Sentiment Mining of Malay Newspaper (SAMNews) Using Artificial Immune System

Mazidah Puteh, Norulhidayah Isa, Sayani Puteh and Nur Amalina Redzuan

***Abstract*—There are sheer volume of rich web resources such as digital newspaper, e-forum, blogs, Facebook and Twitter. Mining the digital text resources may reveal interesting knowledge to respective individuals or organizations. Text mining and sentiment mining or analysis are parts of a new area in sentiment research. Sentiment mining for Malay Newspaper (SAMNews) is constructed based on the artificial immune system called negative selection algorithm which is able to classify the sentiment in newspaper's sentences into the polarity (positive, negative or neutral) intelligently. The sentiment analysis in this project utilized 1000 sentences from newspapers to evaluate the average accuracy. The research used 900 sentences from newspapers as the training data and another 100 as the testing data. The accuracy is achieved at 88.5%. In the future, a comparative study on Artificial Immune System and other techniques or algorithms can be carried out to enhance the performance of the sentiment classification model.**

***Index Terms*— artificial immune system, negative selection algorithm (NSA), text mining, sentiment analysis, sentiment mining, digital text.**

I. Introduction

The advancement of internet technology contributes to the rapid development of knowledge from different part of fields. Hence, there are sheer volume of rich web resources such as digital newspaper, e-forum, blogs, Facebook and Twitter [14]. Mining the text resources may reveal interesting knowledge to respective individuals or organizations. Text mining and sentiment mining are areas in sentiment research [31][6] which have grown as essential methods of knowledge discovery from general and business documents. Currently, one of the active researches on mining is sentiment mining of the textual data.

Sentiment is important in social behavior and it is a way to stimulate cognitive processes in decision making and to carry out strategies in handling situations [9]. Sentiment mining systems are being applied in almost every business and social domain because opinions are central to almost all human activities. For this reason, when we need to make decision, we seek out the opinions of others [5].

Many people like to read, they also like to give comments in many kinds of reading materials such as newspapers, magazines, letters, blogs, and others. Their comments can be considered as bad or sometimes neutral [21]. Experts notice the significances that they can get from the comments. Various analyses have been done in order to investigate about it further.

Recently, many types of analyses have been done to categorize the comments by researchers around the world using personal views on the internet such as Facebook, Twitter, and personal blogs. There are also many kinds of researches about text classification that classify the sentiments, people expression and others. There are many kinds of sentiments and they are hard to analyze because the sentiment have been expressed in symbols such as 'like' and 'unlike' [6]. It is commonly used by people that like to express their emotions or feelings while chatting with their friends. Usually we can see those expressions in social networks likes Facebook, Twitter, MySpace and Friendster.

In sentiment mining system, the researcher can use journals, magazines, newspaper and other kind textual information format to get the sentiment mining data. The aim of the research is to classify the sentiment value in Malay Newspaper using Artificial Immune System (AIS) technique that focuses on Negative Selection Algorithm (NSA).

The organization of the paper is as follows: section II explains related works of sentiment mining and artificial immune system. Section III explains the experiment that has been carried out. Section IV discusses result and section V concludes the findings.

II. RELATED WORKS

A. *Text Mining and Sentiment mining*

Text mining can be loosely described as looking for patterns in a text that contains high-quality information which refers to novelty, relevance and interestingness of the way to write the text. The phrase "text mining" is generally used to denote any system that analyzes large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probably useful information [13]. Normally, the extraction of the information is clearly specified in the text which means it is not hidden. People can easily understand the message or the information from the text that has been written by the author. However, it is

[1] This research is supported in part by Fundamental Research Grant Scheme (FRGS) under the ninth Malaysian Plan (RMK-9), Ministry of Higher Education (MOHE) Malaysia. The grant no is 600-RMI/ST/FRGS 5/3/Fst (208/2010).

F. Mazidah Puteh is with Universiti Teknologi MARA (UiTM). She is now in Computer Science Department, UiTM Terengganu, 23000 Dungun, Terengganu, Malaysia. (corresponding author: phone: 6019-9316631; fax: 609-8400288; e-mail: mazidahputeh@tganu.uitm.edu.my).

S. Norulhidayah Isa is with Universiti Teknologi MARA (UiTM) . She is now in Computer Science Department, UiTM Terengganu, 23000 Dungun, Terengganu, Malaysia. (e-mail: norul955@tganu.uitm.edu.my).

T. Sayani Puteh is with Universiti Kuala Lumpur (UniKL). She is now in Computer Science Department, MIIT, UniKL, 1016 Jalan Sultan Ismail, 50250 Kuala Lumpur, Malaysia (email: sayaniputeh@gmail.com)

4th. Nur Amalina Redzuan is with Universiti Teknologi MARA. She is now in Computer Science Department, UiTM Terengganu, 23000 Dungun, Terengganu, Malaysia.

difficult for a computer to understand and it needs some intelligent ways to make the information be understood.

In order to process this kind of text, it requires the process of structuring the input text such as parsing the text, adding some linguistic features and removing other linguistic features followed by inserting the text into the database.

Sentiment mining is one of the tasks in text mining. The task of sentiment mining is easily described as the classification task that produces the category which represents the sentiment [6]. It is one of the applications of natural language processing, computational linguistics and text analytics that is used to recognize and excerpt some kind of information from the sources [30].

Sentiment mining is a way to enable the computer to recognize and classify the sentiment of some information of what people think automatically [15]. This kind of analysis can make people know whether the comment that has been given by the speaker or the writer is a positive, neutral or negative comment. Comments can consist of judgment or evaluation that is related to the sentiment of a writer during the writing of a document or the sentiment of a respondent when answering a question from an interviewer.

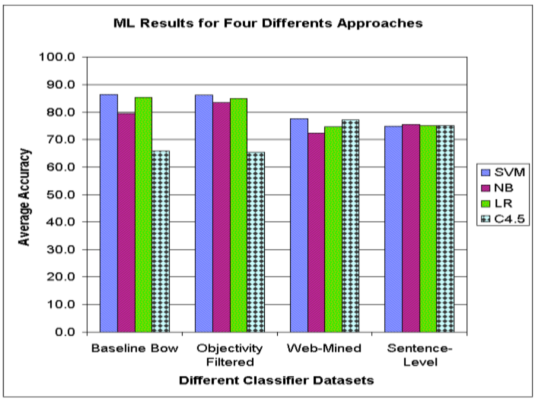
One of the basic tasks in sentiment mining is identifying the polarity of a given existing text in the document, sentence or word. Some studies have used different methods in order to identify the polarity of the existing text [23][1]. Another task that is related to sentiment mining is document-level-sentiment-classification. This task is usually used to classify the whole document that contains opinion as a positive or negative opinion which can be seen in [4]. Other subtask in sentiment mining is classifying messages as opinionated/subjective or factual/objective [1].

B. Machine Learning Algorithm

Many researchers have been using machine learning algorithm to do classification on sentiment mining such as Naïve Bayes and Support Vector Machine (SVM). This is because the performances of the two algorithms are better than other algorithm based on the accuracy [11][29][16].

In the computer science field, SVM is used as a concept to analyze data and recognize pattern that are related to supervised learning methods [22][2].

Fig. 1: Machine Learning Results for four Different Approaches



This kind of machine learning methods has shown a good result in most of the tasks to classify sentiment. Figure 1 shows the comparison of different classifiers that used different kinds of data set.

Preprocessing Technique

Preprocessing is a process to perform a preliminary processing on raw data to prepare it for another processing procedure. It is commonly used as a preliminary sentiment mining practice, data preprocessing transforms the data into a format that will be easily and effectively processed computationally. Data preprocessing consists of the method of cleaning the raw data in order to transform the noisy data into clean ones [10]. Two important techniques that are performed in sentiment mining are stop word removal and stemming.

Stop Word Removal

Stop word removal is the process of removing words that have high frequency which are not important to the sentiment of the sentence. Words such as; ‘a', ‘the', ‘or' are likely to be considered as stop words which have been listed in [7].

There are some stop word studies that have been done in various areas such as sentiment mining, web mining and information retrieval [25][12][13]. The list of stop words that have been identified by the researcher in detecting the sentence-level novelty in Malay words is shown in [3].

Stemming

Stemming or lemmatization is the process of removing the suffixes from a word. In simpler terms, it maps the different “versions” of the word by reducing it to its stem, root or base form. Consider the words; “process”, “processes”, “processing”, “processed”. Such related words may all be grouped into a single root term, “process”, by removing their suffixes.

This preprocessing technique has been used in broad area such as information retrieval that aims to study about how to determine and retrieve a stored information from a corpus, [28]. Although stemming has been studied for English, it also has been utilized in many kinds of languages like Latin [26] and Arabic [18].

One of the algorithms that been used for stemming is called Porter Stemming. Porter stemming is about removing any prefixes, suffixes, or infixes that are contained in the words. [24] For example, the algorithm will remove the prefixes that are usually attached at the beginning of the word likes ‘pre’, suffixes that are attached at the end of the word likes ‘sing’ and infixes that are attached in the middle of the word which came from the word ‘preprocessing’. Table I shows the example of porter algorithm on Malay language.

C. Artificial Immune System

In computer science, Artificial Immune Systems (AIS) is a computational intelligent system inspired by the principles and processes of the vertebrate immune system. The algorithms typically exploit the immune system's characteristics of learning and memory to solve a problem

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE I
AFFIXES REMOVAL USING STEMMING ALGORITHM

Prefix	‘*ber*’, ‘*per*’, ‘*ter*’, ‘*mem*’, ‘*pem*’, ‘*menge*’, ‘*penge*’, ‘*meng*’, ‘*peng*’, ‘*men*’, ‘*pen*’, ‘*me*’, ‘*pe*’, ‘*be*’, ‘*ke*’, ‘*se*’, ‘*te*’, ‘*di*’
Suffix	‘*nya*’, ‘*kan*’, ‘*an*’, ‘*i*’, ‘*kah*’, ‘*lah*’, ‘*pun*’, ‘*ita*’, ‘*man*’, ‘*wan*’, ‘*wati*’, ‘*ku*’, ‘*mu*’
infix	‘el’, ‘er’,’em’, ‘in’,
Prefix and suffix	‘*ber…an*’, ‘*per…an*’, ‘*ter…kan*’, ‘*mem…kan*’, ‘*pem…an*’, ‘*pen…an*’, ‘*men…i*’, ‘*meng…i*’, ‘*menge…kan*’, ‘*penge…an*’, ‘*peng…an*’
Two or more affixes	‘*diper…*’, ‘*…kannya*’, ‘*memper…i*’, ‘*berke…an*’, ‘*men…inya*’, ‘*di…kannya*’

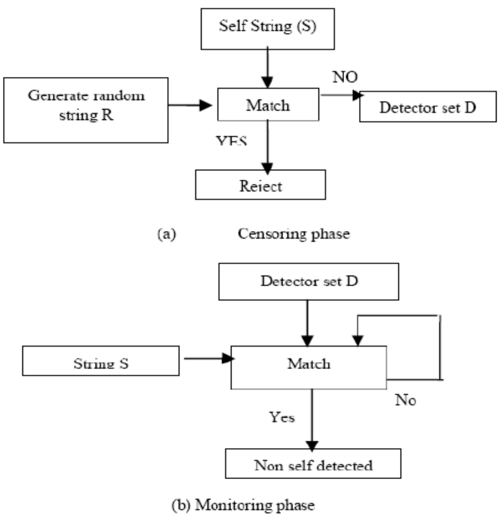
[19][20]. The fundamental concepts of AIS are based on how the lymphocytes which are B-cells and T-cells are matured, adapted, reacted and learn in response to a foreign antigen [8]. There are many models have utilized the idea of immune system in AIS such as negative selection, clonal selection, immune network and danger theory.

Negative Selection Algorithm (NSA)

Negative selection algorithm is used to protect against self-reactive lymphocytes. It has the ability to detect any unknown antigens while not reacting to self-cell. Receptors are made through a pseudo-random genetic rearrangement process during the generation of T-cells in the thymus gland. In the thymus, if there are T-cells that react against self-proteins then it will be destroyed. Only cells that do not drag to self-proteins are allowed to leave the thymus. The T-cells that have matured will circulate all over the body in order to perform immunological functions that will hence protect the body from any foreign antigens [17].

NSA is one of the techniques in AIS. It is a supervised learning algorithm that have been introduced by Forrest et al [27]. This algorithm has been popular in many areas such as computer security, network security and anomalies detection problems [32].

Fig. 2 Negative Selection algorithm [27]



In this algorithm, there are two phases which are censoring and monitoring. In the censoring phase, a set of detector is generated where each of the detectors is a string that does not match any of the protected data. While in the second phase, the protected data will be monitored by comparing them with the existing detector. If a detector is ever activated, a change is known to have occurred. Figure 2 shows the overview of the NSA algorithm that has been used in the system.

III. METHODOLOGY

This research will concentrate on sentiment mining on textual data collected from newspapers. The methodology includes:

A. *Data Preparation*

The activities consisted in this phase is data collection and data representation. This project used data collection from newspapers that were stored in text file document.

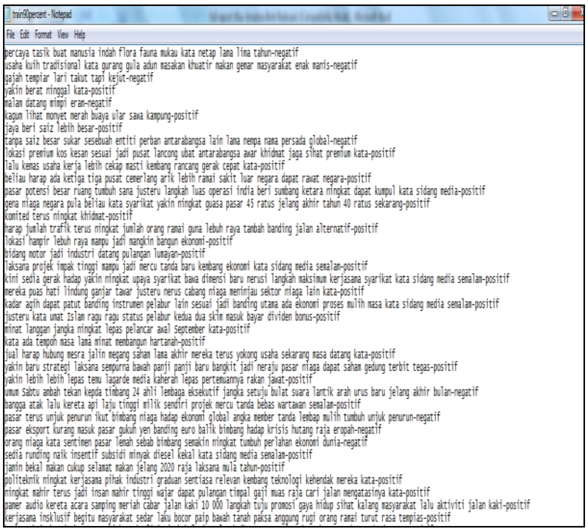

train90percent - Notepad
File Edit Format View Help
percaya tasik buat manusia indah flora fauna mukau kata netap lama lima tahun-negatif
usaha kuih tradisional kata gurang gula adun masakan khuatir makan gemar masyarakat enak manis-negatif
gajah tempiar lari takut tapi kejut-negatif
yakin berat ninggal kata-positif
malam datang mimpi eram-negatif
kagum lihat monyet merah buaya ular sawa kampung-positif
jaya beri saiz lebih besar-positif
tanpa saiz besar sukar sesebuah entiti perban antarabangsa lain lama nempa nama persada global-negatif
lokasi premium kos kesan sesuai jadi pusat lancong ubat antarabangsa awar khidmat jaga sihat premium kata-positif
lalu kemas usaha kerja lebih cekap masti kembang rancang gerak cepat kata-positif
beliau harap ada ketiga tiga pusat cemerlang arik lebih ramai sakit luar negara dapat rawat negara-positif
pasar potensi besar ruang tumbuh sana justeru langkah luas operasi india beri sumbang ketara ningkat dapat kumpul kata sidang media-positif
gena niaga negara pula beliau kata syarikat yakin ningkat guasa pasar 45 ratus jelang akhir tahun 40 ratus sekarang-positif
komited terus ningkat khidmat-positif
harap jumlah trafik terus ningkat jumlah orang ramai guna lebuh raya tambah banding jalan alternatif-positif
lokasi hampir lebuh raya mampu jadi mangkin bangun ekonomi-positif
bidang motor jadi industri datang pulangan lumayan-positif
laksana projek impak tinggi mampu jadi mercu tanda baru kembang ekonomi kata sidang media semalam-positif
kini sedia gerak hadap yakin ningkat upaya syarikat bawa dimensi baru nerusi langkah maksimum kerjasama syarikat kata sidang media semalam-positif
mereka puas hati lindung ganjar tawar justeru nerus cabang niaga meninjau sektor niaga lain kata-positif
kadar agih dapat patut banding instrumen pelabur lain sesuai jadi banding utama ada ekonomi proses mulih masa kata sidang media semalam-positif
justeru kata umat Islam ragu ragu status pelabur kedua dua skim masuk bayar dividen bonus-positif
minat langgan jangka ningkat lepas pelancar awal September kata-positif
kata ada tempoh masa lama minat membangun hartanah-positif
jual harap hubung mesra jalin megang saham lama akhir mereka terus yokong usaha sekarang masa datang kata-positif
yakin baru strategi laksana sempurna bawah panji panji baru bangkit jadi neraju pasar niaga dapat saham gedung terbit tegas-positif
yakin lebih lebih lepas temu lagarde media kaherah lepas pertemuannya rakan jawat-positif
umum Sabtu ambah tekan kepda timbang 24 ahli lembaga eksekutif jangka setuju bulat suara lantik arah urus baru jelang akhir bulan-negatif
bangga atak lalu kereta api laju tinggi milik sendiri projek mercu tanda bebas wartawan semalam-positif
pasar terus unjuk penurun ikut bimbang niaga hadap ekonomi global angka member tanda lembap mulih tumbuh unjuk penurun-negatif
pasar eksport kurang masuk pasar gukuh yen banding euro balik bimbang hadap krisis hutang raja eropah-negatif
orang niaga kata sentimen pasar lemah sebab bimbang semakin ningkat tumbuh perlahan ekonomi dunia-negatif
sedia runding naik insentif subsidi minyak diesel kekal kata sidang media semalam-positif
jamin bekal makan cukup selamat makan jelang 2020 raja laksana mula tahun-positif
politeknik ningkat kerjasama pihak industri graduan sentiasa relevan kembang teknologi kehendak mereka kata-positif
ningkat mahir terus jadi insan mahir tinggi wajar dapat pulangan timpal gaji muas raja cari jalan mengatasinya kata-positif
pamer audio kereta acara samping meriah cabar jalan kaki 10 000 langkah tuju promosi gaya hidup sihat kalang masyarakat lalu aktiviti jalan kaki-positif
kerjasama insklusif begitu masyarakat sedar laku bocor paip bawah tanah paksa anggung rugi orang ramai turut rasa tempias-positif

Fig. 3 Sample of raw data

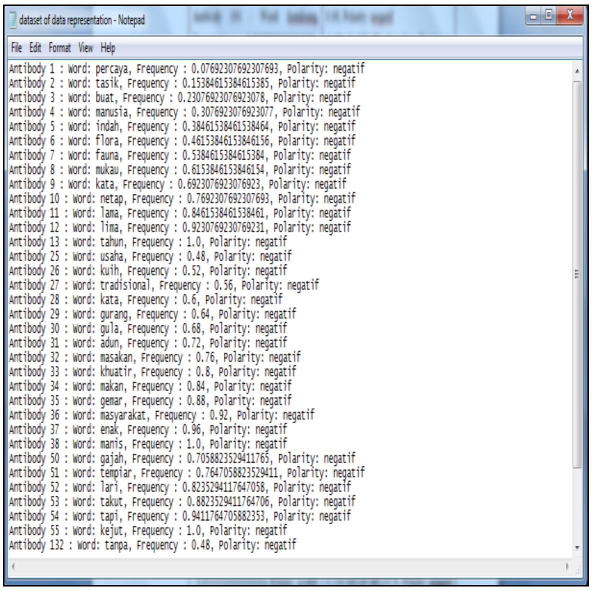

dataset of data representation - Notepad
File Edit Format View Help
Antibody 1 : Word: percaya, Frequency : 0.07692307692307693, Polarity: negatif
Antibody 2 : Word: tasik, Frequency : 0.15384615384615385, Polarity: negatif
Antibody 3 : Word: buat, Frequency : 0.23076923076923078, Polarity: negatif
Antibody 4 : Word: manusia, Frequency : 0.3076923076923077, Polarity: negatif
Antibody 5 : Word: indah, Frequency : 0.38461538461538464, Polarity: negatif
Antibody 6 : Word: flora, Frequency : 0.46153846153846156, Polarity: negatif
Antibody 7 : Word: fauna, Frequency : 0.5384615384615384, Polarity: negatif
Antibody 8 : Word: mukau, Frequency : 0.6153846153846154, Polarity: negatif
Antibody 9 : Word: kata, Frequency : 0.6923076923076923, Polarity: negatif
Antibody 10 : Word: netap, Frequency : 0.7692307692307693, Polarity: negatif
Antibody 11 : Word: lama, Frequency : 0.8461538461538461, Polarity: negatif
Antibody 12 : Word: lima, Frequency : 0.9230769230769231, Polarity: negatif
Antibody 13 : Word: tahun, Frequency : 1.0, Polarity: negatif
Antibody 25 : Word: usaha, Frequency : 0.48, Polarity: negatif
Antibody 26 : Word: kuih, Frequency : 0.52, Polarity: negatif
Antibody 27 : Word: tradisional, Frequency : 0.56, Polarity: negatif
Antibody 28 : Word: kata, Frequency : 0.6, Polarity: negatif
Antibody 29 : Word: gurang, Frequency : 0.64, Polarity: negatif
Antibody 30 : Word: gula, Frequency : 0.68, Polarity: negatif
Antibody 31 : Word: adun, Frequency : 0.72, Polarity: negatif
Antibody 32 : Word: masakan, Frequency : 0.76, Polarity: negatif
Antibody 33 : Word: khuatir, Frequency : 0.8, Polarity: negatif
Antibody 34 : Word: makan, Frequency : 0.84, Polarity: negatif
Antibody 35 : Word: gemar, Frequency : 0.88, Polarity: negatif
Antibody 36 : Word: masyarakat, Frequency : 0.92, Polarity: negatif
Antibody 37 : Word: enak, Frequency : 0.96, Polarity: negatif
Antibody 38 : Word: manis, Frequency : 1.0, Polarity: negatif
Antibody 50 : Word: gajah, Frequency : 0.7058823529411765, Polarity: negatif
Antibody 51 : Word: tempiar, Frequency : 0.7647058823529411, Polarity: negatif
Antibody 52 : Word: lari, Frequency : 0.8235294117647058, Polarity: negatif
Antibody 53 : Word: takut, Frequency : 0.8823529411764706, Polarity: negatif
Antibody 54 : Word: tapi, Frequency : 0.9411764705882353, Polarity: negatif
Antibody 55 : Word: kejut, Frequency : 1.0, Polarity: negatif
Antibody 132 : Word: tanpa, Frequency : 0.48, Polarity: negatif

Fig. 4 Sample of dataset of data representation

The data of this research were collected from the Malaysian newspaper Berita Harian that uses Malay language. The raw data and preprocessed data are shown in Figure 3 and Figure 4. It focused on three (3) topics which are Politics, Natural Disaster and Economy that consists nearly 1000 sentences.

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

B. Preprocessing of Data

In this phase, the research used the technique proposed by [46], as follows:

Stemming

Stemming is the act of trimming the words to their base word. This research will use Reversed Porter algorithm during this process of preprocessing phase.

Stop Word Removal

Stop Word Removal is a process to eliminate any non-sentiment valued words from the data.

Word Tokenizer

Word Tokenizer is a process to make some groups of the same word in each sentence and finds their frequency of occurrences.

C. Reverse Porter Algorithm

The basic idea of this algorithm is to reverse the whole process of Porter algorithm in order to get the result. The main concept of Porter Algorithm is to use a reduction technique where the given words will be reduced to its root form, which can be presented as, w – w' = rW. w is word, w' is suffixes, and rW is root word.

Contradiction to Porter algorithm; the Reverse Porter algorithm is presented as, ArW + Aw' = rS; rS = gW?. ArW is all root words, Aw' is all suffixes, rS are result Word and gW is given word. The process was only done once; the resulting words will be stored in a library so that the engine will not have to repeat all over if it encounters the same word hence reducing computational times greatly.

The first process is adding postfix; the engine will repeat all root words from dictionary and combines the words with all suffixes available. Then, each of the resulting word will be compared with the given word. If there is a match, the engine will return the result. If there is none, the engine will proceed to second sub process which is adding prefixes.

In the second process, the engine will repeat what it has done before with the exception that it will combine all the root words with prefixes instead of suffixes. Then the resulting words will be compared against the given word and if it is a match, the engine will return the result. Otherwise, the engine will proceed to the following sub-process, first letter modification.

In the third process, first letter modification process is unique to Malay language because there are some words that will change when combined with prefixes. For example, word "kering" when combined with prefix "me" will transform into "mengering" instead of a direct combination of both words, "mekering".

Pseudo code of Reverse Porter Algorithm:

- Start: get the word from the input (clean data) (Ex. "memakan").
- Compare with the Library of Generated Word (LGW).
  - False.
    - Add suffixes/prefixes/infixes to the library of Root Words (LRW)(Ex. "me + lari, me+makan")
    - Compare the result with the word again.
    - Repeat until a match is found.
    - Add the matched words to the LGW (Ex, {makan,memakan}).
  - True: return the root word.
- End.

D. Stop Word Removal Algorithm

Stop word removal is a process to find all the words that does not have any sentiment value and remove them from the input. This is because as the research objective is to find sentiment value which is positive or negative from the sentences and if there is no sentiment value word is inserted, it will not bring any effect to the outcome, so it will only increase the time and space costs.

E. Word Tokenizer Algorithm

Word tokenizer algorithm is to discover the occurrence of each word in a sentence. It is a simple method that needs an identifying process of any repeating words and then reducing it to one word with the number of occurance.

Table II and Table III show the data before and after the preprocessing process.

TABLE II
BEFORE PREPROCESSING

Title	Content	Sentiment
Projek RM490j	Ribuan penduduk sekitar daerah Kerian menarik nafas lega apabila Perdana Menteri, Datuk Seri Najib Razak mengumumkan peruntukan RM490 juta bagi mengatasi masalah banjir yang kerap melanda kawasan itu sejak lebih 20 tahun lalu.	Positive
Pas, PR rugi singkir Dr Hassan	KUALA LUMPUR: Bagai retak menanti belah, penyingkiran Datuk Dr Hassan Ali oleh Pas dan Pakatan Rakyat (PR) dianggap satu kerugian bagi kedua-dua pihak, selain membuktikan wujudnya kecelaruan perjuangan serta budaya pentingkan diri selain amalan cantas-mencantas yang sebelum ini dilemparkan terhadap Umno.	Negative

TABLE III
AFTER PREPROCESSING

Title	Content	Sentiment
Projek RM490j	Ribu[1] duduk[1] sekitar[1] tarik[1] nafas[1] lega[1] apabila[1] Perdana[1] Menteri[1], Datuk[1] Seri[1] Najib[1] Razak[1] umum[1] untuk[1] RM490[1] juta[1] bagi[1] atas[1] masalah[1] banjir[1] kerap[1] landa[1] kawasan[1] itu[1] sejak[1] lebih[1] 20[1] tahun[1] lalu[1].	Positive

F. Negative Selection Algorithm (NSA)

After preprocessing, the data is divided into the training and testing dataset. The data is trained using the learning algorithm NSA. The data representation for NSA is constructed into suitable form in a vector space of words as shown below:

Data = {word, frequency, polarity)
Example of data representation;
Word: syukur, Frequency: 0.13333333333333333,
Polarity: positif

The first parameter represents the real word. The second parameter represents the frequency of occurrence of the word in a particular sentence and the last parameter represents the polarity of the sentence that the word belongs to. The data is represented as follows: The sentence "makan coklat sedap suka", has been preprocessed from original sentence, "Kami makan coklat sedap. Kami suka coklat itu". Data representation was;

Word1	{makan,1, positif}
Word2	{coklat,2, positif }

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Word3	{sedap,1, positif }
Word4	{suka,1, positif }

Then, all these data representations are used as the training dataset. The basic idea of the algorithm is to generate a number of detection that is used to identify the self and non-self-data.

1. **Initialize random candidate** of newspaper's data called as antigen
2. **Antigen presentation**: for each antigen, do;
   a) **Compare the word and polarity with antibody**,
      i. if it is false then the antigen will be added as a new antibody (detector) in memory cell
      ii. (match) with the antibody, do affinity measurement (AM)
   b) **Affinity Measurement (AM)**
      Find antibody that match with the word, polarity and frequency of the antigen;
      i. antibody will be skipped
      ii. Otherwise, add the antigen as the new antibody and store into memory cell
1. **Cycle** : Repeat step 1 again

Fig. 5 Steps of Training Process in Negative Selection Algorithm

Figure 5 shows the detail steps in NSA that is used to generate the antibody (detector word) as described below:

1. Define the libraries for positive and negative categories. The words that are clearly defined as positive and classified as positive words will be the library for positive and it will be the same way with negative sentences. The libraries are created so that the keywords or detectors that fit in the positive class will not become the detectors in the negative class. The libraries of both positive and negative categories that have been defined will be matched to the sentences in data training that followed the corpus of Malay word. If a negative word matches one in the positive library, then the word is removed from the sentences. This process will also be implemented with positive sentences where each word in the sentences will be compared with both libraries.

2. Training process continues with training the remaining words in positive and negative sentences so it will become as detectors for positive and negative categories. This detector is called antibody where the first word in the first sentence will be the first detector. After that, the first word is compared to the second word; if they match then the second word is removed and its occurrence is counted; otherwise the second word will be the next detector.
3. This flow will keep running until there are no other antigens that match the memory cell (database of antibody).

IV. RESULT DISCUSSION

This research operated 900 newspaper's sentences in developing the sentiment classification model during training process. This experiment has produced 16, 348 detector words which contains 23, 221 detector words for the positive category and 7, 979 detector words for the negative category.

Positive Library	Negative Library
Antibody 1 : Word: tentu, Frequency : 0.07692307692307693, Polarity: positif Antibody 2 : Word: ratus, Frequency : 0.15384615384615385, Polarity: positif Antibody 3 : Word: isi, Frequency : 0.23076923076923078, Polarity: positif Antibody 4 : Word: rumah, Frequency : 0.3076923076923077, Polarity: positif Antibody 5 : Word: hadap, Frequency : 0.38461538461538464, Polarity: positif Antibody 6 : Word: kena, Frequency : 0.46153846153846156, Polarity: positif Antibody 7 : Word: bil, Frequency : 0.5384615384615384, Polarity: positif Antibody 8 : Word: elektrik, Frequency : 0.6153846153846154, Polarity: positif Antibody 9 : Word: sama, Frequency : 0.6923076923076923, Polarity: positif Antibody 10 : Word: samping, Frequency : 0.7692307692307693, Polarity: positif Antibody 11 : Word: melaksanakan, Frequency : 0.8461538461538461, Polarity: positif Antibody 12 : Word: langkah, Frequency : 0.923076923076923, Polarity: positif Antibody 13 : Word: menjimatkan, Frequency : 1.0, Polarity: positif Antibody 60 : Word: kata, Frequency : 0.14285714285714285, Polarity: positif Antibody 61 : Word: lihat, Frequency : 0.2857142857142857, Polarity: positif Antibody 62 : Word: program, Frequency : 0.42857142857142855, Polarity: positif Antibody 63 : Word: wajar, Frequency : 0.5714285714285714, Polarity: positif Antibody 64 : Word: diperluaskan, Frequency : 0.7142857142857143,	Antibody 14 : Word: salah, Frequency : 0.14285714285714285, Polarity: negatif Antibody 15 : Word: atak, Frequency : 0.2857142857142857, Polarity: negatif Antibody 16 : Word: dapat, Frequency : 0.42857142857142855, Polarity: negatif Antibody 17 : Word: manfaat, Frequency : 0.5714285714285714, Polarity: negatif Antibody 18 : Word: jumlah, Frequency : 0.7142857142857143, Polarity: negatif Antibody 19 : Word: tentu, Frequency :0.8571428571428571, Polarity: negatif Antibody 20 : Word: subsidi, Frequency : 1.0, Polarity: negatif Antibody 21 : Word: cara, Frequency : 0.07692307692307693, Polarity: negatif Antibody 22 : Word: peribadi, Frequency : 0.15384615384615385, Polarity: negatif Antibody 23 : Word: menyifatkan, Frequency : 0.23076923076923078, Polarity: negatif Antibody 24 : Word: fitnah, Frequency : 0.3076923076923077, Polarity: negatif Antibody 25 : Word: laku, Frequency : 0.38461538461538464, Polarity: negatif Antibody 26 : Word: parti, Frequency : 0.46153846153846156, Polarity: negatif Antibody 27 : Word: politik, Frequency : 0.5384615384615384, Polarity: negatif Antibody 28 : Word: hingga, Frequency : 0.6153846153846154, Polarity: negatif Antibody 29 : Word: meresahkan, Frequency : 0.6923076923076923, Polarity: negatif

Fig. 6 Library of Positive and Negative Detectors

Sample of detectors that have been produced can be seen in Figure 6. The performance accuracy of SAMNews is shown in Table IV.

TABLE IV
SAMPLE RESULT OF TESTING FOR ENDING POSITION IN EXPERIMENT III

Total of training and testing Newspaper's sentences	**900:100**
Number of correctly classify(%)	**88.46**
Number of incorrectly classify(%)	**11.54**
Accuracy of testing Newspaper's sentences (%)	**88.46**

V. CONCLUSION

Negative selection algorithm is suitable to categorize the sentences into the sentiments of positive, negative and neutral polarity. The algorithm only recognized and kept words from the newspaper that do not exist in the library yet. The word that existed in the memory cell will be skipped and the process will keep running until there no other words are detected.

One obstacle with this sentiment mining is that the newspaper's data must be in standard language. Some problems will occur in defining an important detector word when the data does not use a standard language. Furthermore, NSA sentiment mining model also needs a clean data to be operated accurately. The strength of the classification model can be enhanced by making some adjustments and improvements.

A comparative study on artificial immune system and other techniques or algorithms is needed to enhance the performance of the sentiment mining classification model.

REFERENCES

[1] A. Abbasi, H. Chen, and A. Salem, (2006). “Sentiment Analysis in Multiple Languages : Feature Selection for Opinion Classification in Web Forums,” ACM Transactions on Information System, vol. 3, pp.2-5.

[2] A. Kennedy, (2006, May). “Sentiment classification of movie reviews using contextual valence shifters,” *Computational Intelligence*, vol. 22, no. 2, pp. 110–125.

[3] A. T. Kwee, F. S. Tsai, and W. Tang, (2009). “Sentence-Level Novelty Detection in English and Malay,” *Nanyang Technology University*, pp. 40–51.

[4] B. Liu, (2010). “Sentiment analysis and subjectivity,” *Handbook of Natural Language Processing*,, pp. 1–38.

[5] B. Liu, (2012). “Sentiment analysis and Opinion Mining,” Morgan & Claypool Publishers.

[6] B. Pang and L. Lee, (2002). “Thumbs up?: sentiment classification using machine learning techniques,” *Conference on Empirical methods*.

[7] C. D. Manning, P. Raghavan, and H. Schütze, (2009). "*An Introduction to Information Retrieval"*, Working Paper on Cambridge UP, pp. 3- 26.

[8] D. Aickelin, U & Dasgupta, (2003). “Artificial Immune System and Their Application,.” in Proceedings of the International Conference on Artificial Immune Systems (ICARIS), pp. 7-10.

[9] D. Consoli (n. d.). “TEXTUAL EMOTIONS RECOGNITION WITH AN INTELLIGENT SOFTWARE OF SENTIMENT ANALYSIS,” *Word Journal Of The International Linguistic Association*, pp. 997–1009.

[10] D. Pyle, (1999). “Data Preprocessing Techniques for Data Mining,” Winter Schools on "Data Mining Techniques and Tools for Knowledge Discovery in Agricultural Datasets”, pp. 139-144.

[11] D. Zhang, S. Li, C. Zhu, and X. Niu, (2010). “A comparison study of multi-class sentiment classification for Chinese reviews,” *Fuzzy Systems and Knowledge Discovery*, pp. 2433–2436.

[12] E. Dragut, F. Fang, P. Sistla, C. Yu, and W. Meng, (2009). “Stop word and related problems in web interface integration,” *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 349–360.

[13] F. Sebastiani, “Machine learning in automated text categorization,” *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, Mar. 2002.

[14] G. Vinodhini and RM. Chandrasekaran, (2012). “Sentiment Analysis and Opinion Mining: A Survey,” *Interpretation Journal Of Advance Research in COmputer Science and Software Engineering*, Vol 2, issues 6, pp282-292.

[15] J. Prager, (2006). “Open-Domain Question–Answering,” *Foundations and Trends® in Information Retrieval*, vol. 1, no. 2, pp. 91–231.

[16] L. Alvim, P. Vilela, E. Motta, and R. L. Milidiú, (n. d.). “Sentiment of Financial News : A Natural Language Processing Approach,” *Learning*, pp. 1–3.

[17] L. N. D. Castro and J. Timmis, (2002). “An Artificial Immune Network for Multimodal Function Optimization,” in Proceedings of IEEE Congres on *Evolutionary Computation*, vol. 1, pp. 699-674.

[18] M. A. H. Omer, (2009). “STEMMING ALGORITHM TO CLASSIFY ARABIC DOCUMENTS,” Symposium on Progress in Information & Communication Technology, pp. 111–115.

[19] M. Puteh, A.R. Hamdan, K.Omar dan A.Abu Bakar 2008. Flexible Immune Network for Mining Heterogeneous Data. LNCS 5132, hlm 232-241 @ Springer-Verlag Berlin Heidelberg.

[20] M. Puteh, A.R. Hamdan, K.Omar dan M.T.H. Mohd Tajuddin 2010. Artificial Immune Network: Classification on Heterogeneous Data, Machine Learning, InTech, ISBN: 978-953-307-033-9

[21] N. Godbole and S. Skiena, (2007). “Large-Scale Sentiment Analysis for News and Blogs,” *Interpretation A Journal Of Bible And Theology*, pp. 765–770.

[22] N. Samsudin and M. Puteh, (2011). “Bess or xbest: Mining the Malaysian online reviews,” in 2011 3rd Conference on *Data Mining and Optimization (DMO)*, pp.38-43.

[23] P. Turney, (2002). “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews,” *Proceedings of the 40th Annual Meeting, pp. 2-5.*

[24] P. Willett, (2006). “The Porter stemming algorithm: then and now,” *Program: electronic library and information systems*, vol. 40, no. 3, pp. 219–223.

[25] R. M. R. Liebregts, (2008, July). “Evaluation of a University-wide Expert Search Engine,”, Thesis for Humanities Department, *Tilburg University*. pp. 70-81.

[26] R. Schinke, M. Greengrass, A. M. Robertson, and P. Willett, (1993). “Journal of Documentation Emerald Article : A STEMMING ALGORITHM FOR LATIN TEXT DATABASES,” *Journal of Documentation, vol. 52, no. 2,* pp. 172-187.

[27] S. Forrest, L. Allen, and A. S. Perelson, (1994). “Self-Nonself Discrimination in a Computer,” *in Proceedings of the IEEE Symposium on Research in Security and Privacy (in press), pp. 3-5.*

[28] T. M. T. Sembok, I. I. H. Uman, and I. N. Rocessing, (2005). “Word Stemming Algorithms and Retrieval Effectiveness in Malay and Arabic Documents Retrieval Systems,” *Cognitive Psychology*, vol. 10, pp. 95–97.

[29] W. Fan, S. Sun, and G. Song, (2011, Apr.). “Sentiment Classification for Chinese Netnews Comments Based on Multiple Classifiers Integration,” *2011 Fourth International Joint Conference on Computational Sciences and Optimization*, pp. 829–834.

[30] W. Wang, (2010.). “Sentiment Analysis of Online Product Reviews with Semi-supervised Topic Sentiment Mixture Model,” *Science*, vol. 5, no. Fskd, pp. 2385–2389.

[31] Y. Mejova, (2009). “Sentiment Analysis : An Overview". Comprehensive Exam Paper, vol. 1, pp. 3-6.

[32] Z. Ji and D. Dasgupta, (2007). “Revisiting Negative Selection Algorithms,” *Evolutionary Computation*, vol. 15, no. 2, pp. 223–251.