Indonesian Journal of Electrical Engineering and Computer Science
Vol. 16, No. 1, October 2019, pp. 465~472
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v16.i1.pp465-472

A taxonomy of Malay social media text

Ruhaila Maskat[1], Yuda Munarko[2]
[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia
[2]Department of Informatics Engineering, University of Muhammadiyah Malang, Indonesia

Article Info

Article history:

Received Nov 1, 2018
Revised Feb 6, 2019
Accepted Mar 15, 2019

Keywords:

Data preprocessing
Malay language
Social media
Taxonomy
Text analytics

ABSTRACT

In this paper, we proposed a preliminary taxonomy of Malay social media text. Performing text analytics on Malay social media text is a challenge. The formal Malay language follows specific spelling and sentence construction rules. However, the Malay language used in social media differs in both aspects. This impedes the accuracy of text analytics. Due to the complexity of Malay social media text, many researches has chosen to focus on classifying the formal Malay language. To the best of our knowledge, we are the first to propose a formal taxonomy for Malay text in social media. Narrow and informal categorisations of Malay social media text can be found amidst efforts to pre-process social media text, yet cherry-picked only some categories to be handled. We have differentiated Malay social media text from the formal Malay language by identifying them as Social Media Malay Language or SMML. They consists of spelling variations, Malay-English mix sentence, Malay-spelling English words, slang-based words, vowel-les words, number suffixes and manner of expression. This taxonomy is expected to serve as a guideline in research and commercial products.

Corresponding Author:

Ruhaila Maskat,
Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA,
Shah Alam, Selangor, Malaysia.
Email: ruhaila@tmsk.uitm.edu.my

1. INTRODUCTION

The use of social media, such as Facebook and Twitter, has recorded significant growth in Malaysia. Facebook account holders have increased from 10.18 million in 2015 to 12.75 million at present [1, 2]. As for Twitter, the 1.5 million account holders in 2015 have grown to 2.2 million to date [1, 2]. Organisations, for-profit and non-profit alike, and governments have now recognised the potential benefits that social media text carries. Text analytics is used to discover knowledge from a large collection of text and often borrows techniques from Natural Language Processing (NLP), Data Mining (DM), Machine Learning (ML) and Information Retrieval (IR) [3]. When dealing with social media, text analytics is used to uncover insights into social networks or groups e.g. sentiment analysis, event detection and customer segmentation [3]. In general, text analytics undergoes phases of Data Acquisition, Pre-processing, Representation and Knowledge Discovery [3]. Pre-processing is the longest phase aimed at transforming data into a form fit to perform analytics [5, 18, 21]. It involves activities of removals, stemming, tokenization, language detection and normalisation. Removals [4] typically involve white spaces, missing values, duplicate reviews, stop words, non-ascii characters and typos which could have an adverse influence on the result. Stemming [19] replaces words with their canonical form for example stand in place of standing, stood and stands. Language detection (LD) [6, 7, 24] is crucial for language-dependent tokenisers and could considerably decrease the size of data extracted. Normalisation [22, 23] converts words into their standard spelling if they are not. Once data is clean, the next phase represents data as numeric vectors (i.e. Bag of Words or Vector Space Model) in preparation

for use in downstream analytics. Finally, knowledge discovery methods would be applied such as supervised classification, clustering, sentiment analysis and event detection.

Malay is a language spoken and written by around 290 million people worldwide [8]. Malay is frequently used in Malaysia, Indonesia, Singapore and Brunei [8]. In Malaysia, the formal Malay language is standardized by an authorized body namely Dewan Bahasa dan Pustaka (DBP). DBP determines the correct spelling and use of Malay language in spoken and written artefacts such as formal speech, formal letters and academic text. Social media, unlike documents, permits very limited text. Users are forced to devise ways to fit more information within the restricted given space, ending up in them altering how text would be presented. Such decisions are often based on what is perceived to be logical by the writer and understanding of the text is left to the reader's own interpretations. Without any authorized body, like the DBP, there is practically no right or wrong with regards to the spelling and usage of Malay language on social media. In this paper, we describe the different kinds of Malay text we found on social media. We named them Social Media Malay Language (SMML) so as to distinguish them from the formal Malay language monitored by the DBP. To our knowledge, there is no formal taxonomy on Malay social media text. The primary objective and contribution of this paper is the taxonomy of SMML. We hope the taxonomy will be utilized in research and commercial tasks.

At the point of writing, no formal taxonomy on Malay social media text has been published. Instead, narrow categorisations were found described within works in text normalization and automatic spell checker [20, 25]. Categorisations appear to be cherry-picked in light of the solution proposed. Misspelled words [13], out-of-vocabulary words [14], ill-formed words and noisy text [14, 15, 17] were used to describe the unconventional condition of Malay social media text. Basri et al. [13] proposed a framework for an automatic spell-checker and corrector for misspelled words. Only slang words from Selangor were handled in this work. Basri et al. also handled the universal character "x" which indicates a negation and twicely duplicated words. Samsudin et al. [14] constructed a set of rules capable of automatically-generating artificial noisy text. These rules were based on an earlier work by DBP as an effort to streamline Short Message Service (SMS) texts [16]. In their work, they dealt with variations of spelling, Selangor and Johor slang words, vowel-less words and R-suffixed words which we identify as a way of expressing a writer's state. Muhamad et al. [15] discussed a conceptual architecture of a Malay text normalisation framework of a work-in-progress where the proposal is to utilise a hybrid dictionary. Their research interest is in spelling variation and Selangor-based slang words. This taxonomy aims at providing a preliminary description of Malay social media text up to the time of writing.

2. METHOD

This taxonomy was constructed via observation on a Malay tweeter corpus. Careful categorisation was done with regards to several identified dimensions. The dimensions are formal spelling-informal pronunciation, second language, trends, geography, small devices and expressions. Formal spelling-informal pronunciation represents the variety of sounds that can be uttered informally on a syllable which differs from its formal spelling. These numerous informal sounds were found adopted into SMML and form our first class of taxonomy (1.0 of Table 1). A second language frequently influences informal communications, often resulting in a mix of two languages in a single sentence. English being the all round second language of most native Malay speakers was discovered mixed with Malay in the corpus in large numbers, thus forming the second class of our taxonomy (2.0 of Table 1). Trends induce the formation of new words or phrases in SMML. Nation-wide event, memes or public figures are common trending triggers. We identified the now widely used English words spelt using Malay phonology which was publicised by a local Malaysian celebrity as our third taxonomic class (3.0 of Table 1). Geography represents how SMML was formed to reflect the unique way different regions pronounce the same Malay word, also known as slang or dialect. We focused on regions in West Malaysia for this study and identified 8 regions with prominent slangs (4.0 of Table 1). Next is small devices which inspire authors to use short spellings to reduce typing time. We found their effect in SMML to produce vowel-less spelling (5.0 of Table 1) and numeric suffixes (6.0 of Table 1). Expressions, such as happy, surprised and annoyed, motivate the use of symbols in social media text. We found universal manners of expressions in SMML as well as a unique one and formed our taxonomic class 7.0 of Table 1.

3. TAXONOMY OF SMML

This section describes further our taxonomy and presents examples. A summary of this taxonomy is shown in Table 1.

Table 1. Taxonomy of Social Media Malay Language (SMML) (*continue)*

- 1.0 Spelling variations
- 2.0 Malay-English mixed sentence
  - 2.1. Direct replacement
    - 2.1.1. English word/phrase in Malay sentence
    - 2.1.2. Malay word/phrase in English sentence
  - 2.2. Combination
- 3.0 English words spelt using Malay phonology
  - 3.1. Consonant
    - 3.1.1. Same as English
      - 3.1.1.1. **b**eau = **b**ola [9]
      - 3.1.1.2. **d**o = **d**ari [9]
      - 3.1.1.3. **f**estival = **f**ikir [9]
      - 3.1.1.4. **g**ain = **g**alah [9]
      - 3.1.1.5. **h**at = **h**abis [9]
      - 3.1.1.6. **j**ob = **j**ari [9]
      - 3.1.1.7. **k**alah = s**k**y [9]
      - 3.1.1.8. c**l**ean = **l**ama [9]
      - 3.1.1.9. **m**oon = **m**akan [9]
      - 3.1.1.10. **n**ote = **n**akal [9]
      - 3.1.1.11. feeli**ng** = **ng**arai [9]
      - 3.1.1.12. ca**ny**on = **ny**aman [9]
      - 3.1.1.13. s**p**y = **p**ola [9]
      - 3.1.1.14. **r**isk = **r**asa
      - 3.1.1.15. **s**ix = **s**aya [9]
      - 3.1.1.16. s**t**y = **t**ari [9]
      - 3.1.1.17. **v**ision = **v**isa [9]
      - 3.1.1.18. **w**e = **w**aktu [9]
      - 3.1.1.19. **y**es = **y**akin [9]
      - 3.1.1.20. **z**ero = **z**aman [9]
    - 3.1.2. Different from English
      - 3.1.2.1. Thick **th** (e.g. brother) = **d** *
      - 3.1.2.2. Thin **th** (e.g. think) = **t** *
      - 3.1.2.3. **c** (e.g. cannot), **ch** (e.g. character) = **k**
      - 3.1.2.4. **ch** (e.g. check) = **c** [9]
      - 3.1.2.5. **sh** (e.g. shock) = **sy** [9]
      - 3.1.2.6. **x** (e.g. max) = **ks**
      - 3.1.2.7. Z-sounding **s** (e.g. is it?, busy) = **z** *
      - 3.1.2.8. American middle **t** (e.g. better, keep it up) = **d** *
  - 3.2. Vowel
    - 3.2.1. Letter A
      - 3.2.1.1. f**a**ther [9], b**u**s = **a**j**a**r [9]
    - 3.2.2. Letter E
      - 3.2.2.1. cl**ay** = s**e**rong, pil**i**h [9]
      - 3.2.2.2. **a**bout [9], p**e**rk, h**u**rt = ap**a**, buk**a** [9], b**e**nd**a**
    - 3.2.3. Letter I
      - 3.2.3.1. s**ee** [9], p**i**n, b**u**s**y**, g**e**ne = b**i**la [9], **i**n**i** [9]
    - 3.2.4. Letter O
      - 3.2.4.1. s**o**le = r**o**da, rump**u**t [9]
      - 3.2.4.2. **o**ff = p**o**hon [9]
      - 3.2.4.3. **awe**some = **o**nak *
    - 3.2.5. Letter U
      - 3.2.5.1. m**oo**n = **u**pah, bar**u**, r**u**mput [9]
  - 3.3. Diphthong
    - 3.3.1. h**ow** = kal**au** [9]
    - 3.3.2. b**ye** = cap**ai** [9]
    - 3.3.3. surv**ey** = murb**ei** [9]
    - 3.3.4. b**oy** = semp**oi** [9]
- 4.0. Regional slang
  - 4.1. West Malaysia
    - 4.1.1. North (Perlis, Kedah, Penang)
      - 4.1.1.1. Letter ‘F’ ↔ letter ‘P’
      - 4.1.1.2. Letter ‘R’ ↔ letter ‘GH’
      - 4.1.1.3. Ending letter ‘A’, no changes
      - 4.1.1.4. Ending letters ‘L’ ↔ ‘I’
      - 4.1.1.5. Ending letters ‘R’ ↔ ‘AQ’
      - 4.1.1.6. Ending letters ‘S’ ↔ ‘IH’
    - 4.1.2. North (Perak)
      - 4.1.2.1. Letter ‘F’ ↔ letter ‘P’
      - 4.1.2.2. Ending letter ‘A’ ↔ ‘E’ or in some instance append with a ‘K’
      - 4.1.2.3. Ending letter ‘AR’ or ‘R’ ↔ ‘OR’
      - 4.1.2.4. Ending letters ‘AS’ ↔ ‘EH’
      - 4.1.2.5. Ending letters ‘L’ ↔ ‘I’

Table 1. Taxonomy of Social Media Malay Language (SMML)

- 4.1.3. East coast (Trengganu)
  - 4.1.3.1. Letter ‘F’ ↔ letter ‘P’
  - 4.1.3.2. Letters ‘IA’ ↔ letter ‘E’ [10]
  - 4.1.3.3. Letters ‘MP’ ↔ letter ‘P’[10]
  - 4.1.3.4. Letters ‘NT’ ↔ letter ‘T’[10]
  - 4.1.3.5. Letters ‘NGK’ ↔ letter ‘K’[10]
  - 4.1.3.6. Letters ‘UA’ ↔ letter ‘O’ [10]
  - 4.1.3.7. Ending letter ‘A’ ↔ ‘E’[10]
  - 4.1.3.8. Ending letters ‘L’ or ‘R’, drop these letters[10]
  - 4.1.3.9. Ending letters ‘AH’ ↔ ‘OH’[10]
  - 4.1.3.10. Ending letters ‘AI’ or ‘AU’ ↔ ‘A’[10]
  - 4.1.3.11. Ending letters ‘AK’ ↔ ‘OK’[10]
  - 4.1.3.12. Ending letters ‘AR’ ↔ ‘OR’[10]
  - 4.1.3.13. Ending letters ‘M’ or ‘N’ ↔ ‘NG’[10]
  - 4.1.3.14. Ending letters ‘P’ or ‘T’ ↔ ‘K’[10]
  - 4.1.3.15. Ending letters ‘S’ ↔ ‘H’[10]
- 4.1.4. East coast (Kelantan)
  - 4.1.4.1. Letter ‘F’ ↔ letter ‘P’
  - 4.1.4.2. Letters ‘IA’ ↔ letter ‘E’ [11]
  - 4.1.4.3. Letters ‘MP’ ↔ letter ‘P’ [11]
  - 4.1.4.4. Letters ‘NT’ ↔ letter ‘T’ [11]
  - 4.1.4.5. Letters ‘NGK’ ↔ letter ‘K’ [11]
  - 4.1.4.6. Letter ‘R’ ↔ letter ‘GH’ [11]
  - 4.1.3.7. Letters ‘UA’ ↔ letter ‘O’ [11]
  - 4.1.4.8. Ending letter ‘A’ ↔ ‘O’ [11]
  - 4.1.4.9. Ending letter ‘P’ or ‘T’ ↔ ‘K’ [11]
  - 4.1.4.10. Ending letter ‘S’ ↔ ‘H’ [11]
  - 4.1.4.11. Ending letters ‘AH’ ↔ ‘OH’ [11]
  - 4.1.4.12. Ending letters ‘AI’ or ‘AU’ ↔ ‘A’ [11]
  - 4.1.4.13. Ending letters ‘AK’ ↔ ‘OK’ [11]
  - 4.1.4.14. Ending letters ‘AM’, ‘AN’ or ‘ANG’ ↔ ‘E’ or ‘AE’ [11]
  - 4.1.4.15. Ending letter ’N’ ↔ ’NG’ [11]
- 4.1.5. Central (Lembah Klang)
  - 4.1.5.1. Ending letter ‘A’ ↔ ‘E’ or in some instance append with a ‘K’
  - 4.1.5.2. Ending letters ‘AR’ ↔ ‘A’
- 4.1.6. Central (Negeri Sembilan)
  - 4.1.6.1. Ending letter ‘A’ ↔ ‘O’
  - 4.1.6.2. Ending letters ‘AS’ ↔ ‘EH’
  - 4.1.6.3. Ending letters ‘T’ ↔ ‘EK’
  - 4.1.6.4. Starting letter ‘E’ ↔ ‘O’
  - 4.1.6.5. In-word letter ‘E’ ↔ ‘O’
- 4.1.7. South (Melaka)
  - 4.1.7.1. Ending letter ‘A’ ↔ ‘E’ or in some instance append with a ‘K’
  - 4.1.7.2. Ending letters ‘AR’ ↔ ‘AU’
- 4.1.8. South (Johor)
  - 4.1.8.1. Ending letter ‘A’ ↔ ‘E’ or in some instance append with a ‘K’
  - 4.1.8.2. Ending letters ‘AR’ ↔ ‘A’ or ‘O’

East

- 4.2. Malaysia

5.0. No-vowel spelling

- 5.1. Multiple letters
- 5.2. Single letter
  - 5.2.1. Similar-sounding
  - 5.2.2. Dissimilar-sounding

6.0. Numeric suffixes

7.0. Conveying expressions

- 7.1. Repeating letters or punctuation marks
- 17.2. Adding the letter ’R’ at a word ending with a vowel

3.1. Spelling Variations

Table 2 shows different forms of spelling used for the same word, with the leftmost being the standard Malay spelling. The variations listed portray the different decisions that the authors have made. One decision is to mimic the common manner by which a word is pronounced by a native speaker. In Malay, formal spelling can differ from actual pronunciation of words, especially letters at the end of a word. For example, in the first line, the way the first letter “a” in the word “apa” is pronounced is different from the “a” at the end. Hence, a letter e is put in place of “a” to form the word “ape”, resembling spoken pronunciation. Another decision is made based on the common manner by which Malay native speakers tend to trim words during informal conversations. The second line shows examples of trimmed words; all holding the same meaning. The phrase

"macam mana" shows trimming of the start of a word, leaving only its end where "macam" becomes "cam" and mana trimmed to "na" or "ne", following the previous decision. Hence, resulting in "camne" and the likes. Words or phrases under this category are noticeably similar at different degrees ("apa" and "ape" being highly similar, "macam mana" and "cane" have very low similarity).

Table 2. Spelling variations

No.	Variations
1	apa = ape
2	macam mana = mcmana, camne, mcmne, cane
3	cerita = cite
4	balik = blik
5	betul = betol, btul, btol
6	boleh = bole, buleh, bule, ble
7	bulat = bolat, bulats, bolats
8	dekat mana = ktmana, ktne, katne
9	kalau = kalo, klu, kalu
10	masuk = masok, msuk
11	mereka = meka
12	sebelum = sblum, sblom
13	tidak ada = takde, tade
14	tidak hendak = tanak, tak nak
15	tidak mahu = tak mahu, tak mau, tamau

3.2. Malay-English Mixed Sentence

English is a second language to most Malaysians after Malay. As a result, it is commonplace to find the mixing of Malay and English words or phrases in the same sentence in informal communications. Categorisedn as a type of SMML are reviews that have both Malay and English words/phrases in a single sentence. This setence can either abide by the English sentence construction rules, obey Malay construction rules or a combination however deemed suited by the author. The words/phrases can be used at the start of a sentence, middle or at the end. These mixed sentences can be grouped into two: direct replacement and combination. Direct replacement can be further grouped into either having English word/phrase in a Malay sentence or replacing some word/phrase in an English sentence with Malay. In Table 3, lines 1 and 2 show the former and in line 3 is an example of the latter. The phrases "improve income" and "Telco provider" substitutes their equivalent Malay phrases. In contrast, Malay words "rakyat" and "untuk" are found embedded in an English sentence. Lines 4 and 5 contain both English and Malay sentence structure in an unobvious combination.

Table 3. Malay-English mixed sentence

No.	Sentences
1	I tak tahu macamana nak *improve income*
2	Dah sampai masa utk tukar *Telco provider*
3	Waive GST for sports equipment the *rakyat* should be encourage *untuk* stay healthy
4	Why items zero gst from all level *ke*?
5	*In my opinion* umat Islam tidak patut dikenakan GST sebab dalam Islam kita dah ada zakat

3.3. English Words Spelt using Malay Phonology

Another type of SMML is writing English words using Malay sound system a.k.a. phonology. Phonology is "the science of speech sounds" . This basically adds playfulness into a message. The International Phonetic Association (IPA) [12] divided Malay phonology [9] into three major groups: consonant, vowel and diphthong. We adopted this categorisation to our taxonomy and adapted it by further dividing the consonants into letters of the alphabet which sound the same in Malay and in English and another group of consonants consisting of unmatching letters used in the approximation of Malay sounds to its English equivalence. For example, the Malay "k" is equated to the English "c", as in "kaktus" and "cactus". Additionally, we observed the existence of newly created approximations, unique to Malay social media text (e.g. "tink" and "think"), and labelled them with an asterisk (i.e. *). Example SMML words/phrases are listed in Table 4 for consonants while vowels and diphthongs are in Table 5.

Table 4. Consonant

No.	Words	English and Malay Sound
1	**b**rother = **b**rader	Same
2	**f**amous = **f**emes	Same
3	**g**uarantee = **g**erenti	Same
4	**h**usband = **h**asben	Same
5	**j**ealous = **j**eles	Same
6	**k**eep it up = **k**ipidap	Same
7	**l**egend = **l**ejen	Same
8	**m**essage = **m**esej	Same
9	**n**aughty = **n**otti	Same
10	to**p**up = to**p**ap	Same
11	**r**elax = **r**ileks	Same
12	**s**chool = **s**kul	Same
13	**t**opup = **t**opap	Same
14	bro**th**er = bra**d**er *	Different
15	**th**ink = **t**ink *	Different
16	**c**annot = **k**enot	Different
17	**ch**aracter = **k**erekter	Different
18	**ch**eck = **c**ek	Different
19	**sh**ock = **sy**ok	Different
20	ma**x** = ma**ks**	Different
21	i**s** it = i**zz**it *	Different
22	keep i**t** **u**p = kipi**d**ap *	Different

Table 5. Vowel & Diphtong

No.	Words	Vowel/Diphtong
1	what = wat *	Vowel
2	jealous = jeles	Vowel
3	husband = hasben	Vowel
4	naughty = noti, notti	Vowel
5	busy = bizi	Vowel
6	keep it up = kipidap	Vowel
7	awesome = ohsem *	Vowel
8	office = ofis	Vowel
9	school = skul	Vowel
10	good = gud	Vowel
11	wow = wau	Diphtong
12	bye = bai	Diphtong
13	hey = hei	Diphtong
14	boy = boi	Diphtong

3.4. Spelling Malay Words based on Regional Slang

Slangs are region-dependent. The slang of some regions may differ slightly from the formal Malay language while others may vary substantially. For example, the word *"besar"* can be written as *"beso"*, *"besa"*, *"bosa"*, *"besaq"* or *"godang"*, depending on which part of Malaysia the slang is adopted from. This kind of SMML use the spelling rules of the formal Malay language in order to write words to the sound of a slang. Table 6 lists example words. We formed our taxonomy based on regions of West and East Malaysia. In this paper, our focus is on West Malaysia slangs. We refine the taxonomy into regions of north, east coast, central and south. Speakers of neighbouring states tend to have similar slangs with some differences, be it prominent or not. For example, northern states such as Perlis, Kedah and Penang are nearly indistinguishable while east coast states, i.e. Trengganu and Kelantan, are notably different. The essence of our taxonomy is in interchangeability. This refers to letters in a formal Malay word that could be replaced to produce a slang word. For example, in Table 1 Item 4.1.1.1. states that letter *"F"* is used interchangeably (↔) with *"P"*. This can be seen in the word *"fikir"* which becomes *"pikir"* when northern slang is applied. Then, Item 4.1.1.5 specifies that letters ending with *"R"* are interchangeable with letters *"AQ"*. As a result, *"pikir"* becomes *"pikiaq"* as how most northern speakers would pronounce and spell in social media. Words that not interchangeable e.g. *"godang"* are not defined in our taxonomy considering that they require a synonym based dictionary.

Table 6. Slang-Based Spelling

Standard Malay	North	North-Perak	East coast - Trengganu	East coast - Kelantan	Central - Klang Valley	Central - Sembilan	South - Melaka	South - Johor
besar	besaq	beso	beso	besar	besa	godang	besau	bes(a/o)
fikir	pikiaq	pikior	pikir	pikir	fikir	pikir	pikir	fikir
raya	ghaya	raye	raye	rayo	raya	ghayo	raye	raya
panas	panaih	paneh	panah	panah	panas	paneh	panas	panas
suka	suka	gemor	suke	suko	suke	suko	suke	suke
pinggan	pinggan	pinggan	pinggang	pingge	pinggan	pinggan	pinggan	pinggan
semut	semut	semut	semuk	semuk	semut	somut	semut	semut
demam	demam	dedor	demang	deme	demam	domam	demam	demam
lemah	lemah	lemah	lemoh	lemoh	lemah	lomah	lemah	lemah
kedekut	kedekut	kedekut	kedekuk	kedekuk	kedekut	kodokut	kedekut	kedekut

3.5. No-vowel Spelling

SMML excessively use short-formed spellings with very few or without any vowels used. Consonants are largely utilised to convey an author's message. Here, all vowels are deliberately left out to save writing time and space. Examples are as listed in Table 7. An SMML word may also consist of a single character that may or may not sound like the original word. An example of the former is the word *"pergi"* which could be represented with the letter *"g"* or *"p"*. The word *"tidak"* is an example of the latter. It means no or not and is represented with the universally-used letter *"x"*.

Table 7. No-vowel Spelling

No.	Words	No. of character	Similar-sounding
1	betul = btl	Multiple	-
2	dah = dh	Multiple	-
3	faham = fhm, phm	Multiple	-
4	jangan = jgn	Multiple	-
5	nama = nm	Multiple	-
6	pada = pd	Multiple	-
7	pula = plk	Multiple	-
8	tanya = tny	Multiple	-
9	tentang = ttg	Multiple	-
10	yang = yg	Multiple	-
11	pergi = g, p	Single	Yes
12	tidak = x	Single	No

3.6. Numeric Suffixes

In Malay, some words and phrases are composed from duplicating words, be it twice or thrice. Twice duplicated words can carry a different meaning from their non-duplicating counterpart or they can indicate pluralism, depending on the context. For example, the word *"hati-hati"* (be careful), *"hati"* (heart) and *"hati-hati"* (hearts). A much used thricely duplicated phrase is the *"ish, ish, ish"*, indicating disbelieve while the word *"ish"* is commonly used to express annoyance towards something. In SMML, these words are often represented with a number attached to the end, e.g. *"ish3"* or *"hati2"*.

3.7. Manner of Expression

There are two manners by which authors can express their messages. The first manner is by repeating a letter in a word e.g. *"I liiiiikkkkeee that"*. This is used in most languages, denoting either pleasure, surprise, resentment etc. The second manner, however, is isolated to SMML. It is by adding the letter *"R"* at the end of a word ending with a vowel. For example, *"aper"*, *"biler"*. The tone of the expression can be both positive or negative, describing excitement or displeasure. This, however, could also be overly used by some authors to the point of normalcy for them.

4. CONCLUSION

As a conclusion, this paper proposed a preliminary formal taxonomy of Malay social media text, namely SMML. SMML is unlike the standard Malay language in the aspects of spelling and sentence construction. The proposed classes are spelling variations, Malay-English mix sentence, Malay-spelling English words, slang-based words, vowel-less words, numeric suffixes and manners of emphasis. This taxonomy is expected to serve as a reference for research and commercial works.

ACKNOWLEDGEMENTS

The authors would like to thank Hamidon Yusof, Ismasabbah Ismail, Brother Sharuzee, Afida Suharti, Ahmad Fadhil Hj Zakariya, Suriyati Rahim, Enni Kesuma Ab. Mukhti, Roshanita Abd Hamid, Fauzi Nusi, Masdi Mohd Din and Hanif Shafiee for their input on regional slangs as native speakers. Our acknowledgement also goes to the Universiti Teknologi MARA Shah Alam, Malaysia and University of Muhammadiyah Malang, Indonesia for the support given.

REFERENCES

[1] "Malay Language", https://en.wikipedia.org/wiki/Malay_language Last accessed December 2018.
[2] Statistica, https://www.statista.com/statistics/490484/number-of-malaysia-facebook-users/ Last accessed 4 July 2018.
[3] Hu, X., & Liu, H. (2012). Text analytics in social media. In Mining text data (pp. 385-414). Springer, Boston, MA.
[4] Derczynski, L., Maynard, D., Aswani, N., & Bontcheva, K. (2013, May). *Microblog-genre noise and impact on semantic annotation accuracy*. In Proceedings of the 24th ACM Conference on Hypertext and Social Media (pp. 21-30). ACM.
[5] Haddi, E., Liu, X., & Shi, Y. (2013). *The role of text pre-processing in sentiment analysis.* Procedia Computer Science, 17, 26-32.
[6] Balazevic, I., Braun, M., & Müller, K. R. (2016). Language Detection For Short Text Messages In Social Media. arXiv preprint arXiv:1608.08515.
[7] Lui, M., & Baldwin, T. (2014). Accurate Language Identification of Twitter Messages. In Proceedings of the 5th workshop on language analysis for social media (LASM) (pp. 17-25).
[8] "Wikipedia: Malay Language," https://en.wikipedia.org/wiki/Malay_language, [Online; accessed 19-July-2018].
[9] "Wikipedia: Help: IPA/Malay," https://en.wikipedia.org/wiki/Help:IPA/Malay, [Online; accessed 19-July-2018].
[10] "Mari Belajar Bahasa Trengganu," http://shahrizal87.blogspot.com/2011/03/marin-belajarn-bahasan-terengganu.html, [Online accessed 19-July-2018].
[11] "Bahasa Orang Kelantan," http://kelateblog.blogspot.com/2013/07/bahasan-orangn-kelantan.html, [Online; accessed 19-July-2018].
[12] "Wikipedia: International Phonetic Alphabet," https://en.wikipedia.org/wiki/Internationaln Phoneticn Alphabet, [Online accessed 19-July-2018].
[13] Basri, S. B., Alfred, R., & On, C. K. (2012, November). Automatic spell checker for Malay blog. *In Control System, Computing and Engineering (ICCSCE*), 2012 IEEE International Conference on (pp. 506-510). IEEE.
[14] Samsudin, N., Puteh, M., Hamdan, A. R., & Nazri, M. Z. A. (2013). Normalization of Noisy Texts in Malaysian Online Reviews. *Journal of ICT*, 12, 147-159.
[15] Muhamad, N. A. B., Idris, N., & Saloot, M. A. (2017, February). *Proposal: A Hybrid Dictionary Modelling Approach for Malay Tweet Normalization*. In Journal of Physics: Conference Series (Vol. 806, No. 1, p. 012008). IOP Publishing.
[16] Dewan Bahasa dan Pustaka, (2008). Panduan Singkatan Khidmat Pesanan Ringkas. Retrieved from http://www.dbp.gov.my/khidmatsms.pdf
[17] Samsudin, N., Puteh, M., Hamdan, A. R., & Nazri, M. Z. A. (2012, July). *Normalization of Common Noisy Terms in Malaysian Online Media.* In Proceedings of the knowledge management international conference (pp. 515-520).
[18] Abdul-Rahman, S., Bakar, A. A., & Mohamed-Hussein, Z. A. (2012). An intelligent data pre-processing of complex datasets. *Intelligent Data Analysis*, 16(2), 305-325.
[19] Rodzman, S. B., Ronie, M. F. I. A., Ismail, N. K., Rahman, N. A., Ahmad, F., & Nor, Z. M. (2018, March). Analyzing Malay Stemmer Performance Towards Fuzzy Logic Ranking Function on Malay Text Corpus. In 2018 Fourth *International Conference on Information Retrieval and Knowledge Management (CAMP)* (pp. 1-6). IEEE.
[20] Nejja, M., & Yousfi, A. (2015). The context in automatic spell correction. Procedia Computer Science, 73, 109-114.
[21] Fan, W., & Gordon, M. D. (2014). *The power of social media analytics.* Commun. Acm, 57(6), 74-81.
[22] Flint, E., Ford, E., Thomas, O., Caines, A., & Buttery, P. (2017, September). *A text normalisation system for non-standard English words.* In Proceedings of the 3rd Workshop on Noisy User-generated Text (pp. 107-115).
[23] Mosquera, A., Gutiérrez, Y., & Moreda, P. (2017). *On evaluating the contribution of text normalisation techniques to sentiment analysis on informal web 2.0 texts.* Procesamiento del Lenguaje Natural, 58, 29-36.
[24] Blum, R., Liutikas, A., Ainslie, A., & Simpson, R. (2016). Automatic Detection of User Language.
[25] Mohammed, N., & Abdellah, Y. (2018). *The vocabulary and the morphology in spell checker*. Procedia Computer Science, 127, 76-81.