Normalization of Common NoisyTerms in Malaysian Online Media

Norlela Samsudin[1], Mazidah Puteh[2],Abdul Razak Hamdan[3] and Mohd Zakree Ahmad Nazri[4]

[1]*Universiti Teknologi MARA Terengganu, norlela@tganu.uitm.edu.my*

[2]*Universiti Teknologi MARA Terengganu, mazidahputeh@tganu.uitm.edu.my*

[3]*Universiti Kebangsaan Malaysia, arh@ftsm.ukm.my*

[4]*Universiti Kebangsaan Malaysia. mzan@ftsm.ukm.my*

ABSTRACT

This paper proposes a normalization technique of noisy terms that occur in Malaysian micro-texts. Noisy terms are common in online messages and influence the results of activities such as text classification and information retrieval. Even though many researchers have study methods to solve this problem, few had looked into the problems using a language other than English. In this study, about 5000 noisy texts were extracted from 15000 documents that were created by the Malaysian. Normalization process was executed using specific translation rules as part or preprocessing steps in opinion mining of movie reviews. The result shows up to 5% improvement in accuracy values of opinion mining.

Keywords: noisy text, text classification, opinion mining, Malaysian online reviews

I. INTRODUCTION

The popularity of Internet has lead to an increase in the number of users using online media such as e-forum, Facebook and Twitter.Individual entry or post in these mediums is also known as micro-text document(Rosa & Ellen, 2009)micro blogging document (Laboreiro et al., 2010) and text language document (Choudhury et al., 2007).One of the characteristics of the contents is innovation of terms used by the author in order to save time and space when delivering information. Nevertheless, the innovation causes difficulties in analyzing these documents due to the use of incorrect abbreviations, incorrect spellings, non standard terminology, missing punctuations and use of slangs. The terms that cause the problem are known as noisy terms.

Knoblock(Knoblock, Lopresti, Roy, & Subramaniam, 2007)defines noisy texts as 'any kind of difference between the surface form of a coded representation of the text and the intended, correct or original text'. The noise causes challenges to normal natural languages processing techniques such as part-of-speech, tagger or parser. In addition, it increases the number of irrelevant features in activities thus confusing statistical analysis tools. Similarly, Yang(Yang, 1995) defines noisy terms as "word combinations that are not meaningful for categorization point of view ... and could cause over fitting and wasteful computation".

The popularity of Internet as medium of communication in Malaysia has increased when the Prime Minister himself uses blog and Facebook to communicate with the nations. It is reported that in June 2009, 64.6% of the Malaysian population uses Internet("Malaysia Internet Usage Stats and Marketing Report,"). That means, a lot of information is available online which may be useful to organizations or their customers. Platforms such as online forums, Twitter and Facebook are used by the Malaysian to express their opinions on a particular item, person or service. Mining these entries may reveal information such as the percentage of people favoringto certain products over the others or percentage of people favoring to a particular political figure.

Nevertheless,noisy terms in these messages need to be processed prior to any text mining or opinion mining activities.The noisy terms may alter the contents of messages thus influence the performance of text classification (Subramaniam, Roy, Faruquie, & Negi, 2009), (Laboreiro et al., 2010), (Dey & Haque, 2009).The use of non-standard abbreviations and invalid sentence formation cause difficulty in understanding the messages. In order to handle the noise in texts, either the system has to be robust or the noise needs to be removed (Subramaniam et al., 2009).

This research analyzes the contents of 15,000 online mediadocuments that were created by Malaysians and suggests the translations for about5000 common noisy terms in this corpus. To the knowledge of the researchers, currently, there is no available dataset on micro-texts that are created by Malaysians. In this project, 5000 e-forum entries, 5000 twitters messagesand 5000 Facebook posts were extracted for academic researches. This dataset may be used to study how the Malaysians communicate which one another in cyber space or to study any hidden patterns in online communication. In addition, at the end of the project the translations of about 5000 common noisy terms in these documents were provided. An experimentconducted using the

translation showed up to 5% improvement in opinion mining activities. This translation may be very useful for text mining activities especially using data from the Malaysian community.

This paper is organized as follows: Section 2 discusses previous works in processing of noisy texts especially in opinion mining projects. Steps which were executed by the researchers in this project are discussed in Section 3. In relation to that, the result and discussion of the experiment are also presented. Lastly, the summary of the main findings and suggestion of future works are presented in Section 4.

II. BACKGROUND

The process of cleaning noisy terms is known as text normalization. Kobuset. al(Kobus, 2008)highlighted three approaches of text normalization i.e spelling checking, translation and speech recognition.Kernighan, Church and Gale (Kernighan, Church, & Gale, 1990) introduced a noisy channel to correct spelling errors. For each word, probability scores of deletion, insertion, substitution and reversal condition were calculated. The word with highest score was considered as the correct spelling substitution. Clark(Clark, 2003)useda similar method to process noisy text with additional involvement of natural language rules. Choudhury et al.(Choudhury et al., 2007) used the translation method and developed a decoder using Hidden Markov Model to suggest a correct term that corresponds to a noisy term. Cook and Stevensonutilized an unsupervised model with about 400 noisy terms as word models.Awet. al(Aw, Zhang, Xiao, & Su, 2006) translated SMS text from English to Chinese with emphasis on the three main characteristics of SMS texts i.e.substitution of words using non-standard acronym, insertion of flavor word, and omission of auxiliary verb and subject pronoun.Kobuset. al(Kobus, 2008) combined the translation and speech recognition approaches and received a better result than using a particular approach only. All of these approaches use dictionaries mainly in English language to aid in the translation process.

Several researchers used different techniques in identifying noisy texts. Laboreiro(Laboreiro et al., 2010) classified each token into 10 categories using SVM method to identify noisy text. Yang (Yang, 1995)use Singular Value Truncated (SVT) method to reduce noisy texts in text categorization using a model that is derived from Linear Least Squares Fit method.

Even though using statistical machine approaches have proven to be successful in identifying noisy terms, a well developed dictionary or a list of correctly spell words is required. Finding a complete combination of English and Malay dictionary was a challenge. Furthermore, as stated by Yang(Yang, 1995), human knowledge to solve the vocabulary gap problem is important. Therefore in this study, human knowledge was utilized to translate noisy terms that exist in the Malaysian micro-text documents.

Noisy term is also known as *bahasarojak* / *bahasa SMS* in Malay literature. In his website, Hussin(Hussin, 2009)highlights seven categories of *bahasarojak* used by Malaysians. Other thant that, a guideline on how to use SMS language in Malay language was published in 2008 after a seminar on the impact of this language toward the Malaysian language. It is one of the main sources of translation rules in this study.

III. METHODOLOGY

A. Data Collection

The first part of the project were extracting online entries and producing a list of noisy terms translations. Figure 1 explains the detail of this process. The step was required since to the researcher knowledge, there is no index that lists the translation of commonly used noisy terms in Malaysian micro-texts.

Online micro-text messages were collected from various source as listed in the following list.

1. 5000 online forum entries which were extracted from several online forum such as http://www.cari.com and http://www.murai.com/. Both online forums are well known in Malaysia and received thousands of entries per day.
2. 5000 Twitter messages which were extracted using keywords such as UMNO, najib, ptptn, PMR, UiTM, UKM and 1Malaysia. These are words that are normally used by the Malaysian only.
3. 5000 Facebook posts which were randomly extracted from entries in the application that were created by the Malaysian.

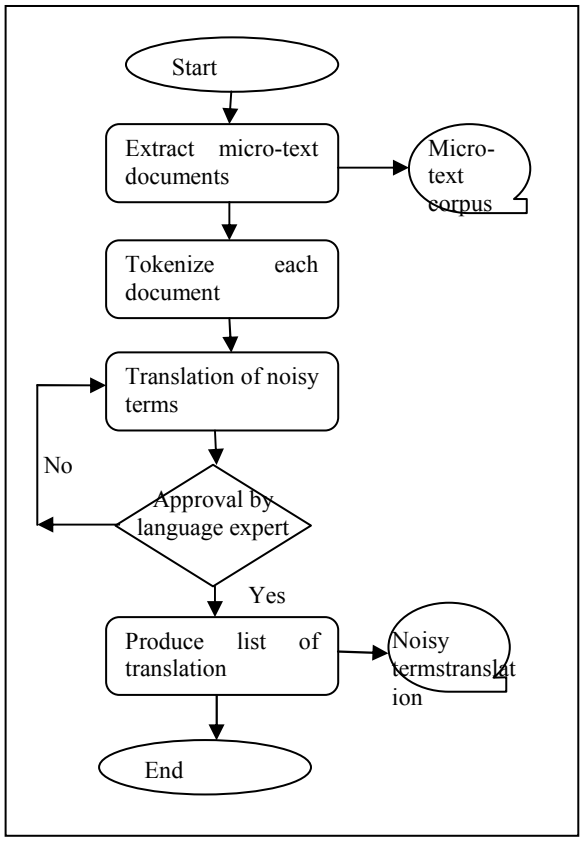

Figure 1: Steps to gather noisy terms translation

Example of entries taken from online forums, Twitter and Facebook messages are listed below.

Example from online forum:

> *"hentahla.....cenelmakinkrg.....hargamakennaik.....cite plaksiaranukgn y keberaparatus da.....hentahhapehape....."*

Example from Twitter message:

> *@krustudios mane anuarzain...=(eh. mane anuarzain?? ManaAnuarZainasdfghjkl?????!ajhsjajhk*

Example from Facebook message:

> *mgunie ad test, mgudpnpon ad...gilaka pa geng2 lecteruruitmnie...x dakanaknkjagaka...bgi test katampasemgu 7 kali baru tau kot...ish!!!!!!!!!!,*

The following characteristics were observed in those messages:

1. The messages are short in nature. On average, each message used less than 30 terms. Among the three platforms, entries from Facebook posts seem to be longer than entries from online forum or twitter messages;
2. The contents of these messages are dominated by noisy terms;
3. The messages were constructed using terms from several languages, mainly theMalay language and English language;
4. The contents hardly follow the language structure of Malay Language or English Language. No clear indication when a sentence starts or ends. Therefore, it is very difficult to identify the components of language syntax such as noun, verb and others.
5. The users used a lot of symbols to express their emotion. Repetition of symbols such as "…." and letters such as"goooooooood morning" are common in these messages.
6. The users also used slang words such as "ler", "moh", "gak" and "lah". These words are unique since different places in Malaysia have their own slang.

Each message was then broken down into smaller unit i.e terms. There were about 32,000 terms in this corpus. About 15,000 terms that exist in three or more documents were selected for the translation process.A term in this context is defined as a group of characters (including punctuation) and numbers that are separated by a blank character or a group of punctuation symbols.

B. Normalization Method

As stated earlier, to the knowledge of the researchers, there is no translation that is created specifically to translate noisy terms which are created by Malaysians. Therefore, before the translation process, the following preliminary activities were carried out:

1. All normal texts with correct spelling in English language or Malay language were removed. Proper nouns such as name of persons, places and brands were also removed.
2. In Twitter messages, words following '@' symbol represent a name of a particular user. Therefore words starts with @ were deleted in Twitter message. Similarly, words following a '#' symbol represent a tag and were deleted.
3. URL links that start with *http://*or *www*...were also deleted.

4. Words which consist of a punctuation symbol or a numberat the beginning or middle of the termwere separated.

Gi2Melaka →gi 2 Melaka,

seriously?dating→ seriously dating

There are several patterns of noisy texts that Malaysians used when writing a micro-text message. These patterns have been explained in detail by several researchers in (Choudhury et al., 2007) and (Pustaka, 2008). Based on these patterns, the following rules were adopted in the translation.

Words that use phonetic spelling: translated to the nearest correct words

cu→ see you,

Hdh2U →hadiah to you

1) **Words with numbers**: A term ended with number 2, which means repetition of the words are spelled correct. Otherwise, the number will be separated from the word.

anak2→anak-anak

5hari →5 hari

2) **Words with slang:** Translate to words with similar meaning

teme→teman, makanang→makanan

pebende→apabenda gua→saya

3) **Use the first character only:** Translate to the proper word.

d→ di, m → meter

u→ you b → be

4) **Use the first character in a phrase:** Translate to the proper word.

sk→salinankepada

al→anaklelaki

5) **Use the first few words:**Translate to the proper term.

no→nombor

tel→telefon

6) **Words that use the consonant only**: Translate to the proper word.

pd→pada,

spt→seperti

7) **Words that use the first and last character**: Translate to the proper word.

yg→ yang,

gram→ gm

8) **Words that use the last syllables:** Translate to the proper word.

mak→emak

dah→sudah

9) **150 translation of noisy terms recommended in (Pustaka, 2008)**: Included in the list even though the term is not the top 5000 noisy terms discovered in the corpus.

ssh→susah

mb→megabait

10) **Abbreviations:**keep if the abbreviation is correct based on reference in (Pustaka, 2008). Translate otherwise.

askm→assalamualaikum

np→kenapa

11) **Smileys :**Translate into its appropriate meaning.

:) → (senyum)

:-o →(terkejut)

12) **Letter repetition:** remove where appropriate. Limit to 3 repetition if the term expresses feeling.

Hehehehehehehe→hehehe

Mmmmmh→mmmh

Ooooooooooooh→oooh

13) **Ambiguous terms:** no translation is made.

ct→Siti,Cutior city.

cube→ cube (English) or cuba (Malay)

Each noisy term was translated by two people based on former experience, discussion with friends and references to (*Kamus Dewan Edisi Ke 4*, 2010) and(*Kamus Dwibahasa Bahasa Inggeris Bahasa Melayu, Edisi 2*, 2002). A third person suggested the translation when conflicts rose. A language expert then checked the validity of all 5000 translations.

C. Experiment

In order to check the impact of the new noisy terms translation, an opinion mining process was conducted using 1000 positive and1000 negative online movie reviews that were created by Malaysians. These reviews were extracted from various forums, Facebook entries and weblogs.The reviews were given to three people,

which identified the category for each reviews either a positive reviewor a negative review.

The researchers believed that high volume of noisy terms in micro-text messages contributed to the classification result. Therefore the translation of noisy texts and removing of stop words were necessary prior to opinion mining process as stated in Figure 2.

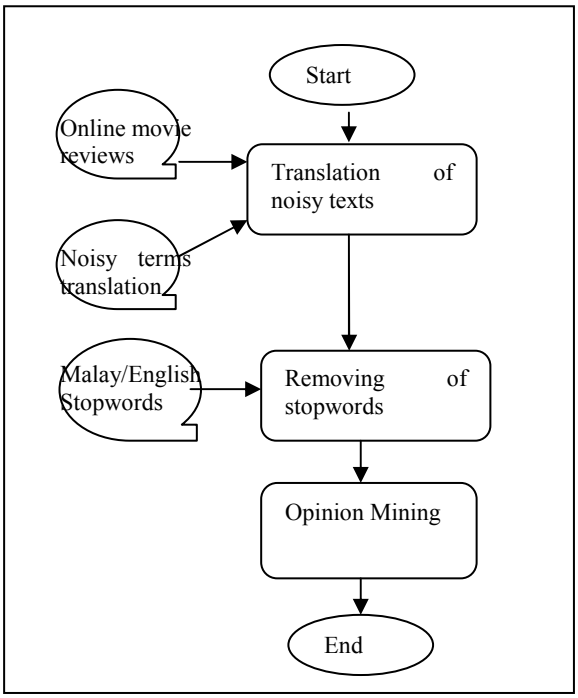

Figure 2: Text classification process

Other than translation of noisy text, removing of stop word is another preprocessing step that is required in text classification. This process reduces the number of features by deleting terms that do not carry any information such as *a, is, or* (English), *kerana*, *dan,* and*atau* (Malay language). Since the researchers have not found any list of stop words that is written in English and Malay language, the list is created from the following sources.

- The English stop word list taken from Armand Dbrahaj website (Brahaj, 2009). The list is selected since the number of words is moderate and he has used it successfully in his project.
- The Malay stop word list was adopted from a list used by (Kwee, Tsai, & Tang, 2009). Nevertheless a few terms that are related to opinion such as '*tidak*' are excluded from the list.

In total, there are 540 stop words in this list.

The opinion mining process was conducted using WEKAsoftware. Three most popular classification models are selected for the experiment i.e k nearest neighbor (IB1 in WEKA), Naïve BayesMultinomial(NB) and Support Vector Machine (SMO in WEKA). Each experiment is conducted using 10 folds cross validation. The accuracy value is used to measure the impact of opinion mining with preprocessing and without pre processing step. The value is calculated using a formula as stated in Equation (1).

$$Accuracy = \frac{number\ of\ correct\ classification}{number\ of\ all\ document} \quad (1)$$

Table 1 shows the values obtained from the experiment without the pre-processing process and with the pre-processing process.

Table 1: Result of experiment

	IB1	**NB**	**SMO**
Without Preprocessing	62.35	80.95	79.55
With Preprocessing	66.60	85.55	82.85
Increment	4.25	4.60	3.30

IV. EVALUATION/DISCUSSION

The result of the experiment shows that the preprocessing activities contribute to the improvement in the accuracies of text mining using kNN, NB and SVM models. All of them used different approaches in identifying the categories of a particular document.

In the training phase of kNN, the training data is indexed using a specific technique that measures the similarity among the documents. In the testing phase, the top k ranking is selected and used to compute the score. Since WEKA use 1 as the k value, the identification of classification is determined by the nearest index. The noisy terms translation had reducedthe number of terms thus improving the similarity values among the documents.

Multinomial Naïve Bayes is a probabilistic learning method. It measures the probability of a particular document being in a particular class based on the frequency of terms that exist in this document. Normalizing the noisy terms increase the number of related terms for a particular class thus increasing the accuracy value of text classification process.

In SMO approach, a hyper plane is created to separate two groups of classes in a vector space. Normalization of noisy texts changes the scores of the documents, thus influencing the margin between both classes. Consequently for significant impact, changes of scores for high number of documents are required. As a result, the percentage of improvement in SVM is less than the percentage of improvement in Naïve Bayes and kNN methods.

V. CONCLUSION

This study shows that high existence of noisy terms do influence the result of text classification of micro-text documents. Human translation approach was used as part of the normalization process. The translations forabout 5000 common noisy terms were introduced since currently there is no suitable dictionary to solve the problem. Opinion mining of Malaysian micro-texts using the noisy terms translation shows improvement in accuracy values for three main text classification models.

Nevertheless, the translation ignores ambiguous terms such as *ct* and *cube*. The proper translations for these words depend on the context where the terms are used. This is an area that needs further analysis in the future. Other than that, suitable technique to automatically translate noisy texts that exist in Malaysians micro-texts will be studied in the future.

REFERENCES

Aw, A., Zhang, M., Xiao, J., & Su, J. (2006). *A phrase-based statistical model for SMS text normalization*. Paper presented at the Proceedings of the COLING/ACL on Main conference poster sessions, Sydney, Australia.

Brahaj, A. (2009). List of English Stop Words, from http://armandbrahaj.blog.al/2009/04/14/list-of-english-stop-words/

Choudhury, M., Saraf, R., Jain, V., Mukherjee, A., Sarkar, S., & Basu, A. (2007). Investigation and modeling of the structure of texting language. *International Journal on Document Analysis and Recognition, 10*(3), 157-174. doi: 10.1007/s10032-007-0054-0

Clark, A. (2003). *Pre-processing very noisy text*. Paper presented at the Proceeding of Workshop Shallow Processing at large Corpora, Lancester.

Dey, L., & Haque, S. K. M. (2009). *Studying the effects of noisy text on text mining applications*. Paper presented at the Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data, Barcelona, Spain.

Hussin, S. (2009). Bahasa SMS Retrieved 27/1/2012, from http://supyanhussin.wordpress.com/2009/07/11/bahasa-sms/

Kamus Dewan Edisi Ke 4. (2010). Kuala Lumpur, Malaysia: Dewan Bahasa Pustaka.

Kamus Dwibahasa Bahasa Inggeris Bahasa Melayu, Edisi 2. (2002). Kuala Lumpur, Malaysia: Dewan Bahasa Pustaka.

Kernighan, M. D., Church, K. W., & Gale, W. A. (1990). *A spelling correction program based on a noisy channel model*. Paper presented at the Proceedings of the 13th conference on Computational linguistics - Volume 2, Helsinki, Finland.

Knoblock, C., Lopresti, D., Roy, S., & Subramaniam, L. (2007). Special issue on noisy text analytics. *International Journal on Document Analysis and Recognition, 10*(3), 127-128. doi: 10.1007/s10032-007-0058-9

Kobus, C., Yvon, F. & Damnati, G. (2008). *Normalizing SMS: are two methaphors better than one?* Paper presented at the 22nd International Conference on Computational Linguistics (Coling 2008), Manchester.

Kwee, A., Tsai, F., & Tang, W. (2009). Sentence-Level Novelty Detection in English and Malay

Advances in Knowledge Discovery and Data Mining. In T. Theeramunkong, B. Kijsirikul, N. Cercone & T.-B. Ho (Eds.), (Vol. 5476, pp. 40-51): Springer Berlin / Heidelberg.

Laboreiro, G., Lu, #237, Sarmento, s., Teixeira, J., Eug, . . . Oliveira, n. (2010). *Tokenizing micro-blogging messages using a text classification approach*. Paper presented at the Proceedings of the fourth workshop on Analytics for noisy unstructured text data, Toronto, ON, Canada.

. Malaysia Internet Usage Stats and Marketing Report. http://www.internetworldstats.com/asia/my.htm

Pustaka, D. B. d. (2008). Panduan Singkatan Khidmat Pesanan Ringkas. Retrieved from http://www.dbp.gov.my/khidmatsms.pdf

Rosa, K. D., & Ellen, J. (2009). Text Classification Methodologies Applied to Micro-Text in Military Chat. 710-714. doi: 10.1109/icmla.2009.49

Subramaniam, L. V., Roy, S., Faruquie, T. A., & Negi, S. (2009). *A survey of types of text noise and techniques to handle noisy text*. Paper presented at the Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data, Barcelona, Spain.

Yang, Y. (1995). *Noise reduction in a statistical approach to text categorization*. Paper presented at the Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Washington, United States.