Journal of Engineering and Applied Sciences 12 (3): 468-474, 2017
ISSN: 1816-949X

Construction of Malay Abbreviation Corpus Based on Social Media Data

Nasiroh Omar, Ahmad Farhan Hamsani, Nur Atiqah Sia Abdullah and Siti Zaleha Zainal Abidin
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA,
Shah Alam, Selangor, Malaysia

Abstract: This study describes a construction of Malay abbreviation corpus by extracting and normalizing selected social media data with multilayer filtration pattern matching technique along with statistical machine translation approach. In this study, one million Malay Lingo user-generated-posts via Twitter and Facebook are extracted for sampling. Each word will undergo pre-processing stage which involves filtration and association and stored in MySQL database table. Then, each word in the corpus is linked with its respective word in existing vocabulary; otherwise, it is considered as abbreviation word will be further processed by using N-grams approach and added to the existing corpus. Based on the result, it can be seen that the longer the length of text, the translation probability is decreased. Furthermore, the style of writing is very important. The lack of space usage to separate in between words will cause more than one word are merged and became out-of-vocabulary word. The worst case is the strange merged word has no link to any recognizable root word in the dictionary. In the first attempt of processing 1000 selected posts from the social media, a lot of uncommon abbreviation words are found. As a result, a lower translation percentage is achieved. Nevertheless when the post uses common abbreviations that exist in the Malay Social Media Corpus then the result of the translation is able to achieve 100% accuracy. Nevertheless, the source of user-generated word is infinite and there is still many ways to improve the combination of NLP techniques in constructing a better and reliable corpus due to the dynamic nature of user's behaviour and their informal ways of writing texts. The corpus is very much needed in analysing public's sentiments in various dimensions such as product-related evaluations and service-oriented feedbacks which are propagated across various platforms of social media.

Key words: Social media, user generated content, abbreviation, corpus, big data

INTRODUCTION

Social media has become one of the most important medium for people to communicate around the world since the existence of Facebook (was founded in 2004) and Twitter (was founded in 2006). People are more to post about their daily life, opinion, feeling or emotion across various social media platforms. It is a norm for people to post their contents several times per day which could influence trust among people and promote risk on individual behaviour while sharing all sorts of information (Wang *et al*., 2016). There are about 55 million status updates every day on Facebook by over 500 million users (Branckaute, 2010). Meanwhile, for Twitter, 50 million tweets per day are made and the number of people communicate via the media is growing tremendously (Twitter, 2011).

The social media data has attractive nature to be considered as favourite dataset for Natural Language Processing (NLP). The data comes in the form of rich text that is unlimited and can be obtained at real-time as well as real-life people's conversation. Nevertheless, for most of the time, the magnanimous data consists of text in non-formal format and hinders data scientists to apply NLP techniques and use them for machine translation application. The data contains excessive noisy texts or out-of-vocabulary words that need to be cleaned before applying related NLP techniques. The massive number of noisy texts poses various challenges to identify the correct out-of-vocabulary words associated with its multiple abbreviations.

Malay language is the tenth most spoken languages of the world. Many studies of text normalization based on English language but yet there is very little research on other languages such as Malay. The scarcity is mainly due to the incomplete or unavailability of digital resources for the language. Inspired by the scarcity this research will make use of unlimited texts produced and construct Malay abbreviation corpus based on the user-generated contents.

Prior to social media, Short Message Service (SMS) was one the dataset source for NLP studies. For example, SMS messages written in English and Mandarin Chinese via public SMS were collected and added to a

Corresponding Author: Nasiroh Omar, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia

corpus in the form of XML and SQL dumps. The use of abbreviations was rampant in SMS texts as published in formal Malaysian standard English newspaper (Joseph *et al*., 2013). A Malay Mixed text Normalization Approach (MyTNA) and text-selection technique was created via machine learning approach (Samsudin *et al*., 2012). The approach was applied on a mixed languages online messages which is not suitable for data mining purpose. MyTNA normalizes noisy texts and the machine learning approach by selecting only relevant features.

Normalizing informal and unstructured texts is one of the prominent steps in dealing with construction corpus of spoken or written words. For example, a personalized chat normalizer for English is constructed and integrated within a multilingual chat system (Aw and Lee, 2012). The challenge in normalizing of non-standard and self-created abbreviation through machine translation technology is normally due to the lack of training data (Jehl, 2010) and variation of abbreviations used. The research to classify variation of abbreviations in Malay words (Kayte and Mundada, 2016) proposes four clusters; parentheses, common formats, definition and specific titles.

The research on machine translation (Bali *et al.,* 2007) shows that statistical machine translation has improved normalization performance. Informal data is used as source language and standard english as target language. Nevertheless, the research has revealed one flaw that is the word can only be normalized if it exists in the training data set. Thus, a linked database to a machine translation system is able to improve the accuracy of the translation results (Aw *et al*., 2006). In addition, understanding the inherit characteristics of noisy texts (Kobus *et al*., 2008) in the form of character deletion, substitution of phonetic, typing errors, merging of word, dropping of word and capitalization/de-capitalization are able to treat noisy text normalization problem in the machine translation process.

Various related researches have been analysed in order to extract pertinent processes involved, challenges faced and potential solutions sought. For instance, 5000 English SMS were collected and translated into phrase-based by using statistical machine translation (Bali *et al*., 2007). Nevertheless, a lesson learnt from the work that it is not suitable since the semantic of the words or sentence is not the concern of this research. Another research collects 3000 French sms and analyses by using multiple approaches to normalize noisy text (Aw *et al*., 2006). Their combination of pre-processing technique and statistical machine translation show improvement in the performance of getting the right abbreviations. However, in the pre-processing part, multiple filtrations (Kayte and Mundada, 2016) retain out-of-vocabulary words. On the other hand, combination of machine translation and natural language processing techniques (Gadde *et al*., 2011) in converting Twitter words into standard english has improved the translation performance. The earliest research on normalizing Malay words creates a dictionary-based system called Noisy term based on data from Malaysian online media. Research shows that the use of stemmer in Malay internet lingo writing style (Cook and Stevenson, 2009; Han *et al*., 2012) causes a loss of original term.

The next study presents the methodology of this research while taking precaution of what must be included and what approach should be avoided in order to follow best practices in constructing the intended abbreviation corpus.

MATERIALS AND METHODS

This study describes the whole process of constructing the Malay abbreviation corpus. The construction starts with provisioning in-vocabulary words, followed by collecting user-generated content from social media, filtering the content and lastly, linking the content with the words designated as out-of-vocabulary and in-vocabulary.

Construction of the Malay corpus starts by pre-processing stage which involves data gathering, extraction, filtration and association. At the end of the pre-processing stage, a Malay corpus is created and stored in MySQL database table. MySQL is chosen due to its popularity as open-source relational database management system. MySQL supports full-text indexing which is more than enough to perform tasks required for the corpus and is very stable (Samsudin *et al*., 2012). In addition, this research implements the Malay abbreviation corpus as Basri *et al*. (2012). Therefore, four steps are performed as follows: provisioning in-vocabulary words, collecting user generated content from Social Media, filtering the content and link the content with out-of-vocabulary and in-vocabulary words.

Provisioning in-vocabulary words: The first step of building a corpus involves provisioning as many formal Malay words as possible to vocabulary collection. This step will be used in later stage to differentiate words in social media posts between in-vocabulary and out-of-vocabulary words. The in-vocabulary words are taken from Bahasa Wordnet (Musa *et al*., 2011) which contains 64,431 unique words and stored into a database table. The collected words are checked against the main source of data from Dewan Bahasa dan Pustaka which is a government body that in charge of coordinating Malay language usage in Malaysia as well as Brunei.

In addition, more words are taken from Wiktionary. Wiktionary is a collaborative constructed resource mainly by non-professional volunteers on the web and also known as Collaborative Knowledge Base (CKB) (MySQL, 2001). Extracting data from CKB requires suitable programming access mechanism. As the data is in Web with HyperText Markup Language (HTML) format, a simple HTML DOM (Document Object Model) parser (Zesch *et al*., 2008) is used to extract all the words. The parser returns the DOM tree of the webpage in an object to be traversed in order to retrieve the intended data. All stored words are in lowercase as uppercase does not have orthographic value in Tweets (Basri *et al*., 2012) shows the snippet of the return object.

To ensure the data consistency, Wordnet database is used before storing the word from Wiktionary to the database. Data from Wiktionary were tagged with root word or affix and the affix is linked to its root word in a tree-type data structure.

Then, all the available in-vocabulary words from Wordnet are gathered and stored in vocabulary database table. Before adding in-vocabulary words from Wiktionary to the vocabulary database, each word will be checked against the existing data in the database. The next step proceeds by collecting User-Generated-Content (UGC) from selected social media. Figure 1 shows the process of provisioning in-vocabulary words into vocabulary collection.

Collecting UGC from social media: In this study, one million user-generated-posts via Twitter and Facebook are extracted for sampling and referred to as Malay Social Media Corpus (MSMC). The sampling data has the highest priority among the compiling criteria in achieving the intended representativeness (Chen and Kan, 2013). The sampling phase is carried out in three stages: defining population of target, sampling frame and data gathering in accordance to the chosen sampling technique (Basri *et al*., 2012). The target population is defined as a Malay lingo or Malay abbreviation. These samples include 100 mixed Twitter and Facebook user IDs and 50 hash-tag topics to cover many different styles of abbreviations in Malay language. Purposive sampling techniques are applied in making decision based on population of interest (Smith, 1976). Thus, the posts do not include user-generated-content that contains only formal Malay language or mixed language.

All user-generated-content are collected via respective social media Application Programming Interface (API). Most of the texts which are collected from the posts are in a form of abbreviations. The collected texts are tallied with the categorization by Basri *et al*. (2012). Table 1 shows part of the abbreviation patterns.

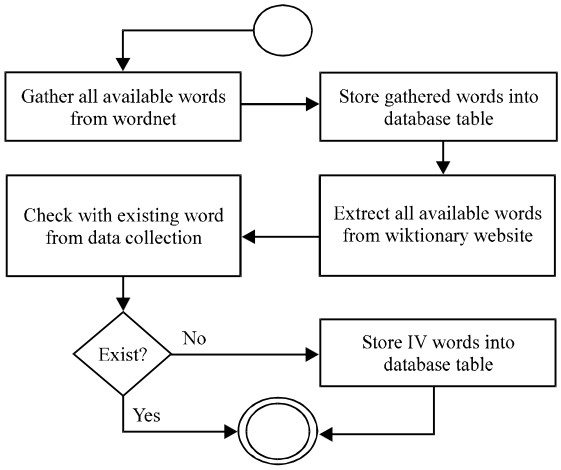

Fig. 1: Process of provisioning In-Vocabulary (IV) words

Abb. 1 and 2 in Table 1 have obvious and predictable properties or pattern that are negation and reduplication. The word 'tidak' is normally used before the verb or adjective to produce a negative sentence. Reduplication can produce new meanings from the original word or to express grammatical functions. In addition, abbreviation can be identified by the use of parentheses and common format (Kayte and Mundada 2016). Identifying common format is based on rules that if sequence of letters is separated by full-stop, sequence of letter with two three or four capital letters, sequence of consonants in upper case and sequence of vowel in upper case can be considered as abbreviation as well. This identification pattern of abbreviation is used during the filtration of user-generated-content.

Filtering UGC: In this phase, the vocabulary collection contains words taken from Wordnet, Wiktionary and User-Generated-Content (UGC) from Twitter and Facebook (Algorithm 1 and 2). The user-generated-content is still in its original and raw form. Therefore, the content consists of scattered unwanted texts (such as hyperlink, user handle (@username) and topic hash-tag (#topic)) and in-vocabulary words. In this study, only abbreviation words are needed and the rest of other words will be ignored. Table 2 shows the type of texts which are filtered.

Snippet of wiktionary return object:

```
[0] ⇒ simple-html-dom-node object
    (
        [nodetype] ⇒ 1
        [tag] ⇒ Array
    (
        [href] ⇒ /w/index. Php? tile = aba&action = edit&redlink = 1
          [class]⇒ new
          [title]⇒ aba (tidak wujud)
    )
```

Table 1: Abbreviation patterns

No. of Abb.	Type of abbreviation	Normal phrase	Example	Meaning
1	Replace tidak with the letter x	Tidak boleh	X boleh	Cannot
2	Alter reduplication	Hari-hari	Hari2	Everyday
3	Eliminate vowel letters	Bangun	Bgn	Get up
4	Eliminate the letter r	Terserang	Terseang	Attacked
5	Eliminate affix	Kekasih	Kasih	Sweetheart
6	Eliminate initial letter	Itu	Tu	That
7	Eliminate last letter	Tidur	Tidu	Sleep
8	Combine words	Macam mana	Camne	How

Table 2: Filtered texts

No. FT	Type of filtered text	Example
1	@ + username	@raditzfarhan
2	# + topic	#AyatRejectKejam
3	Hyperlink	https://www.google.com
4	Special characters	! $ & * / ; < ? {
5	In-vocabulary words	Serang (attack)

Algorithm for translating the word:

Get the input text from the from
Connect to MYSQL database
Perform a full-text searching on MSMC
If a match is found
    Search for IV link
If no match
    Return the same text
Repeat for each word in user input
Display result

The filtrations are carried out in three layers. The first layer involves the filtration of type FT1, FT2 and FT3 as defined in Table 2. The second filtration layer removes type FT4 which are the special characters. Lastly, the third layer removes type FT5 with the help of vocabulary collection. The filtrations processes are carried out in three sequential steps so that the filtered texts contain only the out-of-vocabulary words. The left over words are treated as abbreviation words.

Layer 1 filtration: At this layer one filtration, the user filters texts start with at-sign (@) followed by user name, topics with hash-tag (#) and hyperlink with "http://". A pattern matching technique which is called regular expression (regex) (Smith, 1976), particularly Perl Compatible Regular Expressions (PCRE) is used. The following regex pattern to match the text strings is used: "/^(@|#)([a-zA-Z0-9_]+$)/" and "/^(https?:\/\/)?([\da-z\.-]+)\.([a-z\.]{2,6})([\/\w \.-]*)*\/?$/".

Layer 2 filtration: This layer filtration removes special characters. Using the same method as in layer one, the special characters are filtered using the regex pattern to match special characters: "/[^a-zA-Z0-9]/".

Layer 3 filtration: At this layer filtration, texts match each word in the User-Generated-Content (UGC) that went through layer one and two filtrations with the word in the vocabulary collection. Any matched string is considered as in-vocabulary and the rest considered as out-of-vocabulary words which has higher probability of belong to abbreviation collection.

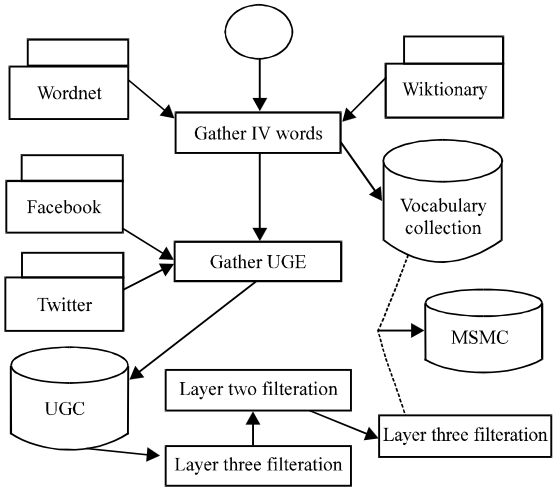

Fig. 2: Process of constructing Malay social media corpus

After went through all the filtration layers, the Out-of-Vocabulary (OOV) words are stored in a database to form the Malay Social Media Corpus. The next study associates the filtered words with In-Vocabulary (IV) words.

Linking OOV with IV: The final step in the construction of Malay Social Media Corpus is to link or associate each out-of-vocabulary words or abbreviation to its respective in-vocabulary word. An intermediate database table is used in order to link between the vocabulary collection and the corpus tables. The links are essential elements for identifying the formal form of Out-of-Vocabulary (OOV) words. There is also a possibility that one OOV word could be linked to more than one In-Vocabulary (IV) word. For instance, the abbreviation 'msk' could mean 'masak' (cook) or 'masuk' (enter). Such ambiguity can be solved using N-grams approach that will be covered in the next section.

Figure 2 shows the overall linking process between the Out-of-Vocabulary (OOV) and In-Vocabulary (IV) words. The first step gathers in-vocabulary words from

Internet which is specifically from Wordnet and Wiktionary. Before storing the in-vocabulary words, redundancy checking between both sources is carried out to ensure just a single in-vocabulary word exists in the data collection. Then, IV words are stored in a single database table.

The association process is carried out in two ways; automatically and manually. Some of the abbreviations are detected automatically when there matched the patterns as described in Table 2 or by parentheses and common abbreviation format. The words that cannot be identified and linked will be connected manually to its root word. If the root word does not exist, the word is added to the vocabulary collection at the same time.

RESULTS AND DISCUSSION

This study explains how to use the Malay Social Media Corpus (MSMC) that has been constructed. In this study, MSMC is used with a web-based machine translation application to show how normalization of an input text is done and return the expected output. The algorithm for translating the word is defined as in.

As explained in the previous study, one out-of-vocabulary word might have linked to >1 in-vocabulary word. In order to determine which in-vocabulary word to be used, a probabilistic N-gram language model is used. In this case, uni-gram model is used where n = 1 due to compare only one term or word. Usually, the whole vocabulary collection will be generated with a Maximum-Likelihood (ML) estimation model and the collection model is linearly interpolated with a maximum-likelihood model for every document and a smoothed document model will be created (Aho, 1980; Neubig, 2012). Maximum-likelihood estimation equation is shown as below to get PML (wi) value where w is each word in the sentence:

$$P_{ML}\left(w_i \middle| w_1 \ldots w_{i-1}\right) = \frac{c\left(w_1 \ldots w_i\right)}{c\left(w_1 \ldots w_{i-1}\right)} \quad (1)$$

where, N = Total vocabulary size; $\lambda_1 = 1-\lambda_{unknown}$; $\lambda_{unknown}$ = Probability of unknown word. Figure 3 shows the overall process of normalization in this study. Figure 4 shows a screenshot of the web-based application created to demonstrate the translation. The graphical user interface is a simple website that has a text area where a user can input the text. The user can enter normal text in the text area with the use of short forms and abbreviations. For this data entry, a text limit is enforced which is up to 200 characters including spaces.

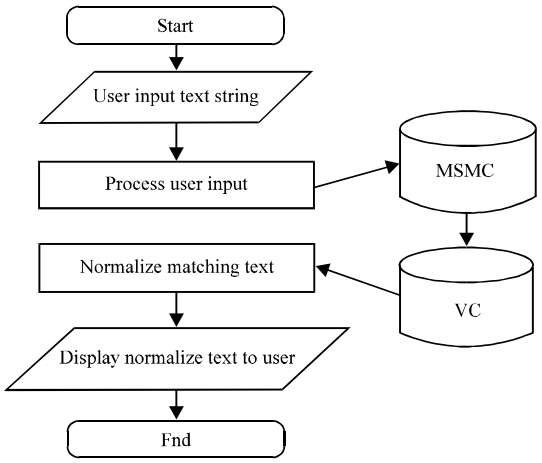

Fig. 3: Normalization process

Fig. 4: Web-based application interface

When the user hits "Translate" button, the application returns the normalize text. Each abbreviation or short form will be translated to its formal words.

Table 3 shows a sample of translation from social media posts with different length of texts. The sample shows the accurateness of the text translation in relation to its formal words. Based on the result, it can be seen that the longer the length of text, the translation probability is decreased. Furthermore, the style of writing is very important. The lack of space usage to separate in between words (e.g., bwhmeja) causes more than one word are merged and became out-of-vocabulary. The worst case is the strange merged word has no link to any recognizable root word. In the first attempt of processing 1000 selected posts from the social media, a lot of uncommon abbreviation words are found. As a result, a lower translation percentage is achieved. Nevertheless when the post uses common abbreviations that exist in the Malay Social Media Corpus then the result of the translation is able to achieve 100% accuracy.

Table 3: Sample inputs with results

Original text	No. of characters	Translated text	Translation percentage
@afdlinshauki alhamdulillah semuanya dr berkat allah	57	@afdlinshauki alhamdulillah semuanya dari berkat allah	100
Jauh sgt percutian ini Ke negeri di bawah bayu As tourist guide, Saya sedia berkhidmat. Visit my	102	Jauh sgt percutian ini Ke negeri di bawah bayu As tourist guide, Saya sedia berkhidmat. Visit my	100
Hapy birhtday k, smoge kowang d'pnjang kn umor n d'mura kn rzeki aw, ika, cuwie td aw, cian ko kn, heee, tp yg sal frog bwhmeja tuwh ak x tau papew, heee tp bezs gak kne kn ko td...^_^...so, jgn buli ak ag aw...heee	220	Hapy birhtday okay, semoga kowang d'pnjang kan umur dan d'mura kan rezeki aw, ika, cuwie tadi aw, kasihan kau kan, heee, tapi yang pasal frog bwhmeja tuwh aku tidak tahu papew, heee tapi bezs juga kena kan kau tadi...^_^...so jangan buli aku ag aw...heee	55.5

CONCLUSION

In summary, social media plays an important part in Natural Language Processing (NLP). The source is infinite and there is still many ways to improve the combination of NLP techniques in constructing a better and reliable corpus. Due to the dynamic nature of user's behaviour and their informal ways of writing texts, the challenge in keeping track of their freshly created abbreviations to the existing corpus is ever growing.

Based on the experience of constructing this Malay Social Media Corpus, there are many ways to improve the capability of the corpus. For example, the medium to store the corpus, MySQL can further extend using NOSQL (Not only SQL) that can greatly boost the searching performance. In addition, other NLP techniques can be explored to enhance the ambiguity issue faced in the application. Moreover, the size of the corpus can be further expand by extracting more social media posts and provisioning more training set data. The number of root words can be increased by extracting online dictionary of Dewan Bahasa dan Pustaka into vocabulary collection once the online data dictionary can be fully accessed.

ACKNOWLEDGEMENTS

The researchers express appreciation to Ministry of Higher Education (MOHE), Malaysia through Fundamental Research Grant Scheme (600-RMI/FRGS 5/3 (163/2013)) and Universiti Teknologi MARA for sponsoring this study.

REFERENCES

Aho, A.V., 1980. Pattern Matching in Strings Formal Language Theory: Perspectives and Open Problems. Academic Press Inc, New York, USA., pp: 325-347.

Aw, A., M. Zhang, J. Xiao and J. Su, 2006. A phrase-based statistical model for SMS text normalization. Proceedings of the International Conference on Poster Sessions COLING-ACL on Main, July 17-18, 2006, ACM, Stroudsburg, PA, USA., pp: 33-40.

Aw, A.T. and L.H. Lee, 2012. Personalized normalization for a multilingual chat system. Proceedings of the International Conference on ACL 2012 System Demonstrations, July 10, 2012, ACM, Stroudsburg, USA., pp: 31-36.

Bali, R.M., C.C. Chong and K.N. Pek, 2007. Identifying and classifying unknown words in Malay texts. Master Thesis, Universiti Sains Malaysia, George Town, Malaysia.

Basri, S.B., R. Alfred and C.K. On, 2012. Automatic spell checker for Malay blog. Proceedings of the 2012 IEEE International Conference on Control System, Computing and Engineering (ICCSCE), November 23-25, 2012, IEEE, Kota Kinabalu, Malaysia, ISBN:978-1-4673-3143-2, pp: 506-510.

Branckaute, F., 2010. Facebook Statistics: The Numbers Game Continues. Airbnb Inc., San Francisco, California.

Chen, T. and M.Y. Kan, 2013. Creating a live, public short message service corpus: The NUS SMS corpus. Lang. Resour. Eval., 47: 299-335.

Cook, P. and S. Stevenson, 2009. An unsupervised model for text message normalization. Proceedings of the Workshop on Computational Approaches to Linguistic Creativity, June 4, 2009, ACM, Stroudsburg, PA, USA., ISBN:978-1-932432-36-7, pp: 71-78.

Gadde, P., R. Goutam, R. Shah, H.S. Bayyarapu and L.V. Subramaniam, 2011. Experiments with artificially generated noise for cleansing noisy text. Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data, September 17-17, 2011, ACM, New York, USA., ISBN:978-1-4503-0685-0, pp: 1-4.

Han, B., P. Cook and T. Baldwin, 2012. Automatically constructing a normalisation dictionary for microblogs. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, July 12-14, 2012, ACM, Stroudsburg, PA, USA., pp: 421-432.

Jehl, L.E., 2010. Machine translation for Twitter. MSc Thesis, University of Edinburgh, Edinburgh, Scotland.

Joseph, C., C. Muthusamy, A.S. Michael and D.S.D. Telajan, 2013. Strategies applied in SMS: An analysis on SMS column in the star newspaper. Asian Soc. Sci., 9: 8-13.

Kayte, S.N. and M.A. Mundada, 2016. Corpus-driven marathi text-to-speech system based on the concatenative synthesis approach. Int. J. Eng. Res. Gen. Sci., 4: 14-20.

Kobus, C., F. Yvon and G. Damnati, 2008. Normalizing SMS: Are two metaphors better than one?. Proceedings of the 22nd International Conference on Computational Linguistics, August 18-22, 2008, ACM, Stroudsburg, PA, USA., pp: 441-448.

Musa, H., R.A. Kadir, A. Azman and M.T. Abdullah, 2011. Syllabification algorithm based on syllable rules matching for Malay language. Proceedings of the 10th WSEAS International Conference on Applied Computer and Applied Computational Science, March 08- 10, 2011, ACM, Wisconsin, USA., ISBN:978-960-474-281-3, pp: 279-286.

MySQL, 2001. Developer zone. MySQL, Cupertino, California. http://dev.mysql.com/

Neubig, G., 2012. Unigram language models. Nara Institute of Science and Technology, Ikoma, Japan. http://www.phontron.com/slides/nlp-programming-en-01-unigramlm.pdf.

Samsudin, N., M. Puteh, A.R. Hamdan and M.Z.A. Nazri, 2012. Normalization of common noisyterms in Malaysian online media. Proceedings of the Knowledge Management International Conference on (KMICe), July 4-6, 2012, Universiti Utara Malaysia, Changlun, Malaysia, pp: 515-520.

Samsudin, N., M. Puteh, A.R. Hamdan and M.Z.A. Nazri, 2013. Mining opinion in online messages. Int. J. Adv. Comput. Sci. Appl., 4: 19-24.

Smith, T.M.F., 1976. The foundations of survey sampling: A review. J. Royal Statist. Soc. Ser., 139: 183-204.

Twitter, 2011. Twitter official blog. Twitter Inc, San Francisco, California. https://blog.twitter.com/2011/numbers

Wang, Y., Q. Min and S. Han, 2016. Understanding the effects of trust and risk on individual behavior toward social media platforms: A meta-analysis of the empirical evidence. Comput. Hum. Behav., 56: 34-44.

Zesch, T., C. Muller and I. Gurevych, 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. Technische Universität Darmstadt, Darmstadt, Germany, pp: 1646-1652.