Identifying Terrorism-Related Postings in Twitter for Malay language

by

Merissa Nazriana binti Mohd Nadzri
13963

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

Terrorism-Related Posting in the Social Media (Twitter)

by

Merissa Nazriana binti Mohd Nadzri

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,

_____
(Professor Dr. Alan Oxley)

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.

_____

MERISSA NAZRIANA BINTI MOHD NADZRI

Table of Contents

List of Figures

List of Tables

ABSTRACT

In this day and age, social media is the fastest way of receiving the latest news. Users have the freedom to post and read any posts that have been posted regarding any topic. Through the social media, Terrorism groups have used this as one of the method to recruit new members. The youths are their main target. A system to detected posting in Twitter is available but only in certain aspect. There is no system to detect terrorism related tweets that are in the Malay language. Therefore, the purpose of the project is to automatically detect terrorism-related tweets in Malay. The social network medium will be Twitter and the tweets that will be focused on terrorism-related in Malay language. The target user for this project would be the people who are using the Twitter. However, the main target user for this project, it would be the youngsters that are using the Twitter. The method that will be used for the program would be Naïve Bayes. Based on the result collected from the pilot study, a couple of patterns can be derived from it. These patterns will be very beneficial in the development of the project.

CHAPTER 1

INTRODUCTION

1.1 Background of Study

Micro blogging or short postings is one of the most popular social networking platforms. Sharing information publicly via social networking sites can be misleading and create problems, especially regarding terrorism. By having a monitoring scheme, we should be able to gather information specifically about terrorism. Since they post the information freely in the social networking sites, therefore it does not breach any privacy of the user. The project will be focused on terrorism related posting in Malay language. The main social networking medium for this is project is Twitter. To detect whether the tweets are terrorism related, a series of keywords will be provided. These keywords will act as a reference for the program to detect terrorism related tweets. The program will alert the user if there is a high possibility of such tweets.

1.2 Problem Statement

Acts of terrorism are always on the news these days. It can also be the topic of conversation in social networking medium. Some of which is information regarding national security. By sharing through these social networking sites, the news spread very quickly. A system to detected posting in Twitter is available but only in certain aspect. There is no system to detect terrorism related tweets that are in the Malay language. Moreover, based on Samuel (2012) researched that there is a lack of awareness on the right use of Internet including social medias by the Malay youths. Furthermore, youths are being targeted by these terrorist groups and the Internet is used as a means of recruitment.

1.3 Objective and Scope of Study

There is no mechanism or program that has been created to automatically detect terrorism related tweets especially in Malay. Therefore the objective of the project is to automatically detect terrorism-related tweets in Malay. The validity of information whether the tweet is a true message or not will not be focused in the project.

The social network medium varies in term of services provided for the user and functionalities that can be used by the user. For example Instagram, the main purpose is to allow users to capture and post picture or video at that very moment, Twitter to share short messages with their followers and Facebook with provide a wide range of functionalities. For the purpose of this project, the social network medium will be Twitter. Twitter can be categorize as a microblog, where the user are able to post or tweet short messages not more than 140 characters. These tweets may contain valuable information that could be about any topic or issues. It could be about politics, world issues, sports updates and much. To narrow down the scope of research, the project will focused on tweets that are terrorism-related. Furthermore, terrorism is one of the most concern issues since September 11. By giving the freedom to user to tweet, the language used by the user is too not bound by any limitation. The tweets can be written in any language of their choice. Quite a number of studies had been conducted in detecting information from tweets in English. However in the Malay language not much study had been completed. Therefore this project will be focusing on terrorism-related tweets in the Malay language. As for the target user for this project, it would be the youngsters that are using the Twitter. Based on Samuel T.K. (2012), the internet was a powerful tool to reach out the young and allowing them to spread their propaganda and attract young individuals who are ‘internet savvy’.

1.4 Relevancy of the project

The project is relevant for the youngster that owns a Twitter account because these youngsters are the prime target for the terrorist group. The system created will helped the youngster to identify whether the tweets are terrorism related. Besides that this project indirectly supports the Southeast Asia Regional Centre for Counter-Terrorism (SEACCT) objective to spread awareness to the youth regarding terrorism.

1.5 Limitations of the project

The project will only be focusing on terrorism related tweets written in Malay. These tweets will be classified by calculating the probability of the terrorism words present in the tweet. The nature of the tweet will not be identified in the project, for example whether if the tweet is just a fabricated threat, personal thoughts or views or etc. The process of lemmatization of the words in the tweet will not be conducted. The words in the tweets will be used as it is. The project will be conducted based on a mock Twitter. The tweets are created for the purpose of the project as there is no available Malay tweet available.

CHAPTER 2

LITERATURE REVIEW OR THEORY

In an overview, Twitter provides a real-time information network. The users may follow the accounts or conversations that they are interested. The short messages that are shared are known as Tweets. Each Tweet is 140 characters long. Besides sharing the Tweets, the users can attach photos, videos and location to the Tweets (Twitter, 2013)[1].

Based on S.Adinaraina and Muhamad Ridzuan from the Correctional Academy of Malaysia[2], there is no internationally agreed upon definition of terrorism. Some of the general definitions of terrorism are acts that intend to create fear, to accomplish ideological goal and the safety of the civilians are taken for granted and will be targeted during the attacks. The definition given by CitizenWorrior[3], is very much similar to the above explanation. That are threat or violence are used as the main element of the activity, the objective of the activities are tied together with political agenda and the civilians are the target of the terrorism.

Artificial Intelligence (AI) is the ability of the computer to act or think, like a human being. In video games, AI is used, where the computer act as another player (TechTerms, 2010). Based on AcedemicRoom[4], AI aims to create the intelligence of the machines or computers, where the machine is an intelligent agent that can learn from its environment and calculates the actions that have the highest amount of success. Some of the researched that are being done under AI are:

- Knowledge representation
- Planning

[1] https://twitter.com/about
[2] http://jpmportal.prison.gov.my/akademi/images/ArkibAKM/note%20terrorism.pdf
[3] http://www.citizenwarrior.com/2007/07/definition-of-terrorism.html
[4] http://www.academicroom.com/topics/definition-of-artificial-intelligence

- Learning
- Natural Language Processing
- Motion and manipulation
- Perception
- Cybernetics and brain simulation

The learning and natural language processing areas will be the main research fields focused upon in the project. Since the AI research begins, machine learning has been the core of the research. There are two type of learning, unsupervised and supervised learning. In unsupervised learning the machine will have the ability to understand and find patterns based on the input given. On the other hand, supervised learning consists of numerical regression and classification. The numerical regression method focused on the relationship between the inputs and the outputs. It will predict on how the output will change depending on the input. Examples from different aspect will be given, from there the machine will determine which category it will belong to, this is known as the classification method. (AcedemicRoom)

The ability to read and understand the languages that humans speak is known as the natural language processing (NLP). A powerful NLP will be able to acquire knowledge directly from written sources for example the Internet text. Information retrieval and machine translation are some of the applications of NLP. Baeza-Yates (2004), stated that information retrieval used the basic NLP techniques. These techniques are tokenization, stopword removal, stemming and text normalization. In tokenization, the text is split into a sequence of tokens. Each token is considered as a word. From here the punctuation, special characters and numbers are removed. Stopwords usually are functional words or acronyms that depend on the specific knowledge domain. It is common that it does not distinguish any subset of the document and have little reflect in the content. Stemming technique on the other hand, obtain the morphological root of every word. For examples, singulars, plurals, verbal forms and etc. Text normalization involves synonym translation and detecting multiword expressions.

There quite a number of successful applications based on NLP such as summarization, information extraction and question answering are some of the applications. In summarization, the gist of the document is identified. Question answering focused on the answer retrieval (passage retrieval) based on the questioned asked. Information extraction on the hand has some similarities with summarization. The text will be normalized and will be later used for relational databases or ontologies. The techniques used in information extraction are mainly text mining and novelty detection. Based on Church K. W and Rau L. F (1995), the four types of information management are retrieval, categorization, extraction and generation. This is familiar to Baez-Yates research.

The text mining technique has similarities with data mining. The difference is that text mining is designed to handle unstructured or semi-structured data such as email, HTML files and full text documents (Fan W. et. all, 2006). In another opinion, Hearst M (2003), the difference between text mining and data mining is from where the information is extracted. In text mining the information or patterns are extracted from natural language text whereas in data mining the information is extracted from structured database. In text mining, one of the challenges is that the information in the unstructured data is not readily accessible to be used in the computers. It is written in human language, thus natural language interpretation is needed (Dorre J. et all, 1999).

Cohen K.B. and Hunter L. (2008) stated that the common approaches in text mining are rule-based (knowledge-based) approaches and statistical (machine-learning-based) approaches. In the rule-based approaches, there are a few ways on how to apply this approach. Rule-based system will make use the knowledge (general or specific) that is present. Another way is that the rule based-system will use hard-coded patterns or use sophisticated linguistic and semantic analyses. On the hand, statistical or machine-learning–based approach is by using classifiers that can operate on any level.

Besides text mining, usage can be made of a corpus, which is a collection of natural language texts. CorpusLinguistic[5] define Corpus as a collection systematically or random collected text of natural language. These texts will be stored and processed electronically. Corpus does not represent the complete language but just large number of texts which can be used to analyse the linguistic rules. Based on the research written by Flowerdew L. (1998), the corpus technique or analysis focussed on the lexico-grammatical patterning of the text and how certain words co-occur together. However, the corpus analysis has moved into textlinguistic approach.

Reuters Corpus, British National Corpus, Celt and Global Terrorism Database are some example of corpus that had been completed by NGO bodies. These corpuses are based on news article. For example, The Global Terrorism Database (GTD) is where all related news articles regarding terrorism are collected.

Figure 1 Reuters Corpus

Figure 2 British National Corpus

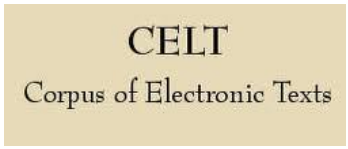


Figure 3 Corpus of Electronic Texts

Figure 4 Global Terrorism Database

[5] http://www.cl2011.org.uk/corpus-linguistics-terms-and-their-meanings.html

Another approach in collecting text is through name entity recognition (NER). NER on the other hand, focussed on the elements in the text rather than analysing the text as a whole. Based on Ganapathy V. and Krishan V. (2005), the NER is involved in extraction where it will locate and classify atomic elements in the text to predefined categories. Steveson M. and Gaizauskas R. (2000) explained NER as "a process of identifying and categorising names in text." Which is similar the explanation given by Ganapathy V. and Krishan V. In NER, lists of common names are used to provide clues and recognising names. Hence, the process does not involve reading the whole text more on the identifying the name pattern. University of Illinois at Urban-Champaign have created a Name Entity Recognizer however this recognizer will only recognize entities that are person, organization, location and miscellaneous. Below is the example program that was created by the University of Illinois.

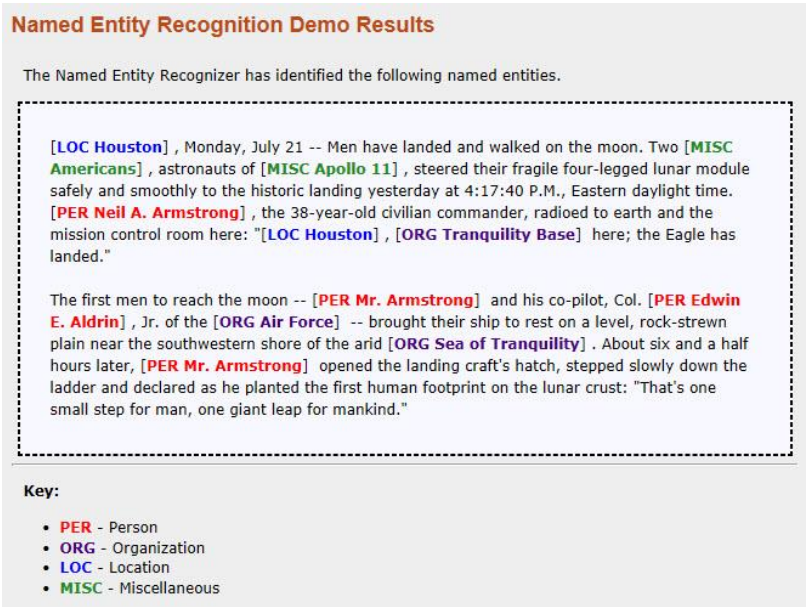


Figure 5 Entity Name Recognizer

Naïve Bayes is another technique for text classification. Based on Zhang (2004) researched one of the most efficient and effective inductive learning algorithm for machine learning and data mining is Naïve Bayes. Rennie (2001), believe that Naïve Bayes is commonly used in text classification. Moreover this technique is suited to perform multiclass text classification. Naïve Bayes uses probability to calculate and classify the text.

The theory of Naïve Bayes:

Assumption: All inputs are independent.

$$P(C|X_1, X_2, \ldots, X_n) = \frac{P(X_1, X_2, \ldots, X_n|C)\,P(C)}{P(X_1, X_2, \ldots, X_n)}$$

As shown above, it is the equation that is used in Naïve Bayes classification. Example of Naïve Bayes Classification in Deciding a Tennis Playing game (ComputerScienceSource, 2010):

- The training table: four attributes will be used to decide whether to play tennis. These attributes are outlook of weather, temperature, humidity and wind. The training table will show the condition of each attribute that will be preferable to play tennis and vice versa.

Day	Outlook	Temperature	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	*Overcast*	*Hot*	*High*	*Weak*	*Yes*
4	*Rain*	*Mild*	*High*	*Weak*	*Yes*
5	*Rain*	*Cool*	*Normal*	*Weak*	*Yes*
6	Rain	Cool	Normal	Strong	No
7	*Overcast*	*Cool*	*Normal*	*Strong*	*Yes*
8	Sunny	Mild	High	Weak	No
9	*Sunny*	*Cool*	*Normal*	*Weak*	*Yes*
10	*Rain*	*Mild*	*Normal*	*Weak*	*Yes*
11	*Sunny*	*Mild*	*Normal*	*Strong*	*Yes*
12	*Overcast*	*Mild*	*High*	*Strong*	*Yes*
13	*Overcast*	*Hot*	*Normal*	*Weak*	*Yes*
14	Rain	Mild	High	Strong	No

Table 1 Example of Training Table

- After creating the training table, the look-up table will be created for each attributes. The look-up table will contain the probability of a tennis game being played based on the attributes. Furthermore based on the training table, there are 9 cases of tennis being played and 5 cases tennis are not being played.

OUTLOOK	Play = Yes	Play = No	Total
Sunny	2/9	3/5	5/14
Overcast	4/9	0/5	4/14
Rain	3/9	2/5	5/14

Table 2 Look-Up Table for Outlook

TEMPERATURE	Play = Yes	Play = No	Total
Hot	2/9	2/5	4/14
Mild	4/9	2/5	6/14
Cool	3/9	1/5	4/14

Table 3 Look-Up Table for Temperature

HUMIDITY	Play = Yes	Play = No	Total
High	3/9	4/5	7/14
Normal	6/9	1/5	7/14

Table 4 Look-Up Table for Humidity

WIND	Play = Yes	Play = No	Total
Strong	3/9	3/5	6/14
Weak	6/9	2/5	8/14

Table 5 Look-Up Table for Wind

- Probabilities for P(C)
  - P(Play=Yes) = 9/14
  - P(Play=No) = 5/14

- A new instance is given. By using the look-up table, the probability for the new instance will be calculated.
  - X = (Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong)

- The look-up table will be used to look for the probability being game:
    - P(Outlook=Sunny | Play=Yes) = 2/9
    - P(Temperature=Cool | Play=Yes) = 3/9
    - P(Humidity=High | Play=Yes) = 3/9
    - P(Wind=Strong | Play=Yes) = 3/9
    - P(Play=Yes) = 9/14
- Then the probability a game will not be played:
    - P(Outlook=Sunny | Play=No) = 3/5
    - P(Temperature=Cool | Play=No) = 1/5
    - P(Humidity=High | Play=No) = 4/5
    - P(Wind=Strong | Play=No) = 3/5
    - P(Play=No) = 5/14
- Then, using those results, multiple all the probabilities for Play=Yes and Play=No such as:
    - P(X|Play=Yes)P(Play=Yes) = (2/9) * (3/9) * (3/9) * (3/9) * (9/14) = 0.0053
    - P(X|Play=No)P(Play=No) = (3/5) * (1/5) * (4/5) * (3/5) * (5/14) = 0.0206
- Lastly is to divide both results by the evidence, or 'P(X)'. The evidence for both equations is the same, and the values can be found within the 'Total' columns of the look-up tables.
    - P(X) = P(Outlook=Sunny) * P(Temperature=Cool) * P(Humidity=High) * P(Wind=Strong)
    - P(X) = (5/14) * (4/14) * (7/14) * (6/14)
    - P(X) = 0.02186
- Then, dividing the results by this value:
    - P(Play=Yes | X) = 0.0053/0.02186 = 0.2424
    - P(Play=No | X) = 0.0206/0.02186 = 0.9421
- Both probabilities will be compared and the highest value and that is answer.
    - P(Play=Yes | X) = 0.2424

- P(Play=No | X) = 0.9421

- Since 0.9421 is greater than 0.2424 then the answer is 'No', a tennis game cannot be played today.

There are many researched and studies that have been conducted by using all the above approaches on the Web pages and social media. Abdul Razak H. and Norlela S. (2011), stated that one of the challenge in online entries is that the quality of the text are 'dirty' and 'noisy'. This is because the written text could contain a mixture of language, use of phonetic spelling, use of slang, and incorrect spelling. The table below is from their researched.

Table 6 Example of Malay Text

Method	*Example of texts*
Mixed of English and Malay words.	*best ctenienjoy glertengok*
Loss of "alpha-case" information makes it very difficult to identify the sentence boundaries, proper nouns and acronyms.	datgkmmg berbaloi2lah! bessstttsgtmse part lawakmmggelak r tp part ygsedih pun cm nknangismmg best rugikluxtgk.nktgkyg 3d tpptgsgtlak so tgkygbesepnye je
Use of slang	• chettttt.. alasanG • kitekurengsukednganparaplakonnye.
	• Yesssszeee
Flexible use of punctuations symbols or do not use the punctuation symbols at all.	• bagiaku... best gak la cite nie.... lawak pun ada... haha.. • xbeslangsung~!!!!!
Use of phonetic spelling	• *b4* → before, • cu → see you, • *r* → are • *2u* → to you.
Use of initial letters only	• *hw* → *homework,* • *wrt* → *with respect to* • *sc* → *stay cool*
Use of a new form of written representation to express emotion	• :) → smiling • *:0* → *shocked* • *:-* → *skeptical.*
Dropping the vowel	• tp→tapi • tgk→tengok
Grammatical errors	• manakenakalangilakasyahyg FK lakonkanmmgtakbape gila2 sgttu. • viewplgcantikmasarapunzelngan u-jinlepak d perahutgk lampu2 berterbangan (x tau apanamabendatuhehe)...

A study was also conducted on using NER in tweets by Ritter A., Clark S., Mausam and Etzioni O in 2011. They conducted this study because the performances of NLP tools are very poor when it is used on the tweets. To improve the performance, they conducted on an experiment to re-build the NLP pipeline from part-of-speech tagging, through chunking, to named-entity recognition.

Even though there is a vast amount of research on NLP, most of this is based on the English language. Based on the paper written by Abdul Razak H. and Norlela S, they confirmed that there is no translation of micro-texts for the Malay language. Moreover the NER and NLP tools for the Malay language are also unavailable. Based on S.A. Noah, A.Y. Amruddin, and N. Omar, due to the non-availability of a lexical database similar to the Wordnet conducting a text similarity is difficult.

Twitter is being widely used all around the world. People tweets about their daily activities, events, and even their comments on the worldly news. To know the authentications of each new that is being shared on Twitter is true or not is hard. A study a have been conducted by Sreenivasan N.D, Lee C.S and Goh D.H (2011), to study on how the Twitter is used when there is a crisis. They concluded that tweets are based on well-informed facts however there quite a number that tweet based on their experience and humour.

CHAPTER 3

METHODOLOGY

The project objective is to automatically detect tweets in Malay that are terrorism-related messages or events. On order to accomplish this objective, Naïve Bayes method will be used as the basis of this project. Each tweet that is posted will be automatically scanned and calculated to identify the probability of the tweet being related to terrorism.

3.1 Research Methods

The research method that will be used in this study is prototyping-based methodology.

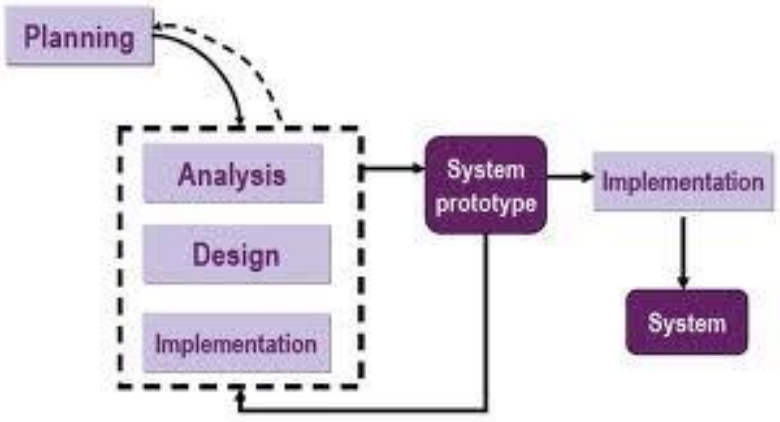


Figure 6 Systems Analysis and Design with UML: An Object-Oriented Approach 3rd Ed (pp. 12-13)

This method is well suited for projects that have unclear user requirements, schedule visibility and a short time schedule. The duration to complete the project is for less than 28 weeks only. With this method a prototype can be developed fast and improvement can be made to achieve a functional program.

Gantt Chart

	FYP 1														FYP 2													
Detail 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																												
Initial Background Study																												
Project Approval																												
Literature Review																												
Submission of Extended Proposal																												
Analysis Phase																												
Research Terrorism Related Words																												
Understanding in Algorithm																												
Tabulation of data																												
Analysis of application process																												
Design Phase																												
Interface Sketch and Design																												
Proposal Defence																												
Design of Application Architecture																												
Submission of Interim Report																												
Implementation Phase																												
Application Development																												
Application Deployment																												
Usability Testing																												
Progress Report Submission																												
Tabulation of usability data & feedback																												
Improvement of prototype																												
Pre-Sedex																												
Viva																												
Project Dissertation Submission																												

The project Gantt chart will cover Final Year Project 1 (FYP1) and Final Year Project 2 (FYP2). The duration of the Gantt chart is for 28 weeks. In the FYP 1, the planning, analysis and design phase will be conducted. While in FYP2, the development, implementation and testing face will be conducted.

Key Milestones

No	Deliverables/Activities	Schedule
1	Title Selection and Proposal	Week 2
2	Project Approval	Week 4
3	Problem Identification	Week 5
4	Extended Proposal	Week 6
5	Requirement Gathering	Week 8
6	Understanding of Algorithm	Week 10
7	Interface Design	Week 12
8	Proposal Defence	Week 12
9	Interim Report	Week 14
10	Architecture and Application Design	Week 15
11	Application Prototype Complete	Week 17
12	Progress Report	Week 20
13	Usability Testing	Week 21
14	Pre-Sedex	Week 24
15	Viva	Week 27
16	Final Dissertation	Week 28

Yellow: Activities Green: Deliverables

The key milestones consist of deliverables that are needed to be presented to the university and the activities that need to be completed to complete the project. The key milestones that have been completed are until the Pre-Sedex (Week 24). For FYP1, all of the key milestones have been achieved. Next key milestone should be achieved are Viva and Final Dissertation.

3.2 Data Collection Methods

The data collection will be used to create the list of terrorism related words that may be used in the tweets. The list will be made based on a few resources that are:

a) Malay Language newspaper or article
b) Online blogs related to Terrorism
c) Malay Dictionary
d) Crowd Sourcing

The objective of the crowd sourcing is to collect and analyse the pattern on how people write certain words for chatting or posting. The words given are related to terrorism. The crowd sourcing form can be found in Appendix 1.

The data collected will not be only ‘clean’ text but also ‘dirty’ and ‘noisy’ text. This is because most of the words used on the social media are ‘dirty’ and ‘noisy’ text. Furthermore, the list will be limited to not more than 100 words as to provide limitation for the project and not to be overwhelmed with the available words that can be gathered.

3.3 Design and Implementation

To build the program, Microsoft Visual Basic will be used. The program will be running on a mock Twitter that is created for the purpose of testing.

3.3.1 Requirement Analysis

The project will be only focusing on the terrorism related tweets in Malay language. The main function of the program is only to detect terrorism related tweet. The validity of the tweet will not be evaluated on the project. For example if the tweet is meant as a joke, false statement or a true statement. As for the collection of data, it will be limited to not more than 100 words as to provide limitation for the project and not to be overwhelmed with the availability of words that can be gathered.

3.3.2 System Architecture

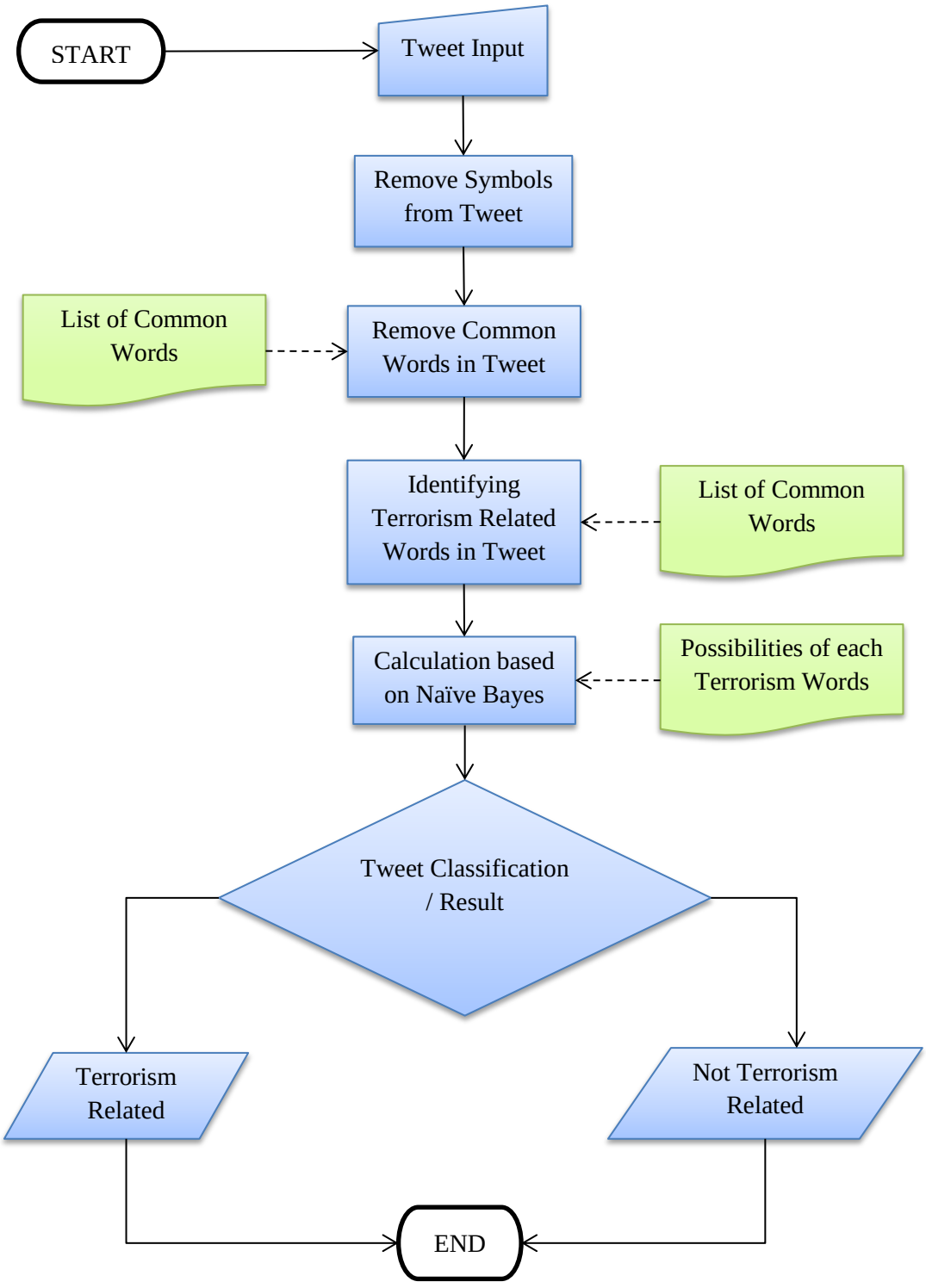


Figure 7 Flow Chart

At the start of the system, the user will manually input the tweet in the text box. From the tweet, the system will remove symbols such as !,@,#,$,%,^, &, *, {,[,},] and etc. The hyphen symbol will not be remove as it is used to write certain words in Malay such as *'bunuh-membunuh'*.

The next process the system will be removing the common words. These common words are mixture of nouns and prepositions. The common words are stored in a text file shown in Figure 8. The motif of removing the common words is to narrow down the number of words present in the selected tweet. It is to increase the efficiency of the system.

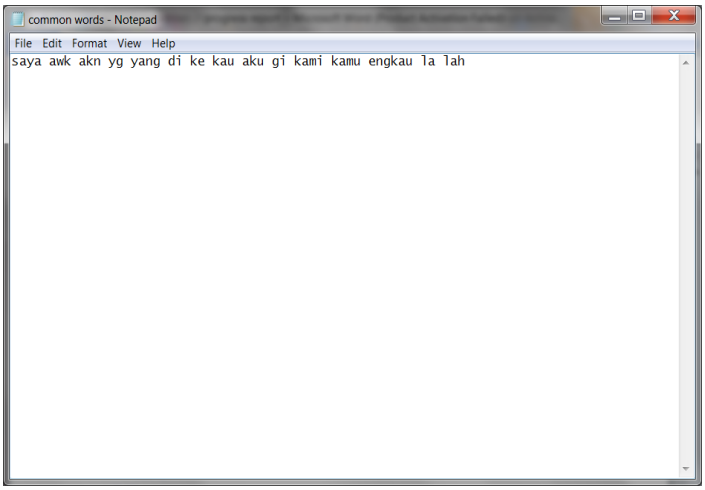


Figure 8 Text File for Common Words

The left over words are then compared to the terrorism related words. These words are listed in a text file as shown in Figure 9. The next process will be to calculate the tweet using Naïve Bayes. Four possibilities will be calculated for each word that are:

i) Possibility of being terrorism related when present
ii) Possibility of not being terrorism related when present
iii) Possibility of being terrorism related when not present
iv) Possibility of not being terrorism related when not present

All of the possibilities are calculated in an Excel spreadsheet. The system will retrieve the data from the Excel spreadsheet.

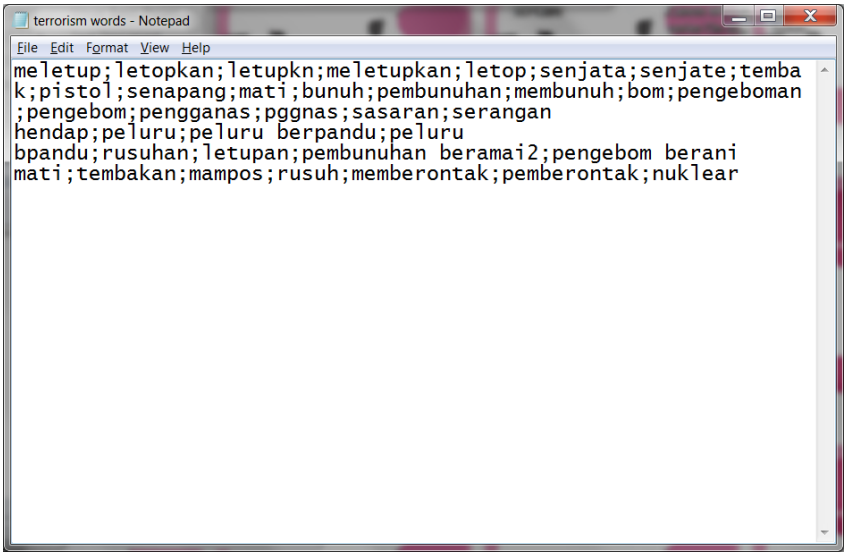


Figure 9 Text File for Terrorism Related Words

Based on the calculations that have been performed using Naïve Bayes, the tweet will go through the classification process. The classification is completed by comparing the result of the possibility of being a terrorism related tweet and the possibility of not being a terrorism related tweet. The tweet will be classified based on the highest possibilities.

Calculation for Tweet Classification:

P(TweetClassificationAsTerrorism) = [P(TerrorismWordPresent-TerrorismRelated) x P(TerrorismWordNotPresent-TerrorismRelated)] x P(TerrorismRelated) / TotalEvidence

P(TweetClassificationAsNotTerrorism) = [P(TerrorismWordPresent-NotTerrorismRelated) x P(TerrorismWordNotPresent-NotTerrorismRelated)] x P(TerrorismRelated) / TotalEvidence

3.4 Data Analysis

To analyse the data (tweet) that have been collected, the Naïve Bayes method will be used to complete the tweet classification. Naïve Bayes is one of the widely use approaches in text classification. This approach is conducted by assuming that the probability of one word occurring in a document is not affected by the probability of another word appearing. As an evaluation on the text categorization, the Precision and Recall method is used against humans' results. Precision will show the accuracy from the predicted values while Recall reflects the relevant reviews that are classified correctly (Abdul Razak H. and Norlela S., 2011).

To conduct the testing phase, 10 sample tweets are created to be used. These tweets will run through the system to be classified. At the same time, these tweets will be given to a few individuals in a questionnaire form to be classified based on their understanding. The list of tweets that are created (including English translation) and the questionnaire given to the individuals can be found in Appendix 2. The results will then be compared to the Precision and Recall results.

3.5 System Development

The system is being developed using Visual Basic 2010. In the current development (prototype) of the system, only 15 common words, 35 terrorism related words and 29 training tweets example will be used as the test phase. After the system main processes are bug free and all error checking have been completed then number of common words, related terrorism words and training tweets will be increased.

The current interface that has been developed is as below:

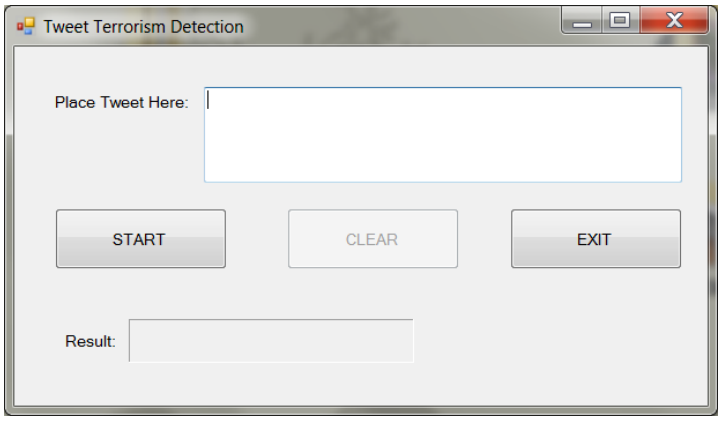


Figure 10Interface for the System

Overview of the system interface referring to Figure 10:

1. Text box

   The function of the text box is to receive input from the user. The tweet will be placed here. The text box is set to only allow 140 characters to be accepted.

2. 'Start' Button

   Upon clicking the start button the program will start the process of the system. The process of tweet classification will be initiated. At first the start button will be disabled and the clear button will be enabled.

3. Result Box

   The result of the tweet classification will be displayed here. It will either display "The tweet is terrorism related" or "The tweet is not terrorism related".

4. ‘Clear’ Button

   The clear button is disabled at the very beginning of the program. It will only be enabled when the start button is click. The clear button is to clear the text box provided and the label box displaying the result. Furthermore, when the button is click the start button will be enabled as well.

5. ‘Exit’ Button

   The exit button is to end and close the program.

CHAPTER 4

RESULT AND DISCUSSION

4.1 Pilot Study

A crowd sourcing was conducted for the pilot study. The objective of the crowd sourcing is to collect the Malay words that the youngster will used based on the terrorism given to them. The crowd sourcing forms were given out to youngster age from 15 to 24. The gender of the youngster does not play an important. The form consist 15 words that are related terrorism. The table below is the summary of the Malay words that have been collected through the crowd sourcing.

Table 7 List of Malay words from Crowd Sourcing

Words in English	Words in Malay
Blast	Letupan, letopan, meletop, letupkan, letopkan, letopkn, letupkn, letup, letop,
Weapon	senjata, senjate
Gunshot	tembak, tembakan, tembakn
Gun	Pistol, senapang
Death	kematian, mati, mampos, mampus, mti
Kill	bunuh, pembunuhan, bunuhn, membunuh
Bom	Bom, pengeboman, pengebom
Terrorist	Pengganas, pggnas
Ambush	Serang hendap, serangan hendap, serangn hendap, serang
Warhead	Peluru, peluru berpandu, peluru bpandu, peluru b'pandu
Suicide Bomber	Pengebom, p'bom, pengebom berani mati, p'bom berani mati, p'bom brani mati
Hijack	Rampasan, rampas, rampasn
Riot	Rusuhan, rusuh, rusuhn, rusuh'n

Genocide	Pembunuhan beramai-ramai, pembunuhan beramai2, pembunuhn beramai2, pembunuhn beramai, pmbunuhan b'ramai2
Rebellion	Pemberontakan, pemberontakn, pmberontakn, berontakn

Based on the table above, a few patterns can be derived from the list of words. One of the obvious patterns among these words is that the consonant of the words remains the same but the vowels changes. The vowels that were used can be different or the vowels are omitted out from the word itself. For example the word '*letupkan*' can be spell as '*letopkan*', '*letupkn*' or '*letopkn*'.

The second pattern can be observed is the addition particle that is added to the word whether it is at the beginning, end or both. However the root word (kata dasar) still remains the same. Based on the table above, the word '*bunuh*' which is the root word can be written as '*pembunuhan*', '*pembunuh*', '*membunuh*' or '*bunuhan*'. The categories for the Malay particle (imbuhan) are as follow:

beR-	beR...an	beR...kan	ter-	ter…i	ter…kan
meN-	meN...i	meN..kan	peR-	peR…an	peN-
peN…an	ke-	ke…an	di-	di…i	di…kan
-an	-man	-wan	-wati	-ita	

Table 8 List of Malay Particle from http://tatabahasabm.tripod.com/bank/tatabahasa.pdf

4.2 Evaluation Testing

The evaluation testing was conducted over 13 people. The user is requested answer 10 sample tweets. The result of the questionnaire is then compared to the system's result. Below are the result based on the questionnaire arrangement:

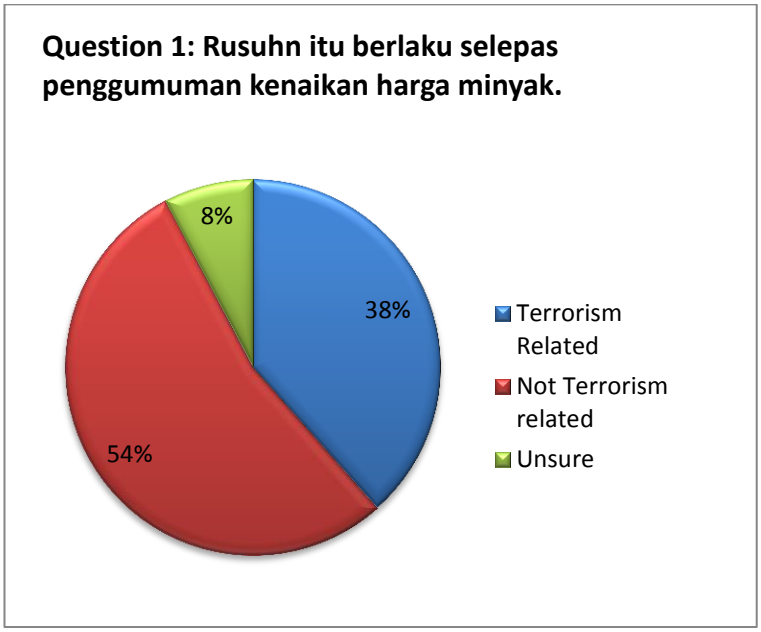


Figure 11 Result of Question 1

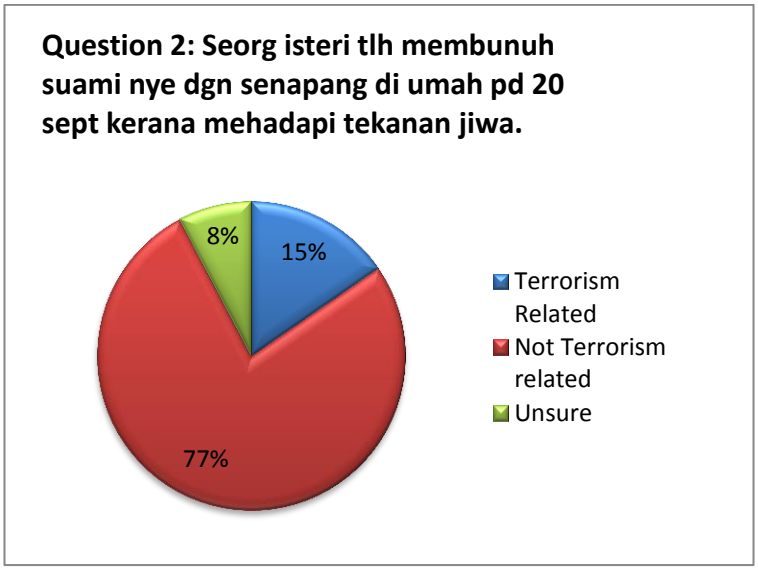


Figure 12 Result of Question 2

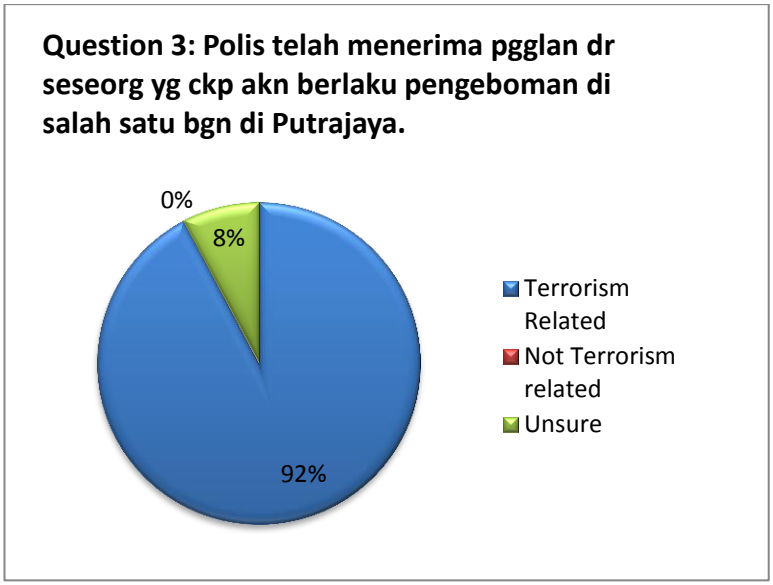


Figure 13 Result of Question 3

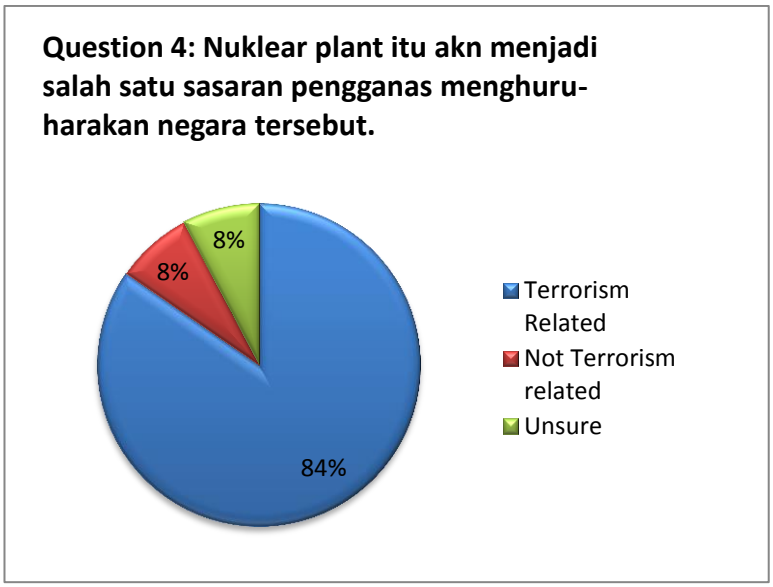


Figure 14 Result of Question 4

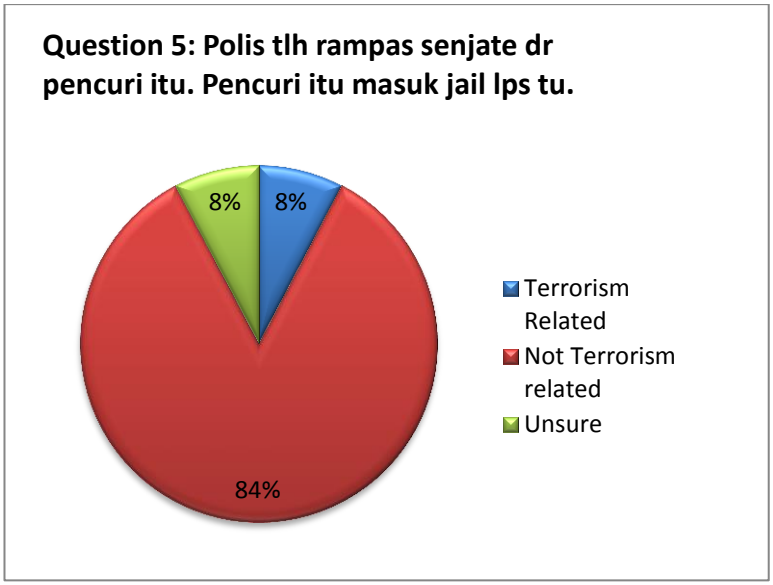


Figure 15 Result of Question 5

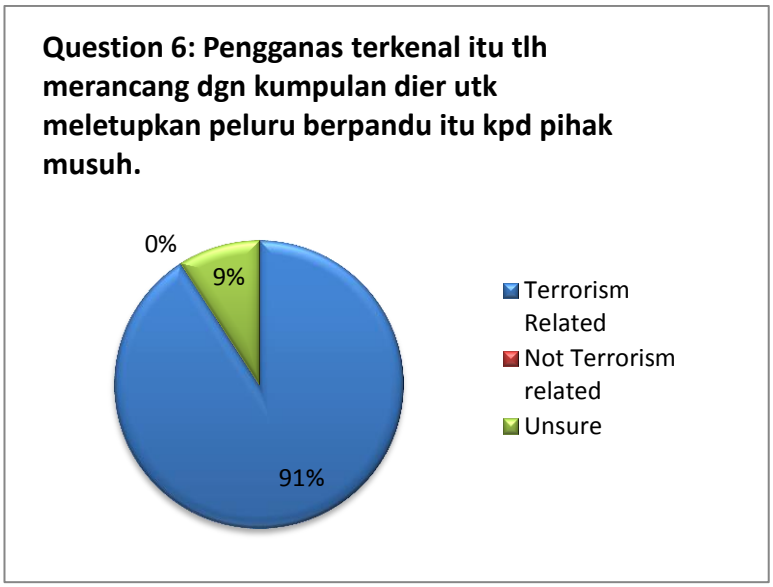


Figure 16 Result of Question 6

Figure 17 Result of Question 7

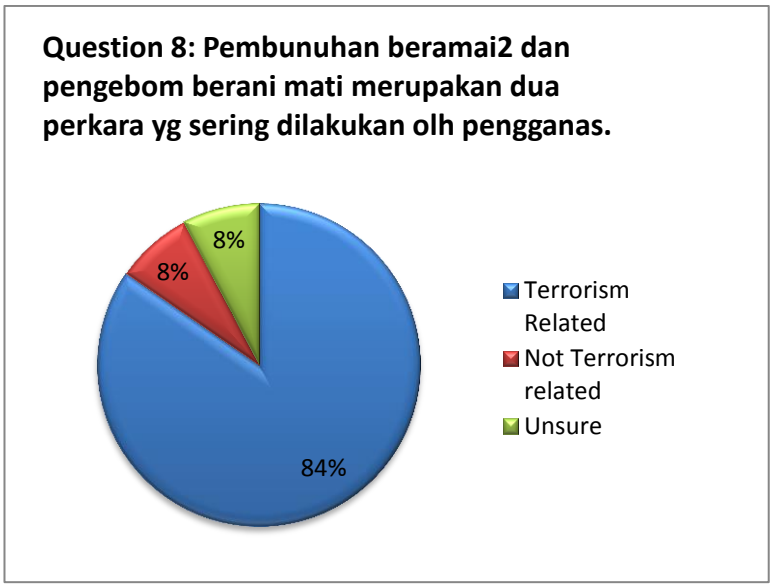


Figure 18 Result of Question 8

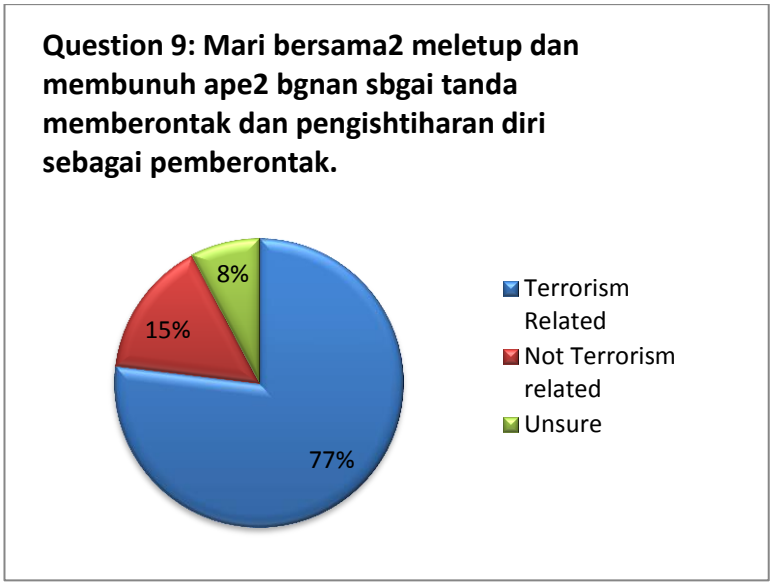


Figure 19 Result of Question 9

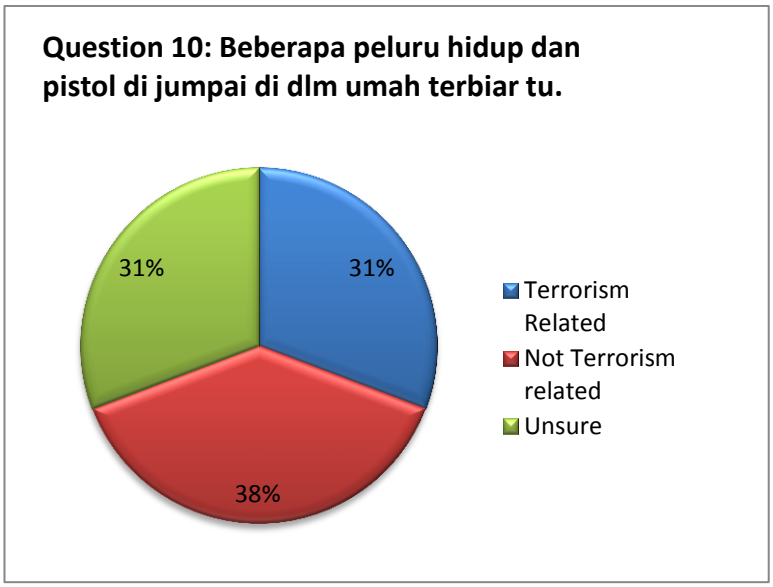


Figure 20 Result of Question 10

Based on the result above it can be concluded that, Question 3, 4, 6, 8 and 9 are terrorism related tweet. On the other hand Question 1, 2, 5, 7 and 10 are non-terrorism related tweets. The classification is completed by following the majority vote.

When the tweets classification were conducted using the system, the result that were gathered are as follow:

i) 3 out 5 Terrorism Related Tweet were categorize as Terrorism related tweet, while 2 out 5 were categorize as non-terrorism related.
ii) 5 out 5 Non-Terrorism Tweet were categorize correctly

By comparing with the Human Evaluation, the system only categorize 2 out of 10 sample tweets wrongly. The 2 sample tweets that were categorize wrongly are Question 6 and Question 9. One of the reasons why these two tweets were categorize differently because the system was not able to detect 'peluru berpandu' as one object. The system will detect both words as a separate object. Even though 'peluru berpandu' is present in the Terrorism Related words text file, it will not be detected. Furthermore by using Naïve Bayes theory, each word is treated independently. The result of categorization is very much related to the testing tweet that was conducted and the list of terrorism words present. There is a high possibility that the testing tweets that were conducted were not sufficient. The Testing Tweet table and the possibilities of each word can be found in Appendix 3.

Despite the fact that only 2 sample tweets were categorized incorrectly, 80% of the sample tweets were categorized correctly. The system can be assumed to be performing quite effectively. Improvement can be made to increase the efficiency of the system in categorizing the tweet.

CONCLUSION

The main objective of the project is to automatically detect terrorism related tweets only written in Malay. There is no available tool that will be able to detect this type of tweets. By the doing this project, it will bring benefit to individual as well as the society. The target users of the project are the Twitter users especially for the youngsters. It will help the youngster from being the victim in involving themselves in Terrorist group.

Future work

For the continuation of the project, the method of extracting the tweet from the text box should be improved. The system should be able to detect two combinations of words that represent a single meaning or object. Based on the result and discussion section, the number of common words, terrorism related words and the training tweets will be increased to provide better result. The codes of the system will be improved to be more efficient and to be able to perform better. The human evaluation results with the system result have been compared. The next step is to conduct the precision and recall testing and will be compared to human evaluation and the system results.

References

Academic Room. (2013). Artificial Intelligence. Retrieved 21 June, 2013, from http://www.academicroom.com/topics/definition-of-artificial-intelligence

Baeza-Yates, R. (2004). Challenges in the Interaction of Information Retrieval and Natural Language Processing. In A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing (Vol. 2945, pp. 445-456): Springer Berlin Heidelberg.

Bekkerman, R., & Allan, J. (2004). Using bigrams in text categorization. Department of Computer Science, University of Massachusetts, Amherst, 1003.

Church, K. W., & Rau, L. F. (1995). Commercial applications of natural language processing. Communications of the ACM, 38(11), 71-79 %@ 0001-0782.

ComputerScienceSource (28 January 2010). Machine Leraning – Naïve Bayes Classifier. Retrieved 19 June, 2013, from http://computersciencesource.wordpress.com/2010/01/28/year-2-machine-learning-naive-bayes-classifier/

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.

Dörre, J., Gerstl, P., & Seiffert, R. (1999, August). Text mining: finding nuggets in mountains of textual data. In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 398-401). ACM.

Flowerdew, L. (1998). Corpus linguistic techniques applied to textlinguistics. System, 26(4), 541-552. doi: http://dx.doi.org/10.1016/S0346-251X(98)00039-6

Hearst, M. (2003). What is text mining. Retrieved February, 7, 2011.

Knowles, G., & Don, Z. M. (2003). Tagging a corpus of Malay texts, and coping with 'syntactic drift'. In Proceedings of the corpus linguistics 2003 conference.

Rennie, J. D. (2001). Improving multi-class text classification with naive Bayes (Doctoral dissertation, Massachusetts Institute of Technology).

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).

Ritter, A., Clark, S., & Etzioni, O. (2011, July). Named entity recognition in tweets: an experimental study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1524-1534). Association for Computational Linguistics.

Samsudin, N., Puteh, M., & Hamdan, A. R. (2011, June). Bess or xbest: Mining the Malaysian online reviews. In Data Mining and Optimization (DMO), 2011 3rd Conference on (pp. 38-43). IEEE.

Samuel T. K. (2012). Reaching the Youth: Countering the Terrorist Narrative. Southeast Asia Regional Centre for Counter-Terrorism, pp. 10-11.

Sreenivasan, N. D., Lee, C. S., & Goh, D. H. L. (2011). Tweet me home: Exploring information use on Twitter in crisis situations. In Online Communities and Social Computing (pp. 120-129). Springer Berlin Heidelberg.

Tan, C. M., Wang, Y. F., & Lee, C. D. (2002). The use of bigrams to enhance text categorization. Information processing & management, 38(4), 529-546.

TechTerms (2010). Artificial Intelligence. Retrieved 21 June, 2013, from http://www.techterms.com/definition/artificial_intelligence

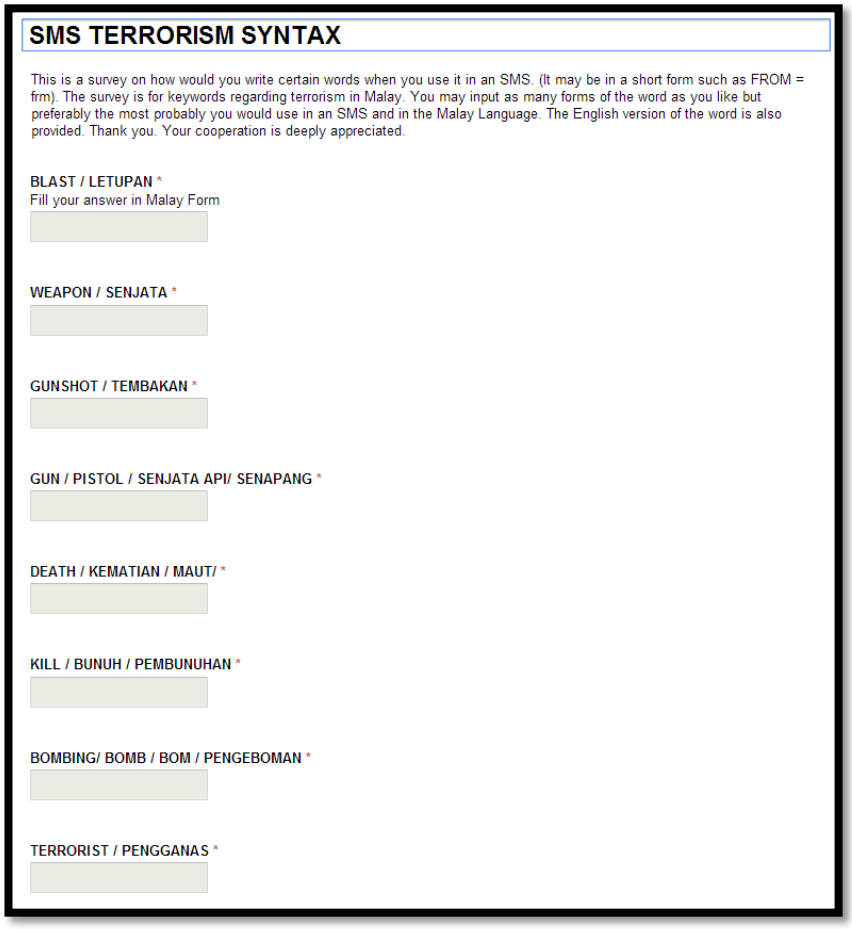
TwitterTM. (2013). About Twitter. Retrieved 21 June, 2013, from https://twitter.com/about

Zhang, H. (2004). The optimality of naive Bayes. A A, 1(2), 3.

Appendix

Appendix 1

Survey to conduct Crowd Sourcing



SMS TERRORISM SYNTAX

This is a survey on how would you write certain words when you use it in an SMS. (It may be in a short form such as FROM = frm). The survey is for keywords regarding terrorism in Malay. You may input as many forms of the word as you like but preferably the most probably you would use in an SMS and in the Malay Language. The English version of the word is also provided. Thank you. Your cooperation is deeply appreciated.

BLAST / LETUPAN *
Fill your answer in Malay Form

WEAPON / SENJATA *

GUNSHOT / TEMBAKAN *

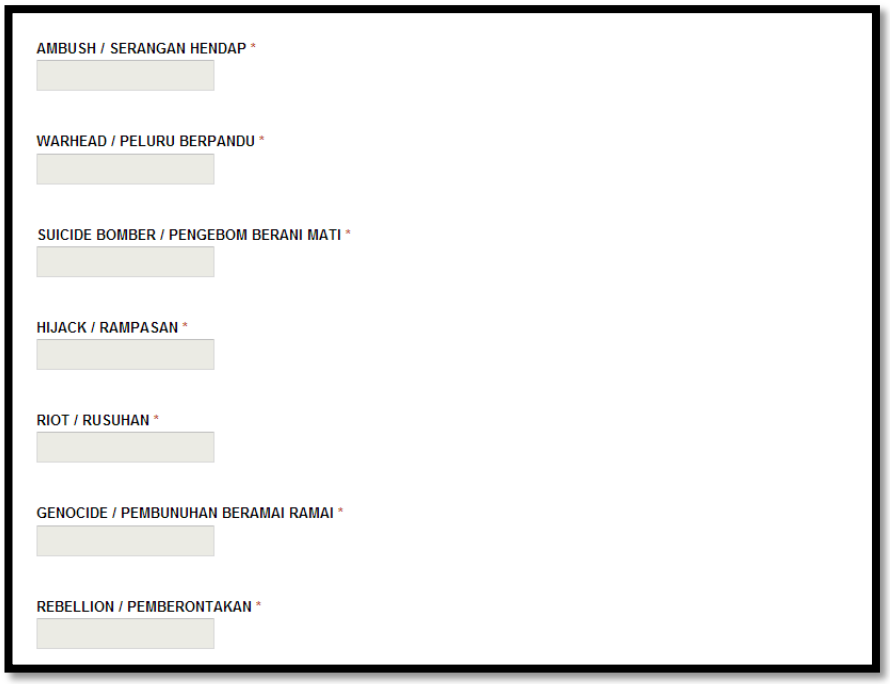
GUN / PISTOL / SENJATA API/ SENAPANG *

DEATH / KEMATIAN / MAUT/ *

KILL / BUNUH / PEMBUNUHAN *

BOMBING/ BOMB / BOM / PENGEBOMAN *

TERRORIST / PENGGANAS *



Appendix 2

List of sample tweets

Terrorism related:

1. *Pembunuhan beramai2 dan pengebom berani mati merupakan dua perkara yg sering dilakukan olh pengganas.*
   Translation: Genocide and suicide bombings are the two things that are done by terrorists.
2. *Polis telah menerima pgglan dr seseorg yg ckp akn berlaku pengeboman di salah satu bgn di Putrajaya.*
   Translation: Police had received a call from someone saying that a bombing would occur in one of the buildings in Putrajaya.
3. *Mari bersama2 meletup dan membunuh ape2 bgnan sbgai tanda memberontak dan pengishtiharan diri sebagai pemberontak.*
   Translation: Lets blow stuff up and kill. We're rebels of the rebellion.

4. *Pengganas terkenal itu tlh merancang dgn kumpulan dier utk meletupkan peluru berpandu itu kpd pihak musuh.*

   Translation: A known terrorist group had planned to launch his missiles to the enemy.

5. *Nuklear plant itu akn menjadi salah satu sasaran pengganas menghuru-harakan negara tersebut.*

   Translation: The nuclear plant will become one of the targets of terrorist to create chaos in that country.

Non-terrorism related

6. *Polis tlh rampas senjate dr pencuri itu. Pencuri itu masuk jail lps tu.*

   Translation: Police had seized weapons of the thief and was imprisoned.

7. *Kucing itu mati terkena letupan bom. Bom itu telah lame tertanam di pdg tersebut.*

   Translation: The cat died in an explosion caused by an old buried bomb.

8. *Rusuhn itu berlaku selepas penggumuman kenaikan harga minyak.*

   Translation: The riots occurred after the announcement of the increase in oil prices.

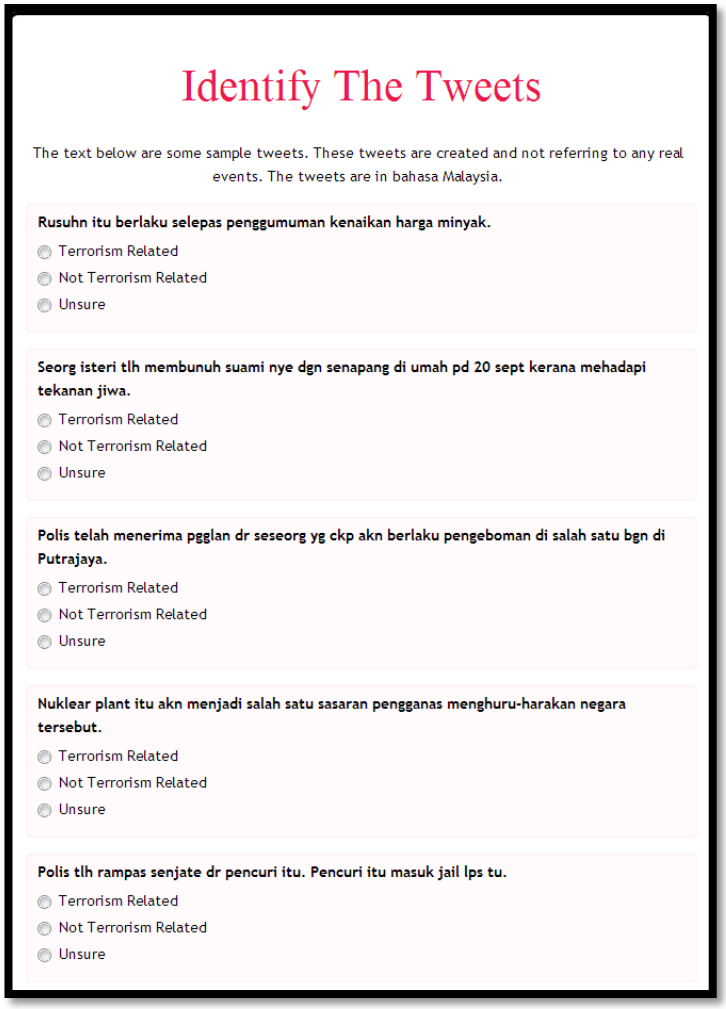
9. *Seorg isteri tlh membunuh suami nye dgn senapang di umah pd 20 sept kerana menghadapi tekanan jiwa.*

   Translation: on 20th sept a depressed wife killed her husband with a rifle.

10. *Beberapa peluru hidup dan pistol di jumpai di dlm umah terbiar tu.*

    Translation: A gun with live ammunitions was found in an abandoned house.

Questionnaire for individuals



Identify The Tweets

The text below are some sample tweets. These tweets are created and not referring to any real events. The tweets are in bahasa Malaysia.

Rusuhn itu berlaku selepas penggumuman kenaikan harga minyak.

- Terrorism Related
- Not Terrorism Related
- Unsure

Seorg isteri tlh membunuh suami nye dgn senapang di umah pd 20 sept kerana mehadapi tekanan jiwa.

- Terrorism Related
- Not Terrorism Related
- Unsure

Polis telah menerima pgglan dr seseorg yg ckp akn berlaku pengeboman di salah satu bgn di Putrajaya.

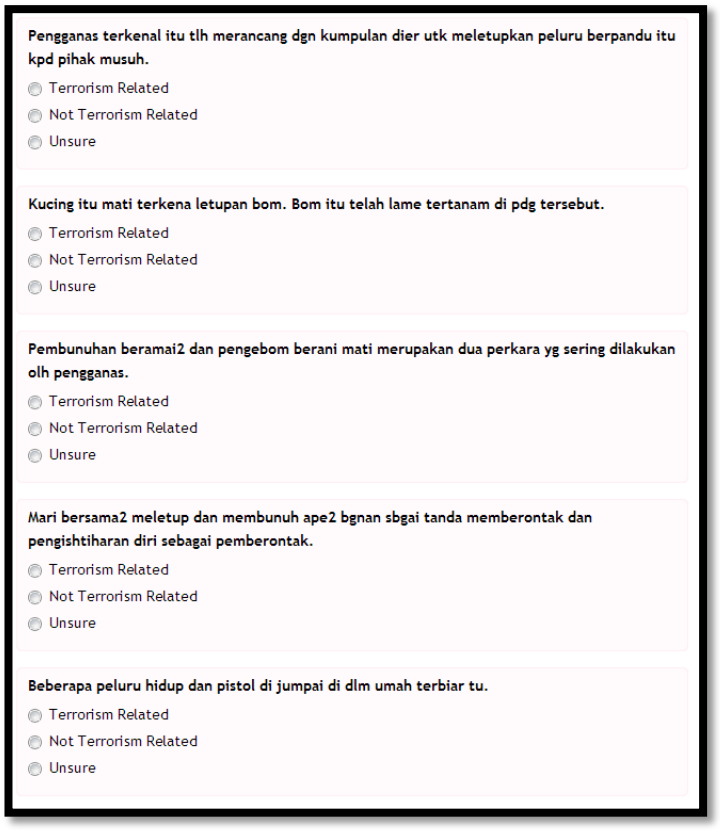
- Terrorism Related
- Not Terrorism Related
- Unsure

Nuklear plant itu akn menjadi salah satu sasaran pengganas menghuru-harakan negara tersebut.

- Terrorism Related
- Not Terrorism Related
- Unsure

Polis tlh rampas senjate dr pencuri itu. Pencuri itu masuk jail lps tu.

- Terrorism Related
- Not Terrorism Related
- Unsure


Pengganas terkenal itu tlh merancang dgn kumpulan dier utk meletupkan peluru berpandu itu kpd pihak musuh.
Terrorism Related
Not Terrorism Related
Unsure
Kucing itu mati terkena letupan bom. Bom itu telah lame tertanam di pdg tersebut.
Terrorism Related
Not Terrorism Related
Unsure
Pembunuhan beramai2 dan pengebom berani mati merupakan dua perkara yg sering dilakukan olh pengganas.
Terrorism Related
Not Terrorism Related
Unsure
Mari bersama2 meletup dan membunuh ape2 bgnan sbgai tanda memberontak dan pengishtiharan diri sebagai pemberontak.
Terrorism Related
Not Terrorism Related
Unsure
Beberapa peluru hidup dan pistol di jumpai di dlm umah terbiar tu.
Terrorism Related
Not Terrorism Related
Unsure

Appendix 3

Training Tweets Table

Training Tweets	meletupkan	letop	senjata	tembak	pistol	senapang	mati	bunuh	pembunuhan	membunuh	bom	pengeboman	pengebom	pengganas	serang hendap	peluru	peluru berpandu	rusuhan	letupan	pembunuhan beramai2	pengebom berani mati	Terrorism
T1											x			x	x						x	Yes
T2	x										x		x							x		Yes
T3		x					x				x								x			Yes
T4			x	x		x				x				x								Yes
T5						x			x					x						x		Yes
T6					x		x								x				x		x	Yes
T7	x							x				x						x		x		Yes
T8									x		x			x	x				x			Yes
T9			x			x				x						x						Yes
T10	x								x		x				x				x			Yes
T11										x		x		x	x						x	Yes
T12													x	x							x	Yes
T13			x	x				x						x		x						Yes
T14										x				x			x					Yes

T15			x					x										x		x	x	Yes
T16			x														x					No
T17	x																					No
T18		x																				No
T19							x				x											No
T20							x			x												No
T21									x													No
T22			x	x				x		x												No
T23				x	x			x														No
T24					x			x										x				No
T25						x				x						x						No
T26														x								No
T27																x						No
T28										x								x				No
T29																			x			No

Possibilities of Each Terrorism Words

T = Yes	T = No		not present	T = Yes	T = No	Total
0.200	0.071		meletup	0.800	0.929	0.138
0.067	0.071		letopkan	0.933	0.929	0.069
0.067	0.071		letupkn	0.933	0.929	0.069
0.200	0.071		meletupkan	0.800	0.929	0.138
0.067	0.071		letop	0.933	0.929	0.069
0.267	0.143		senjata	0.733	0.857	0.207
0.267	0.143		senjate	0.733	0.857	0.207
0.133	0.143		tembak	0.867	0.857	0.138
0.067	0.143		pistol	0.933	0.857	0.103
0.200	0.071		senapang	0.800	0.929	0.138
0.133	0.143		mati	0.867	0.857	0.138
0.200	0.214		bunuh	0.800	0.786	0.207
0.200	0.071		pembunuhan	0.800	0.929	0.138
0.267	0.286		membunuh	0.733	0.714	0.276
0.333	0.071		bom	0.667	0.929	0.207
0.133	0.000		pengeboman	0.867	1.000	0.069
0.133	0.000		pengebom	0.867	1.000	0.069
0.533	0.071		pengganas	0.467	0.929	0.310
0.533	0.071		pggnas	0.467	0.929	0.310
0.533	0.071		sasaran	0.467	0.929	0.310
0.333	0.000		serangan hendap	0.667	1.000	0.172
0.133	0.143		peluru	0.867	0.857	0.138
0.067	0.071		peluru berpandu	0.933	0.929	0.069
0.067	0.071		peluru bpandu	0.933	0.929	0.069
0.133	0.143		rusuhan	0.867	0.857	0.138
0.267	0.071		letupan	0.733	0.929	0.172
0.267	0.000		pembunuhan beramai2	0.733	1.000	0.138
0.333	0.000		pengebom berani mati	0.667	1.000	0.172
0.133	0.143		tembakan	0.867	0.857	0.138
0.133	0.143		mampos	0.867	0.857	0.138
0.133	0.143		rusuh	0.867	0.857	0.138
0.133	0.143		memberontak	0.867	0.857	0.138
0.533	0.071		pemberontak	0.467	0.929	0.310
0.333	0.071		nuklear	0.667	0.929	0.207

Appendix 4

```
Imports System.Text.RegularExpressions
Imports Excel = Microsoft.Office.Interop.Excel

Public Class Form1
    Dim strcommwords() As String
    Dim strterrwords() As String

    Dim PTerrPresYes() As Double = {1}
    Dim PTerrPresNo() As Double = {1}
    Dim PTerrNotPresYes() As Double = {1}
    Dim PTerrNotPresNo() As Double = {1}
    Dim PTotalEvidence() As Double = {1}

    Private Sub Form1_Load(ByVal sender As System.Object, ByVal e As
System.EventArgs) Handles MyBase.Load
        Dim commonwords As System.IO.StreamReader
        commonwords = New System.IO.StreamReader("C:\Users\Merissa
Nazriana\Documents\Visual Studio 2010\Projects\Terrorism Tweet
Detection\Terrorism Tweet Detection\common words.txt")

        Dim temp As String

        Do Until commonwords.Peek = -1
            temp = commonwords.ReadToEnd
            strcommwords = Split(temp)
        Loop

        commonwords.Close()

        Dim terrorismwords As System.IO.StreamReader
        terrorismwords = New System.IO.StreamReader("C:\Users\Merissa
Nazriana\Documents\Visual Studio 2010\Projects\Terrorism Tweet
Detection\Terrorism Tweet Detection\terrorism words.txt")

        Dim temp2 As String

        Do Until terrorismwords.Peek = -1
            temp2 = terrorismwords.ReadToEnd
            strterrwords = Split(temp2, ";")
        Loop

        terrorismwords.Close()

        Dim objXLApp As Excel.Application
        Dim intLoopCounter As Integer
        Dim objXLWb As Excel.Workbook
        Dim objXLWs As Excel.Worksheet
        Dim objRange As Excel.Range

        objXLApp = New Excel.Application
        objXLApp.Workbooks.Open("C:\Users\Merissa
Nazriana\Desktop\Possibilities.xlsx")
        objXLWb = objXLApp.Workbooks(1)
        objXLWs = objXLWb.Worksheets(1)
```

```
        Dim num As Integer = 0
        For intLoopCounter = 2 To
CInt(objXLWs.Cells.SpecialCells(Excel.XlCellType.xlCellTypeLastCell).Row)
            objRange = objXLWs.Range("B" & intLoopCounter)
            ReDim Preserve PTerrPresYes(num)
            PTerrPresYes(num) = objRange.Value
            num = num + 1
        Next intLoopCounter

        Dim num1 As Integer = 0
        For intLoopCounter = 2 To
CInt(objXLWs.Cells.SpecialCells(Excel.XlCellType.xlCellTypeLastCell).Row)
            objRange = objXLWs.Range("C" & intLoopCounter)
            ReDim Preserve PTerrPresNo(num1)
            PTerrPresNo(num1) = objRange.Value
            num1 = num1 + 1
        Next intLoopCounter

        Dim num2 As Integer = 0
        For intLoopCounter = 2 To
CInt(objXLWs.Cells.SpecialCells(Excel.XlCellType.xlCellTypeLastCell).Row)
            objRange = objXLWs.Range("F" & intLoopCounter)
            ReDim Preserve PTerrNotPresYes(num2)
            PTerrNotPresYes(num2) = objRange.Value
            num2 = num2 + 1
        Next intLoopCounter

        Dim num3 As Integer = 0
        For intLoopCounter = 2 To
CInt(objXLWs.Cells.SpecialCells(Excel.XlCellType.xlCellTypeLastCell).Row)
            objRange = objXLWs.Range("G" & intLoopCounter)
            ReDim Preserve PTerrNotPresNo(num3)
            PTerrNotPresNo(num3) = objRange.Value
            num3 = num3 + 1
        Next intLoopCounter

        Dim num4 As Integer = 0
        For intLoopCounter = 2 To
CInt(objXLWs.Cells.SpecialCells(Excel.XlCellType.xlCellTypeLastCell).Row)
            objRange = objXLWs.Range("H" & intLoopCounter)
            ReDim Preserve PTotalEvidence(num4)
            PTotalEvidence(num4) = objRange.Value
            num4 = num4 + 1
        Next intLoopCounter
        objXLApp.Quit()

    End Sub

    Private Sub btnstart_Click(ByVal sender As System.Object, ByVal e As
System.EventArgs) Handles btnstart.Click

        btnstart.Enabled = False
        btnclear.Enabled = True
```

```
Dim strtweet As String 'ori
 Dim strtweet1 As String  'ori with no symbols
 Dim strtweet2() As String  'ori into array
 Dim strtweet3() As String = {} 'common words filtered
 Dim totalevidence As Double = 1.0

 strtweet = txttweet.Text
 strtweet = LCase(strtweet)
 strtweet1 = Regex.Replace(strtweet, "[^A-Za-z2]", " ")


 strtweet2 = Split(strtweet1)
 Dim count As Integer = 0

 For dCount = 0 To UBound(strtweet2)
     Dim countercheck As Integer = 0
     For comcount = 0 To UBound(strcommwords)
         If strtweet2(dCount) <> strcommwords(comcount) Then
             countercheck += 1
         End If
     Next
     If countercheck = strcommwords.Length Then
         ReDim Preserve strtweet3(count)
         strtweet3(count) = strtweet2(dCount)
         count = count + 1
     End If
 Next

 Dim wordcounter(strterrwords.Length - 1) As Integer

 For dCount = 0 To UBound(strtweet3)
     For comcount = 0 To UBound(strterrwords)
         If strtweet3(dCount) = strterrwords(comcount) Then
             wordcounter(comcount) = 1
             totalevidence = totalevidence * PTotalEvidence(comcount)
         End If
     Next
 Next

 'Calculation
 Dim PTerrorism As Double = 15 / 29
 Dim PNotTerrorism As Double = 14 / 29


 Dim CalYes As Double = 1
 Dim CalNo As Double = 1


 For count1 = 0 To wordcounter.Length - 1
     If wordcounter(count1) = 0 Then
         CalYes = CalYes * PTerrNotPresYes(count1)
         CalNo = CalNo * PTerrNotPresNo(count1)
     Else
         CalYes = CalYes * PTerrPresYes(count1)
         CalNo = CalNo * PTerrPresNo(count1)
     End If
```

```
        Next

        CalNo = (CalNo * PNotTerrorism) / totalevidence
        CalYes = (CalYes * PTerrorism) / totalevidence

        If CalNo < CalYes Then
            lblResult.Text = "The tweet is terrorism related"
        Else
            lblResult.Text = "The tweet is not terrorism related"
        End If

    End Sub
    Private Sub btnExt_Click(ByVal sender As System.Object, ByVal e As
System.EventArgs) Handles btnExt.Click
        Me.Close()
    End Sub

    Private Sub btnclear_Click(ByVal sender As System.Object, ByVal e As
System.EventArgs) Handles btnclear.Click
        btnstart.Enabled = True
        btnclear.Enabled = False
        txttweet.Clear()
        lblResult.Text() = " "
    End Sub
End Class
```