic. (Pang et al., 2002). In this research, he took this approach to classify movie reviews into two classes, positive and negative. In the result they have gotten, it stated that Naïve Bayes tends to do the worst and SVMs tend to do the best. The accuracy of sentiment classification is not achieved. It was shown that using unigrams as features in classification performed well with either naïve Bayesian or SVM.

Models such as Tree Kernel and Feature Based Models are used to examine the Sentiment Analysis of Twitter Data which outperformed the unigram or known as n-grams baseline. They conclude that the Twitter sentiment analysis is not that different from sentiment analysis for other genres. (Agarwal et al., 2011)

Artificial Neural Network (ANN) is a mathematical model that interconnects group of artificial neurons. It will process information using a connectionist approach to computation. ANN is used in to find relationship between input and output or to find patterns in data. In research done by (Sharma, A. & Dey, S., 2012) mentioned large dataset application such as product reviews will be suitable with approach that they adopted.

Twitter semantic sentiment analysis done by (Saif et al., 2012) stated that Alchemy API is used due to its better performance in terms of coverage and accuracy. However there is one disadvantage of using this method which is the abstraction level of the concepts retrieved from the entity extractor. For an example, "People" is used equally to describe famous musicians or politicians. In this case, it is too abstract. Given a tweet mentioning "I wish I could go to France to meet President Obama haha" and Alchemy API provide the concept "Person" to represent President Obama while Zemanta identified Obama as the concept of government or politician which is specific. Zemanta produced more accurate and specific concepts to describe entities related to music tracks and bands. They also concluded that results will be more accurate if the datasets used to be analysed are larger and cover wide range of topics.

According to Yie and Lee, there are 4 methods involved in Sentiment Analysis Platform (SAP). Those are text pre-processing, feature classification, polarity classification and summarizing and decision making. Those steps will be adopted in this project.

## *Sanders Analytics*

**Figure 4 : Sanders Analytics**

Tweet Sanders Corpus (Sanders, 2011) dataset will be used for training in this project. In that dataset, it has 5513 hand-classified tweets. These tweets were classified to one of 4 different topics which can be either Apple, Google, Microsoft or Twitter. Each entry contains:

- Tweet id
- Tweet text
- Tweet creation date
- Topic used for sentiment
- Sentiment label: positive, neutral, negative or irrelevant.

Iphone falls under “Apple” topic in the Sanders Analytics dataset. Thus, for this project, Apple product topic will be used.

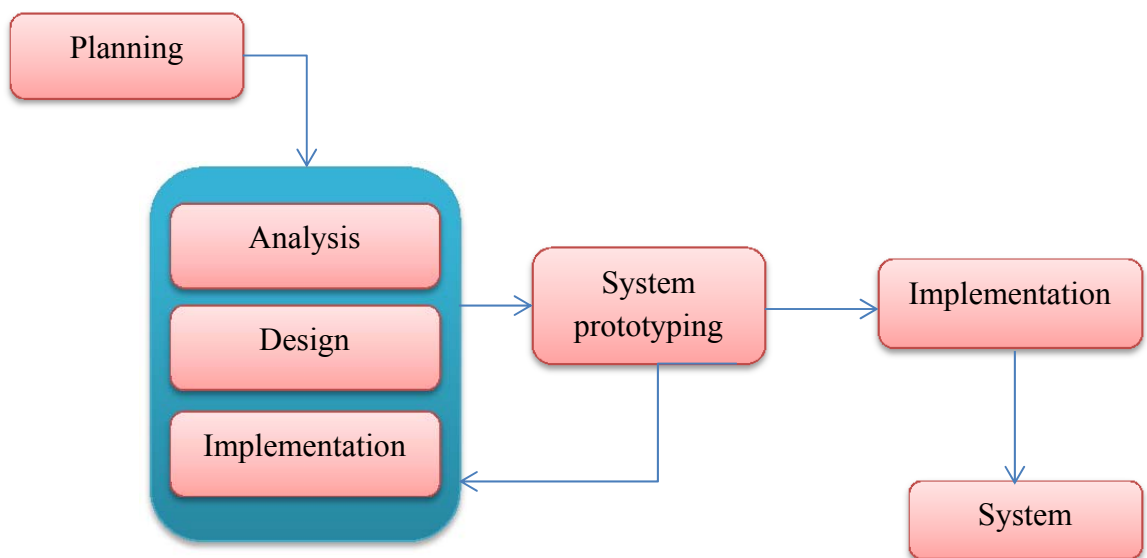
## CHAPTER 3

### METHODOLOGY

The objective of this project is to come up with a system that will be able to automatically generate accurate result based on the Tweets that will be processed by the system. CBR method will be used for this system.

#### 3.1 Research Method

In this project Rapid Application Development (RAD) will be used. A prototyping-based methodology that has been used included analysis, design and implementation that will be performed repeatedly until the system is complete.



**Figure 5 : Prototyping Model**

Prototyping model is chosen because of the development of the system is within 28 weeks including Final Year Project 1 and Final Year Project 2. By using this model, it is fast and it is easier to test the system functionality.

## **Planning**

In the planning stage, basic research is done. Research on past research papers, sample online programmes and all materials that are related to the project are collected. Research papers that are made available online were used to strengthen this project.

## **Analysis**

Analysis over the collected information during the planning stage is done. Deeper research has been done. This is to ensure that the project will have concrete base. Methods that were found in research papers done by previous works have been taken into consideration before deciding on the chosen method.

## **Design**

Prototype of the system will be done in this phase. The first draft of the design is done. This is to allow continuous development of the system. In this stage, enhancement of the prototype will be done continuously. This is because to ensure that the system will be able to generate a system that could generate conclusion.

## **Implementation**

The actual project will be built in this phase. In this stage, it will involve dealing with programming and testing. The codes will be done on Microsoft Visual Basic 2010 Express. Testing of the system will be done once the system is done. It will be tested to ensure that the system is functioning well. Continuous improvement will be done over the prototype.

### **3.2 Data Collection**

Below are the sources gathered for data collection to implement this project.

1) Twitter

Tweets collected will be used for testing for human evaluation.

2) Sanders Analytics

Sanders Analytics on topic regarding “Apple” will be used in this project for testing.

3) Positive Lexicon

Positive lexicon is built by taking positive words from Winspiration.

4) Negative Lexicon

Negative lexicon is built by taking negative words from Enchanted Learning.

### 3.3 Gantt-Chart

	FYP 1														FYP 2														
Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
Planning Phase																													
Problem Identification	■	■																											
Background Study		■	■																										
Project Approval				■																									
Literature Review				■	■	■																							
Analysis Phase																													
Submission of Extended Proposal						■																							
Research Work				■	■	■	■	■	■	■																			
Design Phase																													
Interface												■																	
Proposal Defence													■																
Design of System Architecture														■															
Submission of Interim Report														■															
Implementation Phase																													
System Development															■	■	■	■											
System Implementation																			■										
Usability Testing																				■	■								
Progress Report Submission																						■							
Improvement of prototype																							■	■	■				
Pre-Sedex																								■					
Viva																										■			
Project Dissertation Submission																											■	■	

Table 4 : Gantt-Chart



The project Gantt Chart covers both Final Year Project 1 (FYP1) and Final Year Project 2 (FYP2). The duration for both of these FYP is 28 weeks. In the FYP1, planning, analysis and design phases were done. Currently, in FYP2, development, implementation and testing phase will be conducted according to the planned timeframe.

### 3.4 Key Milestones

No	Deliverables / Activities	Week
1	Title Selection and Proposal	2
2	Project Approval	4
3	Problem Identification	5
4	Extended Proposal	6
5	Interface Design	11
6	Proposal Defense	12
7	Interim Report	14
8	System Architecture	13-14
9	Progress Report	20
10	Usability Testing	20-21
11	Pre-Sedex	24
12	Viva	27
13	Final Dissertation	28

**Table 5 : Key Milestones**



Activities



Deliverables

As stated in the Table 5, the key milestones are the important deliverables that need to be executed and presented to the University together with the activities that have been planned throughout these 28 weeks. Currently, deliverables and activities of FYP2 should be achieved in order to complete this project successfully.

### 3.5 Design and Implementation

Microsoft Visual Basic 2010 is used to develop the system.

#### 3.5.1 Requirement Analysis

This system mainly will focus on identifying whether the tweet falls under positive sentiment, negative sentiment or neutral sentiment. Tweets dataset for the training and testing purposes will be needed. The dataset will be narrowed down to 100 tweets relating to the subject.

Below are the system requirements that will be needed to come up and implement the system successfully.

REQUIREMENTS	TOOLS
<b>HARDWARE</b>	
Any desktop or any personal computer with standard hardware	
<b>SOFTWARE</b>	
Operating System	Microsoft Windows 7
Application Development Tools	Microsoft Visual Basic 2010 Express
Database	Notepad / Microsoft Excel 2010 / Microsoft Access 2010
Application for Documentation	Microsoft Word 2010

**Figure 6 : System Requirements**

Microsoft Windows will be the major platform of this project. Implementation over other platform may need some amendments.

This system is focused on product tweets that are made unprotected in Twitter. It will only be able to detect simple sentences. This project will be focusing on one sentence with one feature. For an example, “iPhone sound system is the best”. In this sentence, it is obvious that the feature stated is the sound system with positive sentiment.

### **3.5.2 System Architecture**

In conducting the process of finding the sentiment data, there will involve 4 processes which are text pre-processing, feature classification, polarity classification and summarizing and decision making. According to Chen and Lee (2011) these steps are the main method that will need to be done.

Methods used by Chen and Lee (2011) will be adopted in this system. Text pre-processing is eliminating unwanted words, correcting misspelled words and finding important words in a sentence. Eliminating unwanted words are such as stop words and emoticons.

Feature identification is done by detecting a feature in one particular sentence. For an example, from the sentence “The screen of this television is big”. From this sentence, it is obvious that screen is the feature of the television.

Sentiment identification is a process to determine the expressed text is positive, negative or neutral. Words such as “not” and “no” will be taken in this step.

The last step is result and summarizing when all of the steps above completed. In this step, the system will be able to generate output in graphical formats. From this summary, the reader will know the product is positively accepted or otherwise and which features are lacking in performance. Illustration is as followed.

## Tweets on Product Features

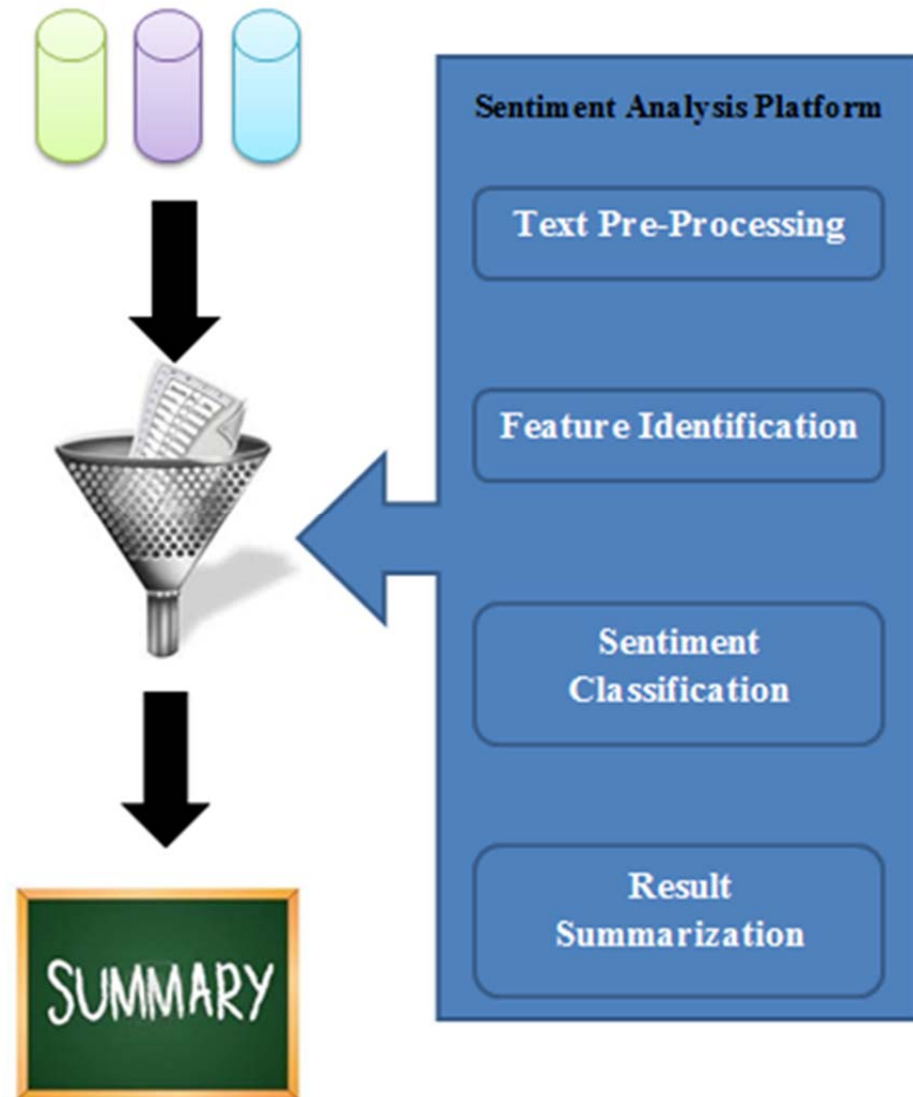
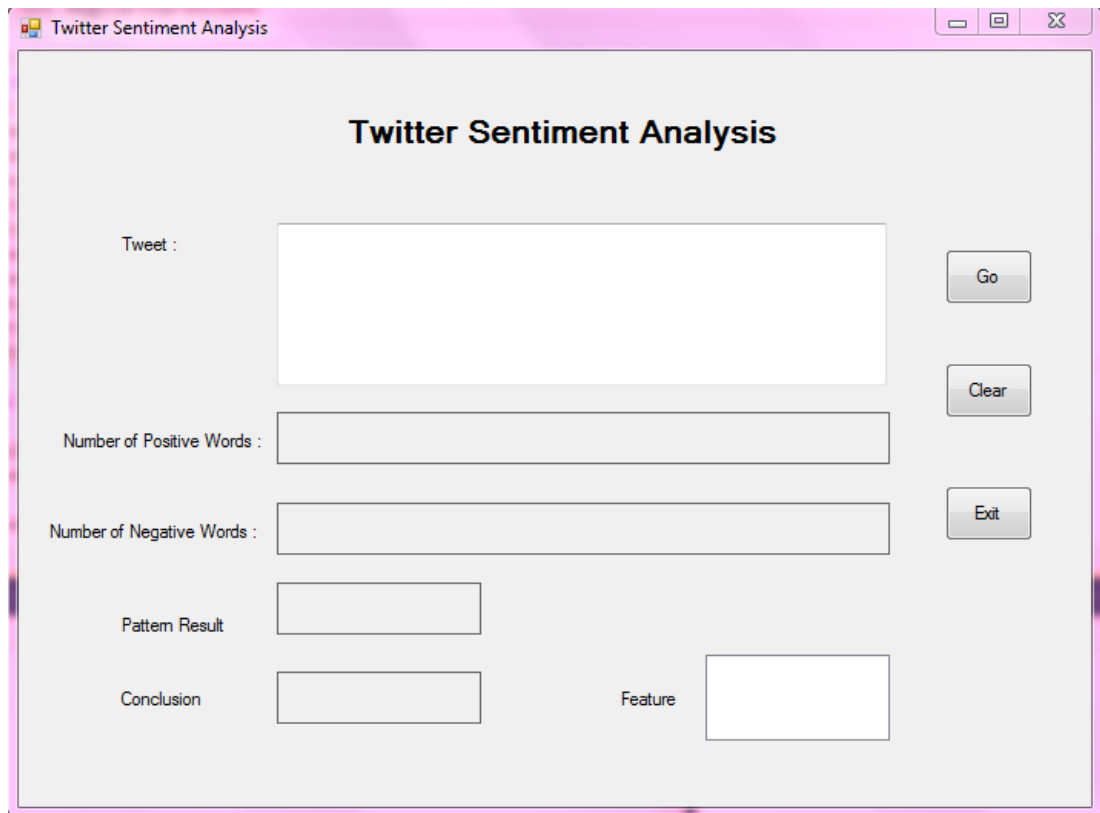


Figure 7 : Flow for System Architecture

### 3.5.3 System Development

The system is developed using Visual Basic 2010. In the current development of the system, there are 377 positive words in the Positive Lexicons, 229 negative words in the Negative Lexicons.



The screenshot shows a Windows application window titled "Twitter Sentiment Analysis". The window has a light gray background and a title bar with standard Windows window controls (minimize, maximize, close). The main content area is titled "Twitter Sentiment Analysis" in bold black text. Below the title, there is a "Tweet :" label followed by a large white text input box. To the right of this box are three buttons: "Go", "Clear", and "Exit", arranged vertically. Below the "Tweet" box, there are two labels: "Number of Positive Words :" and "Number of Negative Words :", each followed by a white text input box. At the bottom of the interface, there are four labels: "Pattern Result", "Conclusion", and "Feature", each followed by a white text input box. The "Feature" label is positioned to the left of its corresponding input box, while the others are to the right.

**Figure 8 : System Interface**

Tweet will be copied and pasted in the Tweet textbox. The textbox has limit of character that can be accessed by the system. It can only contain not more than 140 characters. The textbox is strictly for text based messages. Others such as pictures and images regardless of the various formats will not be processed.

Under the Positive Result section, the system will produce statistical result. By this it means that the system can indicate how many positive words found in a particular tweet. The positive words will be extracted from Positive Lexicons that the has been compiled in Notepad.

For the Negative Result section, the system will produce another statistical result. It will prompt the user with numbers of negative words found in the tweet. Negative words will be extracted from the Negative Lexicons that is linked to the system.

After the completion of processing, the conclusion section will display the actual result of the tweet. This means that, the system will be able to automatically generate the sentiment for a particular tweet. The system will conclude whether the tweet is positive, negative or neutral sentiment.

Overview of function of each elements:

**1. Program Load**

In this section, the positive and negative lexicons are opened. The lexicons were saved in Notepad with the extension “.txt”. It will be connected to the location where the files are stored. The lexicons will be stored in an array respectively. The lexicons will be split by spaces.

**2. Textbox**

In this system, there is only one textbox. The textbox will receive input from the user. This textbox will be limited to 140 characters which follows the Twitter format.

### **3. “Go” Button**

The button “Go” will launch the system. Any tweets that contain other than alphabets will be replaced with space. This is to ensure that no emoticons are considered. For an example, if a tweet says, “I do not like Apple brand ☹”.

With the emoticons at the end of the tweet indicate the level of dissatisfaction. The writer must be really disappointed with the Apple brand that he has to strengthen his dissatisfaction by adding the sad face at the end of the tweet. This part of the coding calculates the positive and negative words. This part is essential because the numbers of positive and negative words will influence the conclusion result.

### **4. “Clear” Button**

The clear button will reset the interface. The textbox will be emptied, the labels to calculate the statistical result will be zero.

### **5. “Exit” Button**

The exit button will exit the execution and the program will stop loading.

### 3.5.4 System Operations

The system works with the algorithm that is set in it. The algorithm for the system of Sentiment analysis on Product Tweets will be adopted by a research done by Ohana et al. (2012). Their research paper is using CBR method on large amount of text documents. This varies from Sentiment analysis on Product Tweets. This is because of the limitation on the characters being used. For this system, it limits the character to 140 characters while for Ohana et al. (2012), they were using texts of documents.

#### 5) Databases

Lexicon is known as a catalogue or a dictionary of a given language's words. These lexicons act as databases to retrieve words and to compare it with the Tweet input. There are 5 lexicons used in this project.

Below are the lexicons that have been used in this project:

##### 1) Positive and negative lexicons.

good	bad
wonderful	disaster
fantastic	terrible
love	poor
excited	loath
amazing	hate

**Table 6 : Positive and Negative Lexicons**

These lexicons will contain a set of words that represents positive and negative. The positive lexicons were taken from positive words that were



published in Winspiration (2006) while negative lexicons were taken from negative words that were published in Enchanted Learning (2012). Both positive and negative words were compiled in a Notepad separated by “Enter”. The lexicons were arranged alphabetically. These lexicons will act as databases and will be connected to Microsoft Visual Basic 2010.

## 2) Feature Lexicon

This feature lexicon gives the word into meaningful sentences and it will be appear at the list box in the system.

Screen
Camera
Wi-Fi Direct
Technology
Memory
Weight

**Table 7 : Feature Lexicon**

Under this dictionary, the features of the product will be focused. In Table 7 is the example of feature lexicons. All the features will be included in this lexicon. There are currently 15 features being used in this project. Feature lexicon will be stored in Notepad separated by “Enter”. It will be linked to Microsoft Visual Basic 2010. This feature lexicon acts as a database which only stores features of a gadget.

### 3) Case Base Lexicons

All the possible patterns to generate outcome are saved under Case Base Lexicons. In this project, both positive and negative patterns are separated into two different lexicons. However, for the initial testing, the system will only be able to read up to either 6 positive words or 6 negative words.

#### 3.1) Positive Case Base Lexicon

There are 48 possible patterns in this positive case base lexicon. These patterns are saved in Notepad. It is separated by "Enter". Positive is labelled as 1. The format in of this lexicon is in binary. Result will be generated by the pattern stored in this lexicon. This case base lexicon will be extracted into the system and compare with the current event Tweet. If the pattern could not be found in the lexicon, the system will not able to generate outcome based on the pattern. However, the system is intelligent and will generate outcome based on the frequency of the word.

```

1
11
111
110
101
011
1111
1110
1101
1011
0111
11111
10111
11011
11101
11110
01111
01110
11100
11001
10011
00111
10101
01011
01101
111111
011111
101111
110111
111011
111101
111110
011110
010111
011011
011101
001111
100111
110011
111001
111100

```

**Table 8 : Positive Pattern Case Base**

### 3.2) Negative Case Base Lexicon

There are 48 possible patterns in negative case base lexicon. These patterns are saved in Notepad. It is separated by “Enter”. Negative is labelled as 0. The format in of this lexicon is in binary. Result will be generated by the pattern stored in this lexicon. This case base lexicon will be extracted into the system and compare with the current event pasted in the textbox provided. If the pattern could not be found in the lexicon, the system will not able to generate outcome based on the pattern. However, the system is able to generate outcome based on the frequency of the word.

0
00
000
001
010
100
0000
1000
0100
0010
0001
00000
10000
01000
00100
00010
00001
11000
10100
10010
10001
00011
00110
01100
11000
000000
100000
010000
001000
000100
000010
000001
000011
000110
001100
011000
110000
101000
100100
100010
100001

**Table 9 : Negative Pattern Case Base**

### **By using the Case-Based Approach:**

Below is the Algorithm that is used in research done by (Ohana et al., 2012).

#### **Populate the Case Base**

##### **Input:**

- T, set of labelled out-of-domain tweets for training.
- L, set of all available sentiment lexicons.
- $f(L,t)$ , unsupervised tweet sentiment classifier using lexicon L as input.

##### **Output:**

- CB, the populated case base.

CB  $\leftarrow$  {}

for all tweets  $t$  in T do

S  $\leftarrow$  {}

for all  $L_i$  in L do

make prediction using  $f(L_i, t)$

if prediction is correct then

S  $\leftarrow$  S  $\cup$   $L_i$

end if

end for

if S  $\neq$  {} then

compute case description  $x(t)$

CB  $\leftarrow$  CB  $\cup$  ( $x(t), S$ )

end if

end for

## CHAPTER 4

### RESULT AND DISCUSSION

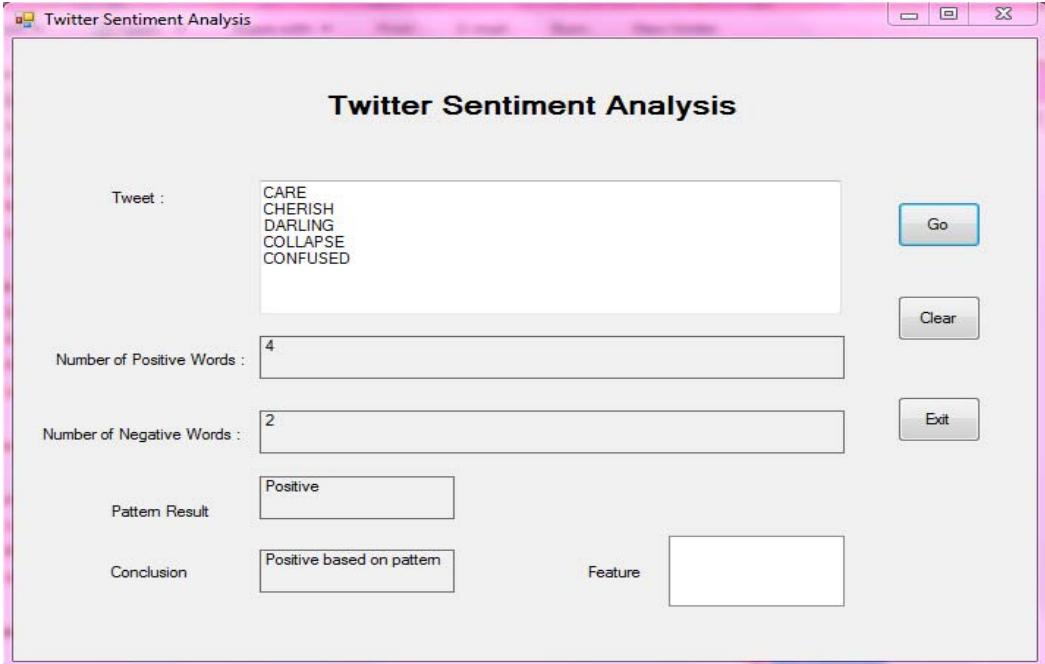
The system will generate 2 possible answers. It could be based on:

- i. Pattern stored in either positive case base lexicon or negative case base lexicon.
- ii. Amount of frequent occurrence.

#### 4.1 Case Base Pattern Testing:

In this section, the testing part is to test the pattern stored in case base lexicons.

The result should be based on the patterns that are stored in the lexicons.



The screenshot displays a software window titled "Twitter Sentiment Analysis". The interface includes a text input field for a tweet containing the words "CARE", "CHERISH", "DARLING", "COLLAPSE", and "CONFUSED". Below this, the system has calculated "Number of Positive Words" as 4 and "Number of Negative Words" as 2. The "Pattern Result" is shown as "Positive", and the "Conclusion" is "Positive based on pattern". There are also buttons for "Go", "Clear", and "Exit", and a "Feature" field.

Field	Value
Tweet	CARE CHERISH DARLING COLLAPSE CONFUSED
Number of Positive Words	4
Number of Negative Words	2
Pattern Result	Positive
Conclusion	Positive based on pattern
Feature	

**Figure 9 : Result Based on Pattern**

As stated in Figure 9, the result is generated based on the patterns that are stored in the lexicon. The “Pattern Result” indicates that there is a pattern detected in the pattern lexicon. The pattern lexicon detected can be either from the positive case base or negative case base.

Twitter Sentiment Analysis

**Twitter Sentiment Analysis**

Tweet : CARE  
CHERISH  
DARLING  
COLLAPSE  
CONFUSED  
CLUMSY

Number of Positive Words : 4

Number of Negative Words : 3

Pattern Result : Pattern not detected

Conclusion : Positive based on number of words

Feature :

Go

Clear

Exit

**Figure 10 : Result Based on Frequent Words**

However, if the first step could not be executed which the pattern could not be identified, the system will indicate that the pattern does not exist. The system will find alternative way to generate result. Figure 10 above is showing the result based on the calculation of the frequent word. In this system, alternative way to generate the output is by calculating the frequent words in the textbox. The words in the textbox will be compared with the both positive and negative lexicons. The system will count the numbers of words that exist in both the lexicons. If the positive word is greater than the negative word, thus the answer is positive. If the negative word is

greater than the positive word, the result will be negative. The result will be neutral when the word does not exist in either positive lexicon or negative lexicon.

#### **4.2 Human Evaluation Testing**

In order to test the pattern functionality, human evaluation test should be done. However, the sample tweets involved in human evaluation testing will be created and labelled beforehand.

Below is the sample of tweets that are labelled in bracket with system pattern:

- 1) I love iphone. The colour is chic. It enhances the resolution but the sound is bad. (positive)
- 2) I love iphone not samsung. It is so cool!!!!!!!!!!!! (positive)
- 3) The iphone camera is fab dope but the screen is lousy (positive)
- 4) The keypad is small and it stresses me out!!!! (negative)
- 5) Damn! The weight is superb. I really hate the sound. It is so annoying and it makes me angry. (negative)

**Answer : P P P N N**

Survey is done on 7 of specific respondents that is not native English speaker but understand all the words stated. They were asked to label the tweets mentioned above.

Respondent 1 : P P P N N

Respondent 2 : P P P N N

Respondent 3 : P P N N N

Respondent 4 : N P N N N

Respondent 5 : N P N N N



Respondent 6 : P P P N N

Respondent 7 : P P P N N

For the Question 1, the system generated positive answer. From the 7 respondents, 5 out of 7 of the respondents feel that the sentence is positive while 2 out of 7 find the sentence is negative.

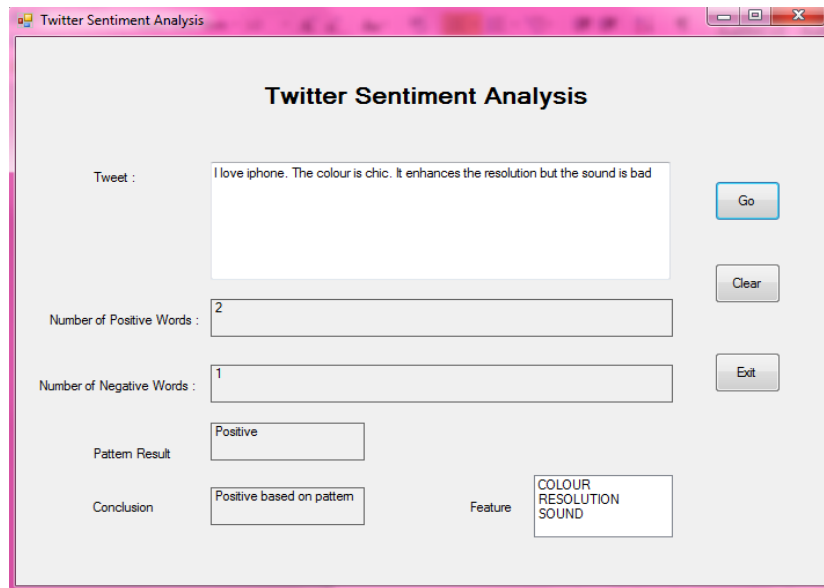
Question 2, the system generated positive answer. All of the respondents feel that the sentence is positive. None of the respondent answered negative.

The system generated positive answer for Question 3. 4 out of 7 answered the sentence is positive while the other 3 answered negative.

For the Question 4, the system generated negative answer. For this set of question, all of the respondents answered negative.

The last Question, the system generated negative answer. All the respondents answered negative for the statement.

Based on the human evaluation testing, the only differences that can be found are at Question 1 and Question 3. Observation made in the human evaluation is, when there is a negative word at the end of the tweet, respondents will label the tweet as negative regardless of the amount of positive words before. This is because the intelligence and perspective of one person differs from the other.



**Figure 11 : Human Evaluation Testing (Positive outcome)**

For the future work, Precision and Recall method will be done. This is to ensure the accuracy of the system is achieved. However, as stated by Euzenat (2007), precision and recall share the same value. The alignment could be very close to the expected result and some could be far. Nevertheless, Precision and Recall will be conducted once the system is completely done.

## **CHAPTER 5**

### **CONCLUSION**

We live in the era where technology is successfully growing and information is accessible anywhere and anytime. Having a system that could automatically produce result over a statement that is posted on Twitter could improve the quality of the outcome. For an example, a company does not need to spend expensive cost over the Research and Development of their product. Instead, with the advancement of technology as machine learning, artificial intelligence could help the company finding loop hole on their product. With that, better quality of the products and services could be improved from time to time.

Sentiment Analysis on Product Tweets is focusing on simple tweet. Direct tweet is said by having one feature in one tweet or a sentence that has emotion stated such as “love”, “hate” or “bad”. However, there are some limitations to the system. It will be mentioned below.

#### **5.1 Limitation of the Project**

Sentiment Analysis on Product Tweets will be focusing on product such as Iphone. This could be one of the limitations of this system. It will detect tweets that are related to Iphone only. Other than that, emoticons such as smiley “☺”, numbers, special characters and symbols that will indicate the level (strong or weak) of satisfaction will not be taken into consideration in this project. Tweets may contain special characters. However, once pasted in the system, special characters and numbers will be discarded automatically. The system will be able to eliminate those special characters.

Plus, this system will not be able to detect navigating sentences and will not be able to detect sentences that have linguistic differences and cultural factors. With the development of 14 weeks, only direct sentences will be taken into consideration.

## **5.2 Recommendation**

This system will be improved if navigating sentences could be incorporated.

## REFERENCES

- Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In Proceedings of the Workshop on Languages in Social Media (pp. 30-38). Association for Computational Linguistics. Retrieved 10 June 2013, from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.207.5100&rep=rep1&type=pdf#page=40>
- Chen, Y. Y., & Lee, K. V. (2011). User-Centered Sentiment Analysis on Customer Product Review. World Applied Sciences Journal, 12, 32-38. Retrieved 10 June 2013, from <http://idosi.org/wasj/wasj12%28CA&KM%29/6.pdf>
- Dave, K., Lawrence, S., & Pennock, D. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. Paper presented at the International World Wide Web Conference, Budapest, Hungary.
- Dennis, A., Wixom, B. H., & Tegarden, D. P. (2010). Systems Analysis and Design with UML: An Object-oriented Approach (3rd Ed): John Wiley & Sons, Incorporated.
- Enchanted Learning (2012). Negative Vocabulary Word List. Retrieved 14 October 2013, from <http://www.enchantedlearning.com/wordlist/negativewords.shtml>

Euzenat, J. (2007, January). Semantic Precision and Recall for Ontology Alignment Evaluation. In IJCAI (pp. 348-353). Retrieved 20 July 2013, from <http://hal.inria.fr/docs/00/81/78/06/PDF/IJCAI07-054.pdf>

Meeker, M. (December 2013). A Social Media Boom Begins In Africa. Using Mobile Phones, Africans Join The Global Conversation. Retrieved 10 June 2013, from <http://www.un.org/africarenewal/magazine/december-2010/social-media-boom-begins-africa#sthash.neiK9msC.dpuf>

Negnevitsky, M. (2005). Artificial intelligence: a guide to intelligent systems. Pearson Education.

Ohana, B., Delany, S. J., & Tierney, B. (2012). A case-based approach to cross domain sentiment classification. In Case-Based Reasoning Research and Development (pp. 284-296). Springer Berlin Heidelberg. Retrieved 13 June 2013, from [http://arrow.dit.ie/cgi/viewcontent.cgi?article=1118&context=scschcomcon&sei-redir=1&referer=http%3A%2F%2Fscholar.google.com%2Fscholar%3Fq%3DA%2BCase-Based%2BApproach%2Bto%2BCross%2BDomain%2BSentiment%2BClassification%26btnG%3D%26hl%3Den%26as\\_sdt%3D0%252C5#search=%22Case-Based%20Approach%20Cross%20Domain%20Sentiment%20Classification%22](http://arrow.dit.ie/cgi/viewcontent.cgi?article=1118&context=scschcomcon&sei-redir=1&referer=http%3A%2F%2Fscholar.google.com%2Fscholar%3Fq%3DA%2BCase-Based%2BApproach%2Bto%2BCross%2BDomain%2BSentiment%2BClassification%26btnG%3D%26hl%3Den%26as_sdt%3D0%252C5#search=%22Case-Based%20Approach%20Cross%20Domain%20Sentiment%20Classification%22).

Osimo, D., Mureddu, F. (2010). Research Challenge on Opinion Mining and Sentiment Analysis. The CROSSROAD Roadmap on ICT for Governance and Policy Modeling.

Oxford Dictionaries. (2013). Retrieved 10 June 2013, from

<http://oxforddictionaries.com>

Pang, B., & Lee, L. (2004, July). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics (p. 271). Retrieved 9 June 2013, from [http://acl.ldc.upenn.edu/acl2004/main/pdf/319\\_pdf\\_2-col.pdf](http://acl.ldc.upenn.edu/acl2004/main/pdf/319_pdf_2-col.pdf)

Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics. Retrieved 14 June 2013, from <http://acl.ldc.upenn.edu/acl2002/EMNLP/pdfs/EMNLP219.pdf>

Pantic, M. (n.d.). Machine Learning. Course 395. Introduction to Machine Learning & Case Based Reasoning. Retrieved 16 June 2013, from <http://ibug.doc.ic.ac.uk/media/uploads/documents/courses/syllabus-CBR.pdf>

Paltoglou, G., & Thelwall, M. (2012). Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media. ACM Transactions on Intelligent Systems and Technology (TIST), 3(4), 66.

- Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2), 143-157. Retrieved 12 June 2013, from [www.researchgate.net/publication/222406047\\_Sentiment\\_analysis\\_A\\_combined\\_approach/file/d912f51333facab04a.pdf](http://www.researchgate.net/publication/222406047_Sentiment_analysis_A_combined_approach/file/d912f51333facab04a.pdf)
- Saif, H., He, Y., & Alani, H. (2012). Semantic sentiment analysis of twitter. In *The Semantic Web–ISWC 2012* (pp. 508-524). Springer Berlin Heidelberg. Sanders,
- Sanders, N. (2011). Sanders-Twitter Sentiment Corpus. Retrieved 9 July 2013, from <http://www.sananalytics.com/lab/twitter-sentiment/>
- Sharma, A. & Dey, S. (2012, Dec). A Document-Level Sentiment Analysis Approach Using Artificial Neural Network and Sentiment Lexicons. *Applied Computing Review* Dec. 2012, Vol. 12, No. 4. Retrieved 29 November 2013, from [http://delivery.acm.org/10.1145/2440000/2432552/p67-sharma.pdf?ip=203.135.190.8&id=2432552&acc=ACTIVE%20SERVICE&key=C2716FEBFA981EF1896690448681A712208491652C4B6B09&CFID=266647692&CFTOKEN=33986430&\\_acm\\_=1385895705\\_9c305bb31233826d4dbf7babf775e2bf](http://delivery.acm.org/10.1145/2440000/2432552/p67-sharma.pdf?ip=203.135.190.8&id=2432552&acc=ACTIVE%20SERVICE&key=C2716FEBFA981EF1896690448681A712208491652C4B6B09&CFID=266647692&CFTOKEN=33986430&_acm_=1385895705_9c305bb31233826d4dbf7babf775e2bf)
- Turney, P.D. (2002, July). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Paper presented at the 40th anniversary meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA. pp. 417-424. Retrieved 29 July 2013, from [nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?action=rtdoc&an=8914166](http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?action=rtdoc&an=8914166)
- TwitterTM. (2013). About Twitter. Retrieved 21 June, 2013, from <https://twitter.com/about>



Winspiration. (14 May 2006). List of Positive Words. Retrieved 12 October 2013,  
from <http://www.winspiration.co.uk/positive.htm>

## APPENDIX

1. The coding of the system as for 2 December 2013

```
Imports System.Text.RegularExpressions
```

```
Public Class Form1
```

```
    Dim tweets As String
```

```
    Dim tweets1 As String
```

```
    Dim positiveLexi As String
```

```
    Dim negativeLexi As String
```

```
    Dim featureLexi As String
```

```
    Dim cbpositiveLexi As String
```

```
    Dim cbnegativeLexi As String
```

```
    Dim positiveLexiArray() As String
```

```
    Dim negativeLexiArray() As String
```

```
    Dim featureLexiArray() As String
```

```
    Dim cbpositiveLexiArray() As String
```

```
    Dim cbnegativeLexiArray() As String
```

```
    Private Sub Form1_Load(ByVal sender As System.Object, ByVal e  
As System.EventArgs) Handles MyBase.Load
```

```
        Dim positive As System.IO.StreamReader
```

```
        positive = New System.IO.StreamReader("C:\Users\Ira  
Iryani\Desktop\FYP2\positive_lexicon.txt")
```

```
        positiveLexi = positive.ReadToEnd()
```

```
        positiveLexiArray = positiveLexi.Split()
```

```
        positive.Close()
```

```
Dim negative As System.IO.StreamReader
negative = New System.IO.StreamReader("C:\Users\Ira
Iryani\Desktop\FYP2\negative_lexicon.txt")
```

```
negativeLexi = negative.ReadToEnd()
```

```
negativeLexiArray = negativeLexi.Split()
negative.Close()
```

```
Dim feature As System.IO.StreamReader
feature = New System.IO.StreamReader("C:\Users\Ira
Iryani\Desktop\FYP2\features_lexicon.txt")
```

```
featureLexi = feature.ReadToEnd()
featureLexiArray = featureLexi.Split()
feature.Close()
```

```
Dim cb As System.IO.StreamReader
cb = New System.IO.StreamReader("C:\Users\Ira
Iryani\Desktop\FYP2\casebasepattern_positive.txt")
```

```
cbpositiveLexi = cb.ReadToEnd()
```

```
cbpositiveLexiArray = cbpositiveLexi.Split()
cb.Close()
```

```
Dim cb2 As System.IO.StreamReader
cb2 = New System.IO.StreamReader("C:\Users\Ira
Iryani\Desktop\FYP2\casebasepattern_negative.txt")
```

```
cbnegativeLexi = cb2.ReadToEnd()
```

```
cbnegativeLexiArray = cbnegativeLexi.Split()  
cb2.Close()
```

```
End Sub
```

```
Private Sub btnGo_Click(ByVal sender As System.Object, ByVal e  
As System.EventArgs) Handles btnGo.Click
```

```
    tweets = txtTweet.Text 'tweets  
    tweets = UCase(tweets)
```

```
    Dim positiveInt As Integer = 0  
    Dim negativeInt As Integer = 0
```

```
    Dim TweetArray() As String  
    Dim TweetArray1() As String  
    Dim cbString As String = ""
```

```
    tweets1 = Regex.Replace(tweets, "[^a-zA-Z]", " ")
```

```
    TweetArray = Split(tweets1) 'split word by word  
    Dim counter As Integer = 0
```

```
    'to throw away array space that is empty  
    For count = 0 To UBound(TweetArray)
```

```
        If TweetArray(count) <> "" Then  
            ReDim Preserve TweetArray1(counter)
```

```

        TweetArray1(counter) = TweetArray(count)
        counter += 1
    End If
Next

For count = 0 To UBound(TweetArray1)

    For tcnt = 0 To UBound(featureLexiArray)

        If TweetArray1(count) = featureLexiArray(tcnt) Then
            lst1.Items.Add(featureLexiArray(tcnt))
        End If

    Next

    'positive words
    Dim cbcount As Integer = 0

    For tcount2 = 0 To UBound(positiveLexiArray)

        If TweetArray1(count) = positiveLexiArray(tcount2)
Then
            positiveInt += 1

            cbString &= "1"

        End If

    Next

    'negative words

```

```

For tcount3 = 0 To UBound(negativeLexiArray)

    If TweetArray1(count) = negativeLexiArray(tcount3)

Then
        negativeInt += 1

        cbString &= "0"
    End If

Next

Next
lblResultPositiveSentiment.Text = positiveInt
lblResultNegativeSentiment.Text = negativeInt

Dim answerCount As String

If positiveInt = 0 And negativeInt = 0 Then
    'lblConclusionAnswer.Text = "Neutral"
    answerCount = "Neutral"

ElseIf positiveInt > negativeInt Then
    'lblConclusionAnswer.Text = "Positive"
    answerCount = "Positive"

ElseIf negativeInt > positiveInt Then
    'lblConclusionAnswer.Text = "Negative"
    answerCount = "Negative"

End If

Dim answerCb As String

```

```

If cbString <> Nothing Then
    For kire = 0 To UBound(cbpositiveLexiArray)
        If cbString = cbpositiveLexiArray(kire) Then
            answerCb = "Positive"
        End If
    Next

    For kire = 0 To UBound(cbnegativeLexiArray)
        If cbString = cbnegativeLexiArray(kire) Then
            answerCb = "Negative"
        End If
    Next

Else
    answerCb = "Neutral"
End If

If answerCount = answerCb Then
    lblConclusionAnswer.Text = answerCb
    lblPatternResult.Text = answerCb
Else
    lblPatternResult.Text = "Pattern not detected"
    lblConclusionAnswer.Text = answerCount & " based on
number of words"

End If

End Sub

Private Sub btnClear_Click(ByVal sender As System.Object, ByVal
e As System.EventArgs) Handles btnClear.Click

    txtTweet.Text = ""

```

```
lblResultNegativeSentiment.Text = "0"  
lblResultPositiveSentiment.Text = "0"  
lblConclusionAnswer.Text = ""  
lblPatternResult.Text = ""  
lst1.Items.Clear()
```

```
End Sub
```

```
Private Sub btnExit_Click(sender As System.Object, e As  
System.EventArgs) Handles btnExit.Click  
Me.Close()
```

```
End Sub
```

```
End Class
```