International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-3, January 2020

Tryna b Kewl: Textual Analytics of Distorted Words among Malaysian Millennials on Twitter

Nur Nashatul Nasuha Nazman, Kee-Man Chuah, Su-Hie Ting

Abstract*: The presence of Malaysian millennials on social media platforms is increasingly gaining attention particularly on Twitter. Language wise, many of them are predominantly using English and Malay in their tweets but with a touch of their own "styles" in various morphological aspects. This trend eventually leads to a rampant use of distorted vocabulary, churning out many non-standard words. This study aims to address the need in classifying the types of morphological distortions of words that are widely used among the Malaysian millennials and identify the reasons behind such trend. A total of 50 active Twitter users from Malaysia aged 18 to 30 years old were randomly chosen for this study. From each user, 20 tweets of longer than 5 words were selected for lexical analysis, giving a sum of 1000 tweets (8443 words in total). Then, interviews were conducted on 30 participants to gauge the factors of using those non-standard words. The findings revealed that the words were largely distorted in terms of its inflections so as to fit some sounds. Also, most distorted words were deliberately coined so that the millennials would appear trendy, while some were merely following the usage without knowing the actual word. This study has shown that the use of distorted words among Malaysian Twitter users did not hinder effective communication.

Keywords: morphological distortion, textual analytics, Twitter

I. INTRODUCTION

Social media is a common tool of communication in the digital age. Among them, Twitter is one of the popular platforms that has been widely used. It is a free social networking microblogging service that allows registered members to broadcast short posts called tweets. Tweets were initially limited to 140 characters because to the constraints of Twitter's Short Message Service (SMS) delivery system, but it was later increased to 280 characters in November 2017 or approximately 40 words. Gligoric, Anderson and West [1] studied the consequence of the switch and found out that despite the increase of characters, tweets produced are still predominantly concise.

Revised Manuscript Received on January 30, 2020.
***** Correspondence Author
Nur Nashatul Nazman*, Department of Language and Communication, University Malaysia Sarawak, Malaysia.
Kee-Man Chuah, Lecturer, Department of Language and Communication, University Malaysia Sarawak, Malaysia.
Su-Hie Ting, Associate Professor, Department of Faculty of Language and Communication, University Malaysia Sarawak, Malaysia.

The restriction imposed by Twitter, therefore, has indirectly trained many Twitter users to be more economical and creative in their expression of thoughts, creating a plethora of newly-formed "words", in order to ensure everything fits in one tweet.

Due to Twitter's microblogging method, users are required to make short, frequent posts to a microblog. Microblogging in Twitter may include hashtags (auto tagging of topics or words), mentions (mentioning or links to other Twitter users) or to other links from web pages, images or videos [2, 3]. From this platform, it is great if people want to quickly record their thought, opinions, ideas and create awareness. The main advantage of this platform is user-friendly and easier for users to be connected, given that the interaction is live and in real time [4]. In contrast, the character-limit set by the platform has prompted the creation of many "alien words" for different reasons [5]. This leaves ample room for further investigation to be done in relation to the content generated by the Twitter users.

Microblogging platforms such as Twitter have been the subject to many studies in the recent years. The studies in this area can be grouped into automatic sentiment analysis and opinion mining as it presents a huge source of data representing the opinions of a significant, yet totally random. Studies in the area of sentiment analysis [1, 6, 7, 8] depends upon random extraction of tweets and identify specific sentiment that researchers would like to uncover ranging from political inclination to prediction of suicidal thoughts. The second group of studies on opinion mining [5, 9] focuses on extracting keywords from tweets that match the intended target such as preference towards a product or reviews. Opinion mining is used by most marketing consultants as well as business corporations in obtaining the market trends.

The problem identified through the review of such studies is that sentiment analysis and opinion mining were done solely based on well-constructed words that convey such meaning. The thematic interpretation of extracted Twitter content has resulted in conflicting findings. This shortcoming is largely due to the informal language use, the presence of non-textual content and the use of slang words and abbreviations, that impede the accuracy of the mining process. Although the studies reviewed were involving English tweets, the growing trend of Malaysian twitter users in using slang words and abbreviations has also been studied [10]. With approximately 3.5 million users and more than 2 million tweets being generated in Malay language (Bahasa Melayu) apart from English on a daily basis [11],

the potential of using Twitter's large corpus of user-generated text-based contents for linguistic analysis is largely untapped. This study, therefore, aims to investigate the types of distortion used by Malaysian Twitter users in "coining" new words and link those to their corresponding factors. The non-standard words, in this study, refer to the words that cannot be found in the dictionary.

The following questions are addressed in this study:

1. What are the non-standard words that are widely used by Malaysian millennials on Twitter?
2. What are the types of morphological distortion identified in the non-standard words used by Malaysian millennials on Twitter?
3. What are the factors that influence the morphological distortions on the non-standard words used by Malaysian millennials on Twitter?

II. REVIEW OF RELATED STUDIES

It is undeniable that written language on the Internet is not same as on paper. Language use on the Internet is more informal. Users start to use abbreviations, acronyms, emoticons and distorted words [12]. This helps users that use tiny keys on their mobile phones and for sure helping in limitation of words in Twitter especially by using abbreviations, acronyms and distorted words. For emoticons, it is a useful element for non-verbal language to communicate. It is describing emotions, facial expressions or physical actions without using words. The style of communication is become more open and more informal. In additions, Internet is the medium for new words and new meanings.

No one can find non-standard words in a dictionary, nor can one find their pronunciation by an application of ordinary "letter-to-sound" rules as explained by Sproat et al. [12]. Non-standard words usually drop vowels in words but sometimes consonants are drop too. Although non-standard words are not in the dictionary, but a lot of people are using it in communication especially in the Internet. In this study, non-standard word refers to those which are not found in dictionary.

This study is based upon the major tenets of cognitive linguistics. Cognitive linguistics, one of the modern schools in linguistics today, explores the process of perception of the reality and concentrates on studying and understanding how the language reflects our experiences of the world. It places central importance on the role of meaning, conceptual processes and embodied experience in the study of language and the mind and the way in which they [13]. In this sense, it argues that the cognitive ability to use language is closely connected to some other abilities which include the abilities to perceive, conceptualize and categorize the information. In this study, a specific reference to cognitive semantics is used.

In analysing the factors that contribute to the morphological distortions of Malay words on Twitter, concepts like Conceptual Metaphor, Metonymy and Emblem within Cognitive Semantics are taken. In Conceptual Metaphor, there is a cross-domain conceptual mapping generally from a concrete source domain to an abstract target domain. In Metonymy, there is also a conceptual mapping between a source and a target; both being a part of a single domain. An Emblem can be defined as a "stereotypical conceptual prototype" which represents an "abstract quality or attribute". The term "cognitive semantics" is somewhat misleading, as it may suggest that semantics is a separate module within the linguistics model, next to "cognitive syntax", "cognitive morphology", "cognitive pragmatics" [14].

However, based on Langacker [15], cognitive linguistics does not adopt a modular view on language such as all structures in language, ranging from morphemes, to words to syntactic patterns are considered as inherently meaningful and moreover, as being of the same kind like symbolic form-meaning pairings, called symbolic units (Langacker's Cognitive Grammar or Construction Grammar). In addition, grammar is defined as a structured inventory of such form-meaning pairs. For lexical items and morphemes, as based on de Saussure, assuming such a form-meaning pair is quite uncontroversial. If one were pushed to summarize cognitive semantics in a number of keywords, some proper candidates would be conceptualisation, construal, image schemas and prototype-structured categorisation. All these principles are not restricted to lexical items, but underlie linguistic structures at all levels, ranging from morphemes, lexical items, semi-open idioms, to semantically highly schematic grammatical patterns.

This study employs Kempson's theory for analysis data. Kempson [16] defined there are three ways that already been used by linguists and philosophers for giving an explanation for exact meaning; (a) explaining word meaning, (b) explaining word meaning characteristic and (c) explaining relationship process. In this modern world, language also has impact. For example, in the expansion of meaning and narrowing meaning in Bahasa Melayu vocabulary. Dictionary compiler has problem in choosing exact meaning for a new meaning for a word.

III. METHOD

Qualitative research design was employed to address the objectives of this study. The main methodology used was textual analytics with the assistance of corpus linguistics knowledge. Due to the nature of the required data, qualitative judgement is necessary in order for the researchers to collect accurate textual evidences for analysis.

For this purpose, the researchers chose tweets from random public users, as long as the user is a Malaysian millennial and the tweets are either in English or Malay. So, the first thing to look up is when the last tweets are posted. This is important to make sure the users are active. Their last tweets must be posted within the last month. Next, the researchers checked the number of tweets they produced (this is to ensure they are real and active users and not spam bots). Tweets with emoji, symbols or emoticons were not counted as the main focus for this research was to find text-based expressions.

After 1000 tweets were selected, the researchers analyzed the morphological distortions from the tweets. Tweets were copied for analysis purpose.

All non-standard words within the compiled tweets were gathered and analysed using a textual analytics software (namely AntConc).

The analysis generated the frequency count of each word and only subsequently filtered to top 100 words. The researcher then classified or coded each word according to its category.

A. Sample and Sampling Method

The main population of this study was the Twitter users from Malaysia. According to the statistics, Malaysia is home to 3.5 million Twitter users. However, beginning February 2014, only 2.8% of these users were active and approximately 162.4 million tweets were published in that time period. This means that Malaysian users tweeting roughly 5.4 million tweets per day. In selecting the sample, two sets of participants were chosen for the two phases of the study.

In phase one, 50 Twitter users from Malaysia were randomly chosen from the pool of users that the researcher had compiled. In order to be selected, the Twitter users must fit the following criteria:

i. Aged 18 to 30 years old.
ii. Has posted more than 50 tweets with words.
iii .Active for the last one month.
iv. Tweets in Bahasa Melayu and English only.

In phase two, 30 participants were randomly selected from aged 18 to 30 for an in-depth interview on the usage of abbreviations and coined words in social digital discourse. The participants were randomly selected, and they must be active users of Twitter.

B. Data Analysis Procedures

The collected Tweets were processed and textually analysed using AntConc softwafre. The most highly used distorted words were ranked and categorised according to the list that the researchers have set as a guideline.

Table- I: General guideline for coding of distorted words

TYPES	EXAMPLE
Deletion of vowels	can → cn; now→ nw
Direct use of acronym	Laugh out loud → LOL; by the way → BTW
Reduced phrase/ partial deletion	Trying to → tryna; going to → gonna
Vowels changing	What → Whut; coffee → coffi
Consonant changing	Orang → owang; fish → pish
Coined	cool → kewl; kedekut → kedek

The researchers only focused on top 20 of non-standard (distorted) words. Non-standard words were listed and ranked according to their frequency count. The words were then used for semi-structured interview in phase two; semi-structured interview. The objective of this phase was to know the factors for such usage.

IV. RESULTS AND DISCUSSION

The analysis of gathered tweets began with the frequency count of the most common non-standard words used by the participants as shown in Table II.

Table- II: Top 20 distorted words used by participants

RANK	WORD	FREQUENCY
1.	Ni	102
2.	Aq	90
3.	Lak	74
4.	Abt	70
5.	Dak	57
6.	Fon	54
7.	Gonna	49
8.	Pls	48
9.	Ko	47
10.	Amek	45
11.	De	43
12.	Gotta	42
13.	R	38
14.	RN	37
15.	Dy	33
16.	Yr	30
17.	Coz	23
18.	Owg	11
19.	Nw	8
20.	Kedek	4

Based on the results, the word “ni” topped the chart with 102 times in the gathered data. This is followed by “aq” and “lak” with 90 and 74 times respectively. The distorted word with lower frequency count is “kedek”, which appeared only 4 times. The gap between all the frequency is quite close. All the words from the list obviously are non-standard words.

The top 5 of non-standard words in the list are “ni”, “aq”, “lak”, and “dak”. These words are categorised by partial deletion as they were formed by dropping part of the full word as in “ini”, “aku”, “pula”, and “budak” respectively. On the other hand, “aq” and “lak” are been shortened and replaced with another letter at the back of words.

While for the words “abt” (about) and “nw” (now), the users ignored the vowels in the middle. This is a good example of vowels deletion, which is rather common in the Internet language. Interesting, consonant changing is also popular such as the chase of “owang”, where “r” was replaced with “w”. This trend is noted to be rather popular among Malaysian millennials, as to indicate cuteness or friendliness. There is also partial deletion that drops only the end of the word such as “kedek” for “kedekut”. It is unclear why only -ut was dropped but the consensus among the users during the interview indicated that the word could be coined just for fun or novelty.

During the data collection process, the researchers also found other unique words such as “sumer” (coined), “cenni” (reduced phrase), “nan” (partial deletion) and “lek” (partial deletion). The frequency of used not really high but they exist among the Malaysian Twitter community.

The original words are “semua”, “macam ini”, “dengan”, and “relax”. These words despite their lower usage were still able to be comprehended by the respondents during the interview.

This findings from this study rather aligned to those found by Ahmed [17] and Sproat et al. [17]. According Ahmed, lexical normalization is widely used for certain reasons. For example, limitation in social media and to keep it simple for time saving. In addition, the non-standard usage of a mixture of both unintentional misspelling and intentionally-created words for various reasons [18].

During the semi-structured interviews, these two factors were repeatedly mentioned by participants. These proven that the limitation of characters and the intention to save time are the main contributing factors for various distorted words formation in Twitter. 15 participants of the interview session agreed that by using non-standard words, it is simpler, and they sound trendy. The style of communication has become more open and more informal. In addition, social media are the platforms for new words and new meanings. Although new words are formed, they are acceptable among Twitter users and widely used. Thus, social media are regarded as a good source to contribute to the forming of language [19].

From the previous research, Sproat et al. [17] also proved that normalised words are helping users to save their time in communication using technology, but nowadays normalisation of words become trends because of the existence of social media. As the social media become a part of almost people in world, the existence of normalised words is raising especially for lexical normalisation.

It is rather clear that Twitter has contributed to non-standard words usage among Malaysian millennials in Twitter. In social media, this is reported as “normal” way to communicate with each other and nowadays normalization of words become trends [17]. This platform is a free format of messages and writing because users are freely to tweet. As stipulated by Jansen et al. [20], the linguistic structure of tweets and its pattern of natural language expression by users is prove that user is feel free to tweet. Twitter consider as powerful social platforms as it not also influence people to communicate using it, but it also build trust between users because they have to have trust in using this platform with families, friends or strangers in Twitter.

In this study, 5 of the non-standard words in the list are “ni”, “aq”, “lak”, “abt” and “dak”. These words are not only removed their characters but also replaced with another letter. This shows that language in Twitter is not same as on paper as the language use in Twitter is more informal. The Twitter users tend to use more abbreviations, acronyms, emoticons and normalised words. This helps users that use tiny keys on their mobile phones and for sure helping in limitation of words in Twitter especially by using abbreviations, acronyms and normalised words. The normalisation task is challenging because it has similarities with spell checking but differs in that ill-formedness in text messages is often intentional, whether due to the desire to save characters/keystrokes, for social identity, or due to convention in this text sub-genre.

Word class for “ni” and “aq” are pronoun, “lak” is adverb, “abt” is preposition and “dak” is noun. This shows that pronouns largely turned into non-standard words. Based on the data, this is the list of word class with the number of users. Interestingly, from the list, it shown that Twitter users like to normalise pronoun more than other word class.

It is also apparent in this study that Malaysian youths use non-standard words in Twitter because they create non-standard words from sound imitation. However, this finding is not in line with the findings found by Penell and Liu [21]. They found that non-standard words actually widen the misinterpretation and the wrong way to pronounce words. Most of the participants agreed that one of the reasons why they use non-standard word is the sound imitation as it is easier to understand compared to its original spelling for some people. Clearly, the mentioned factors do influence the way words are coined and distorted by Malaysian millennials on Twitter.

V. CONCLUSION

The findings from the textual analytics of Twitter was conducted to answer the first research objective, which is to identify the non-standard words which are coined by Malaysian youths on Twitter. It showed that top 20 of most used of non-standard words among Malaysian youth. All the non-standard were listed according to their rank. In addition, the findings in semi-structured interviews revealed the factors of why Malaysian youths use non-standard words in Twitter. The participant revealed that the reasons why they used non-standard words in Twitter are save time, limitation of characters, imitate the sounds and to make it simpler and trendy.

In order to fulfil this study area, some interventions can be done. During the process of collecting data for the corpus analysis, the researcher found some unique and rare words. From the observation, only a few of them used it in their Tweets. For sure, it is not typos or misspelling because they kept used it in Twitter. So, in the future study, the researcher can look through sociolinguistics point of view. Although they all belong to Twitter community and used non-standard words, they still had their differences. This differences that future research need to cover. Twitter users might be divided according to their groups. For example, hobbies, interest or environment. This will give answer to why they use certain non-standard words in their Twitter. From this future research, researcher can also look into why they use these words such as for their identity or secret code in conversation with their group [22].

This research is built for finding factors on why Malaysian millennials use non-standard words in Twitter and identify the non-standard words in Twitter. The findings fulfil the existing gap in context of Malaysian millennials. Furthermore, this research provided a corpus of non-standard words in Twitter and listed which words are usually normalized by Malaysian youth in Twitter. A combination of corpus analysis of Twitter and semi-structured interview gave a useful method and ways for investigation and compiling the data. The study not only filled the previous research but also increased the understanding of why they used non-standard words in Twitter among Malaysian youths.

ACKNOWLEDGMENT

We would like to thank Universiti Malaysia Sarawak all the participating respondents for supporting this research work.

REFERENCES

1. Gligorić, K., Anderson, A., & West, R. (2018, June). How constraints affect content: The case of Twitter's switch from 140 to 280 characters. In Twelfth International AAAI Conference on Web and Social Media.
2. Lowenthal, Patrick R., and Joanna C. Dunlap. "From pixel on a screen to real person in your students' lives: Establishing social presence using digital storytelling." The Internet and Higher Education 13.1-2 (2010): 70-72.
3. Chuah, K. M. (2013). Aplikasi media sosial dalam pembelajaran Bahasa Inggeris: Persepsi pelajar universiti. Issues in Language Studies, 2(1), 56-63.
4. Honeycutt, C., & Herring, S. C. (2009). Beyond microblogging. Conversation and collaboration In 42nd Hawaii International Conference on System Sciences (pp. 1-10).
5. Goldberg, Y. (2016). A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research, 57, 345-420.
6. R. Balamurugan, S. Pushpa. The Public Sentiment and Emotional Variations in Social Media using Twitter Dataset. International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol. 8, No. 12, 2019, 2327 -2333
7. Ghiassi, M., Zimbra, D., & Lee, S. (2016). Targeted twitter sentiment analysis for brands using supervised feature engineering and the dynamic architecture for artificial neural networks. Journal of Management Information Systems, 33(4), 1034-1058.
8. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In Proceedings of the workshop on languages in social media(pp. 30-38). Association for Computational Linguistics.
9. Deli, R. M., Man, C. K., & Razali, N. T. (2019). A Corpus-Based Analysis of Lexical Verbs in L2 Professional Engineering Writing. International Journal of Asian Social Science, 9(8), 461-472.
10. Tan, Y. F., Lam, H. S., Azlan, A., & Soo, W. K. (2016, April). Sentiment Analysis for Telco Popularity on Twitter Big Data Using a Novel Malaysian Dictionary. In ICADIWT (pp. 112-125).
11. Malaysian Communications and Multimedia Commission (2017). Malaysian internet users survey. Retrieved from https://www.mcmc.gov.my/skmmgovmy/media/General/pdf/MCMC-Internet-Users-Survey-2017.pdf
12. Sproat, R., Black, A. W., Chen, S., Kumar, S., Ostendorf, M., & Richards, C. (2001). Normalization of non-standard words. Computer speech & language, 15(3), 287-333.
13. Evans, V. A. (2007). Cognitive linguistics. Edinburgh: Edinburgh University Press.
14. Langlotz, A. (2006). Idiomatic Creativity. Amsterdam: John Benjamins Publishing Company.
15. Langacker, Ronald W. "Cognitive grammar." Concise History of the Language Sciences. Pergamon, 1995. 364-368.
16. Kempson, R. (2019). 11 Formal semantics and representationalism. Semantics-Foundations, History and Methods, 273.
17. Ahmed, B. (2015, July). Lexical Normalisation of Twitter Data. In Science and Information Conference (SAI), 2015 (pp. 326-328). IEEE.
18. Liu, F., Weng, F., Wang, B., & Liu, Y. (2011, June). Insertion, deletion, or substitution?: normalizing text messages without pre-categorization nor supervision. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2 (pp. 71-76). Association for Computational Linguistics.
19. Thorne, S. L., Black, R. W., & Sykes, J. M. (2009). Second language use, socialization, and learning in Internet interest communities and online gaming. The modern language journal, 93(s1), 802-821.
20. Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. Journal of the American society for information science and technology, 60(11), 2169-2188.
21. Pennell, D. L., & Liu, Y. (2010, March). Normalization of text messages for text-to-speech. In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 4842-4845). IEEE.
22. Man, C. K. (2014, August). Word's up with WhatsApp: the use of instant messaging in consciousness-raising of academic vocabulary. In 23rd MELTA and 12th Asia TEFL International Conference (pp. 28-30).

AUTHORS PROFILE

Nur Nashatul Nazman is a Master of Arts candidate at the Faculty of Language and Communication, Universiti Malaysia Sarawak, Malaysia. Her main interest is in textual analytics and corpus linguistics.

Kee-Man Chuah is currently a lecturer at the Faculty of Language and Communication, Universiti Malaysia Sarawak, Malaysia. His research works mainly centres on computational linguistics and educational technology.

Su-Hie Ting is currently an Associate Professor at the Faculty of Language and Communication, Universiti Malaysia Sarawak, Malaysia. Her main research interests are sociolinguistics and applied linguistics.