International Journal on Advanced Science Engineering Information Technology

Vol.10 (2020) No. 4
ISSN: 2088-5334

Categorization of Malay Social Media Text and Normalization of Spelling Variations and Vowel-less Words

Ruhaila Maskat[a,1], Nurazzah Abdul Rahman[#2]

[a] *Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia*
E-mail: [1]ruhaila tmsk.uitm.edu.my; [2]nurazzah@tmsk.uitm.edu.my

***Abstract*— As more data are being introduced, it brings along with it missing values, inconsistencies, and heterogeneities, or so-called unclean aspects. Text analytics relies on clean data to produce reliable results. Pre-processing is an essential phase in text analytics, specifically language detection and normalization. The problem with conducting text analytics on Malay social media text is how substantially it has transformed from formal Malay in terms of spelling and construction, making it difficult to process them. Recent advances have shown works to normalize yet cherry-picked specific types of Malay social media text where their descriptions were listed in simple and narrow categorizations. A formal categorization is necessary to provide significant description of the different patterns of Malay social media text, allowing the selection of suitable methods in handling them. In this paper, we propose an inexhaustive formal categorization for Malay social media text based on inherent nature. We refer to them as Social Media Malay Language (SMML) to differentiate them from the standard Malay language. They are *spelling variations*, *Malay-English mix sentences*, *loan words/phrases*, *slang-based words*, and *vowel-less words*. Also, in this work, we conducted a normalization on two of the SMML categories, spelling variations, and vowel-less words, using two similarity matching techniques (i.e., nGram Tversky Index and Levenshtein). Our result shows that similarity-matching techniques can detect both categories, but a more sophisticated technique is necessary to improve the precision score. The normalization of the rest of the categories is extensive research works.**

***Keywords*— text analytics; social media; data pre-processing; normalization; malay language.**

I. INTRODUCTION

Pre-processing or data cleaning is the longest phase in any data analytics cycle; text analytics included [1]. Besides dealing with missing values, duplicate reviews, unwanted foreign language reviews, and handling typos [2], cleaning in text analytics has two essential steps. They are the *detection of language* and the *normalization of text*.

Malay is a language spoken by around 290 million people worldwide [3], with 12.75 million owning a social media account [4]. Malay speakers can be found living in Malaysia, Brunei, Singapore, and Indonesia [3]. More organizations, profit and non-profit alike, are showing a growing interest in analyzing textual data from social media. It promises great rewards in applications such as customer intelligence [5] and smart city [6].

The problem with conducting text analytics on Malay social media text is how substantially it has transformed from formal Malay in terms of spelling and construction, making it difficult to process them. From our observation, we identified several traits of this new transformation and proposed a categorization. We simply refer to them as Social Media Malay Language (SMML), differentiating them from the standard Malay language curated by the Dewan Bahasa and Pustaka (DBP). DBP is an authorized body responsible for regulating and standardizing spelling and usage of the Malay language. The traits are as follows:

- the use of various forms of spelling (Category 1)
- mixing of Malay and English words in a sentence (Category 2)
- writing English words or phrases spelled using Malay phonology (Category 3)
- spelling Malay words based on regional slang (Category 4)
- spelling words without the use of vowels (Category 5).

A formal categorization is necessary to provide a high-level understanding of the different patterns of Malay social media text, allowing the selection of suitable methods in handling them. To the best of our knowledge, no proposal for formally categorizing text in Malay social media has been done. Thus, we proposed a formal categorization of Malay social media text. This categorization is not exhaustive and shall undergo revisions as new patterns are found. In this paper, we also present a study on normalizing two of the SMML categories using commonly-used similarity matching techniques. They are spelling variant words (Category 1) and vowel-less words (Category 5).

Normalization of the rest of the categories is open for future research works.

Text analytics deals with discovering knowledge from a large collection of text. Often techniques used in text analytics can be found from Natural Language Processing (NLP), Data Mining (DM), Machine Learning (ML), and Information Retrieval (IR) [7]. Text analytics dealing with social media corpus aims at uncovering insights into social networks or groups such as sentiment analysis [8], event detection, and customer segmentation [7]. Text analytics generally consists of three primary steps [7], namely preprocessing, representation, and knowledge discovery (Figure 1). Preprocessing deals with the "cleaning" of data [9]. Often this is the most prolonged phase in any data analytics-related work [1]. The aim is to rid the corpus of white space, missing values, duplicate reviews, stop words, non-ASCII characters, and typos that could negatively affect the result of analytics [2]. Stemming [10] reduces words into their base form, where they are considered as one single feature, for example, "walking," "walked" and "walks" are stemmed to "walk." Language detection (LD) in pre-processing helps to reduce the extracted corpus size by filtering out unrelated text based on the language used [11], [12]. The use of LD is critical when the selection of tokenizers is language-dependent [11], [12]. Normalization is the task of transforming words spelled in non-standard forms to their standard forms for use in Natural Language Processing (NLP) tasks. Representation involves modeling documents by transforming them into numeric vectors such as Bag of Words (BoW) or Vector Space Model (VSM). The linguistic structure is completely ignored through this. This is to prepare the corpus for machine learning or data mining techniques to be applied. Finally, several types of knowledge discovery may take place: supervised classification, clustering, sentiment analysis, and event detection.

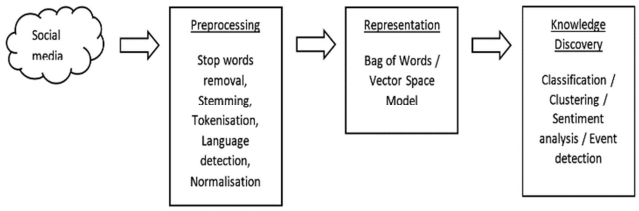

Fig. 1 General steps of text analytics [5]

From our observation, we identified several traits of Malay language used in social media. We categorized them and simply referred to them as SMML to distinguish them from the standard Malay language. SMML is unlike standard Malay, which is taught in academic institutions, used in formal written correspondences, and most importantly monitored by an authorized body for its correct spelling and use, such as Malaysia's DBP. SMML grew from the informal perception of the public or on what many language users view to be acceptable. Although the underlying principles of SMML stem from the formal Malay language, however, without any predetermined guideline by a body of authority, there is practically no right or wrong in terms of its spelling and usage, and many times, a word is left for individual interpretation. A primary influencing factor on how SMML came to adopt its new characteristics is the limited allowable text of social media. This induces users to improvise the presentation of text to fit more information in a restricted given space.

We categorize SMML into the following:

- Category 1: Spelling variants.
- Category 2: Malay-English mixed sentence.
- Category 3: Loan words/phrases.
- Category 4: Slang-based spelling.
- Category 5: Vowel-less spelling.

A. *Category 1 (Spelling variants)*

Table I shows the different forms of spelling used for the same word, with the leftmost word being the standard Malay spelling. Like most languages, the Malay language also depends on vowels to form sounds. We observed that SMML's spelling variants are influenced mainly by the informal pronunciation of native speakers. An example is the formal word of "baca" when spoken will lose the letter *a* at the end and replaced with a letter *e*, which better represents the sound that a human speaker makes.

Another example is the interchangeable use of the letters *u* and *o* due to the same reason. A more complex variant can be found when a two-word phrase is used, e.g. "macam mana", "tidak ada", "tidak hendak", "tidak mahu" and "tidak tahu". The result can be a single word which bears the sound alike the formal two-word phrase. For example, "macam mana" can be spelt as "camne", "camner", "cane." Some, like the word "cane", can be confused by a language detector to be an English word and thus will be excluded from being used in analytics.

TABLE I
SPELLING VARIANTS

baca, bace, bc
balik, blik, blk
betul, betol, btul, btol
macam mana, mcmana, camne, camner, mcmner, cane
tidak ada, takde, xde, tade
tidak hendak, tanak, tak nak, xnak
tidak mahu, tak mahu, tak mau, tamau
tidak tahu, tak tahu, tak tau, tatau

B. *Category 2 (Malay-English mixed sentence)*

The second category describes words where Malay and English can be found used in the construction of a single sentence. This sentence can abide either by the English construction rules, Malay's, or both English and Malay combined wherever deemed possible by the human writer. This may occur at the start of the sentence, middle, or end. Based on a sentence's dominant language, we consider a word is *local* if it agrees with the language, and a word is *foreign* if it is otherwise. In Table II, line 1 shows a sentence built upon Malay construction rules, and the word "improve" is a foreign word used to replace the Malay word with the same meaning.

Conversely, line 2 displays an English sentence with replacements of Malay words, i.e., "rakyat" and "untuk". Line 3 shows a changing in a sentence's tone from English to Malay as it is being read with just the adding of "ye" at the end. Line 4 shows a right combination of English and

Malay. In all these cases, a direct translation from one language to another may result in a loss of meaning.

C. Category 3 (Loan words/phrases)

The third SMML category is loan words/phrases. A "loan word" is *"a word adopted from a foreign language with little or no modification"* [13]. These SMML loan words/phrases have a unique characteristic. Often triggered by trends, they use Malay language's phonology to spell its original words or phrases that are often in English being a second formally-taught language in Malaysia. For example, "kipidap" comes from the phrase "keep it up," which was popularized by a local public figure. Another example is "brader" originally from "brother." The Malay language, although spelled using the Roman alphabet, follows a different method than English to spell words. This is due to the differences in how vowels sound in Malay in contrast with English. Additionally, vowels in Malay tend to hold a single sound, unlike English, which may have multiple sounds from the same vowel, e.g., "hurt," "bus," "busy." We propose a categorization based on the mapping of Malay to English phonemes. The phonemes are divided into sounds of consonant, vowel, and diphthong. Some may have similar-sounding words between English and Malay, while others differ. Table III lists the formal English spelling on the left and its equivalent spelling in SMML on its right, along with the type of phoneme it was categorized and if they are similar-sounding

TABLE II
MALAY-ENGLISH MIXED SENTENCE

1.	I tak tahu macamana nak ***improve***.
2.	Waive GST for sports equipment the ***rakyat*** should be encourage ***untuk*** stay healthy.
3.	Why items zero gst from all level ***ye***?
4.	Hahahaha nanti ***I*** bagi ***free lecture on gst*** untuk ***you***

TABLE III
LOAN WORDS

English	SMML	Speech sound type	Similar sound
relax	**r**ileks	Consonant	Same
s**ch**ool	s**k**ul	Consonant	Same
to**p**up	to**p**ap	Consonant	Same
bro**th**er	bra**d**er	Consonant	Different
think	**t**ink	Consonant	Different
j**eal**ous	j**el**es	Vowel	Same
h**u**sb**a**nd	h**a**sb**e**n	Vowel	Same
w**ow**	w**au**	Diphthong	Same
b**ye**	h**ai**	Diphthong	Same
b**oy**	h**oi**	Diphthong	Same

D. Category 4 (Slang-based spelling)

Table IV shows the fourth SMML category, which is slang-based spellings. Slangs tend to be region-dependent. Some regions differ slightly from the formal Malay language, while others vary considerably. This trait of SMML uses the formal Malay spelling convention to write words to the sound of a slang. We categorize them into groups of North, North (Perak), East Coast (Trengganu), East Coast (Kelantan), Central (Klang Valley), Central (N. Sembilan), South (Melaka) and South (Johor). On the left-hand side of the equivalent word pairs in Table IV are the standard Malay spellings. Since spelling in SMML differs with slang, a single word can be spelled differently depending on the writer's adopted slang. For example, the word "besar" can be written as "beso" or "besa", suggesting the former to be used by people in the southern region of West Malaysia and the latter from the central region. Rules can be constructed to describe these transformational norms.

TABLE IV
SLANG-BASED SPELLING

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

E. Category 5 (Vowel-less spelling)

The last characteristic of SMML is short-formed spelling (Table V). No vowels are used, only consonants for this type of SMML. Vowels are purposely dropped to save space and time. SMML words may also consist of a single character. This character may or may not sound like the original word. For example, the word "tidak" meaning "no" or "not" would frequently be represented with the letter *x*.

TABLE V
VOWEL-LESS SPELLING

pergi = g or p
tidak = x
jangan = jgn
yang = yg
nama = nm
pada = pd

These categories are not isolated from each other. We observed that social media authors do combine multiple categories. For example, "x suko" is a combination between Categories 4 and 5. Due to the limitation of space, further details of these categories will be covered in another article. This paper is structured like the following. Section 2 presents relevant works and discusses our experiments, hypotheses, and techniques. Our results are described in Section 3. We conclude in the final section.

II. Materials and Method

In this section, we describe related works, our corpus, and a series of experiments we conducted. These experiments aim to investigate the use of similarity-matching techniques to pre-process spelling variant (Category 1) and vowel-less (Category 5) types of SMML. Normalization on the rest of the categories is open research works.

At the point of writing, we have not seen any proposals that formally categorize Malay social media text. Narrow categorizations were found described within works to normalize the text or efforts to check and correct spelling errors automatically. Malay social media text has been labeled as misspelled words [14], out-of-vocabulary words [14], [15] ill-formed words, and noisy text [16]. A shared assumption of these works is that of the availability of a dictionary of standard Malay spellings to replace these "rogue" text. The output is a text spelled in standard Malay.

Muhamad et al. [17] discussed a conceptual architecture of a Malay text normalization framework of a work-in-progress. This work utilized a hybrid dictionary approach and identified three types of noisy Malay Twitter messages. They are colloquial language, novel words, and interjections into standard Malay language. A language model is used along with n-gram at the heart of the normalization process.

Saloot et al. [16] presented an unsupervised normalization system. The first phase of the system involves generating candidate words using six different methods, while the second phase makes a language model probability score of each candidate. The highest score will be used to replace the noisy text. The six candidate generation methods are: producing probable phonemes, one-edit distance, two-edit distance, Malay-to-English translation, and heuristic rules.

In another related work by Saloot et al. [18], a narrow categorization was presented, consisting of repeated-lettered words, words not found in a predefined vocabulary which could have been misspelled, words with special characters added with and abbreviated words.

Samsudin et al. [15] constructed a set of rules capable of automatically-generating artificial noisy text. These rules were based on an earlier work by DBP [19] also authorized to produce a guideline on how Short Message Service (SMS) text should be used in official correspondences as well as on TV channels in Malaysia. This was an effort to streamline SMS messages. At the time of writing, we do not see extensive use of this guideline due to the decline in SMS usage.

Basri et al. [14] proposed an automatic spell-checker and corrector of misspelled words. A dictionary is maintained of wrongly-spelled words used during spelling correction, and newly-found misspelled words are later added to it. In the instance that a misspelled word cannot be matched with any words in the dictionary, the Levenshtein coefficient is utilized to find possible candidates. In this work, only Selangor slang words are handled. Additionally, English words are removed, leaving only non-English words.

A. Corpus

Our corpus has 6,269 reviews from Twitter covering over four years, from 2014 to 2018. After duplicate removal, the size is 6,241. These duplicates came from retweets. The corpus consists of mostly Malay reviews with SMML embedded and a handful of full English reviews. As a baseline, we would like to know how well the Malay language can be detected when SMML is present. We use Tika Language Detector on our corpus and discover that Malay reviews are also detected as Indonesian, as reported by Ranaivo-Malacon [20]. Table VI shows the statistics of our corpus.

TABLE VI
Languages Detected

Language	Mixed language	Number of reviews	Percentage
Malay (ms)	[ms]	1,490	23.8
	[ms, id]	535	8.6
	[ms, en]	54	0.86
	[ms, others]	352	5.6
		2,431	39
English (en)	[en]	819	13
	[en, id]	48	0.77
	[en, others]	253	4.1
		1,120	18
Indonesia (id)	[id]	1,721	27.6
	[id, others]	315	5
		2,036	33
Undefined	-	102	1.6
Others	-	552	8.8

Several interesting points can be found here. Firstly, Malay tweets are constantly being categorized as Indonesian due to the close similarity in their features; for example, the same form of prefixes and suffixes (ke-an, me-an, pe-an) are used. So similar they are, differentiating Malay and Indonesian has become a research area along with other closely-similar languages such as European Spanish and Mexican Spanish and Portuguese dialects are spoken in Europe and Brazil [21]. These false negatives ([id], [id,others], [en,id]) sums up to a significant amount of 2,619 (42%). Next, tweets from Malaysia would either use complete English or a mix of English and Malay. This mixing of language, also known as code-mixing [22], was categorized as [ms,en], [en,id] and [en,others], totaling to 355 (5.7%). Others consist of reviews which have been detected as having written in languages other than Malay, Indonesian or English. This includes Tagalog (80 reviews), Somali (54 reviews), German (25 reviews), and Norwegian (18) at the top. Inclusive of other languages, this totals to 552 (8.8%). A sum of 102 reviews (1.6%) has been categorized as Undefined where the language cannot be determined. Both Undefined and Others make up 654 of

total reviews (10.5%) and the size of the Bag of Words (BoW) is 1,576.

We highlight the following pre-processing issues.

- Language detection: Malay and Indonesian reviews need to be correctly differentiated.
- Normalization: English and Malay reviews must be normalized to only a single language.
- Normalization: Spelling variant words, slang-based words, and vowel-less words must be normalized to standard Malay spelling.
- Normalization: Malay-spelling English words must be normalized to standard English and then to standard Malay since often these words are used in Malay-constructed sentences.

Our experiment addresses the normalizing of SMML's spelling variant and vowel-less words. We directed our investigation on reviews under the Undefined and Other categories where no language matches the former (i.e., Undefined), and not even one review was identified as Malay in the latter (i.e., Others). Reviews successfully categorized as either [ms] or [id] or any of their combinations indicate true positives since Indonesian and Malay share similar features. Considering that our goal is not to differentiate them, they are excluded from our investigation. English reviews, [en] and [en, others], consisted of code-mixing text and were also excluded.

B. Ground Truth

From the Undefined and Others subsets, we have identified and manually annotated similar words by the service of a native Malay human speaker. The ground truth set size is 1,176 pairs of similar words. We constructed a similar bi-directional pair, for example, A = B and B = A to capture all permutations. Each pair is labeled with Similarity = Y. Words are of different lengths. The ground truth consists of 435 unique words without any stemming conducted. Stemming is unnecessary as we wish to keep the actual meaning intact and simply identify the different forms of spelling a word may have. This approach preserves the context of the review for further downstream analysis. Table VII shows the details.

TABLE VII
HUMAN-ANNOTATED GROUND TRUTH

Similar word pairs	1,176
Unique words	435

C. Research Question and Hypothesis

1) Research question: Can SMML spelling variant (Category 1) and vowel-less (Category 5) words be normalized using similarity-based techniques?

2) Hypothesis: Spelling variant and vowel-less words have characters that can be different from their standard spelling by varying degrees. Hence, we view normalizing them as the problem of finding the similarity between two words. The more similar the words are, the lesser their distance between one another. The goal is to find a set of optimum parameters of commonly-used similarity matching techniques that produce a good performance.

D. Techniques

Similarity matching techniques generate measurements of similarity normalized between 0 and 1. If the distance is the underlying concept, then 0 indicates an exact match. On the other hand, if the similarity is chosen, then the exact match would be represented by 1. In this work, we adopted the use of distance. A distance or similarity threshold specifies the value of what is similar and what is not. Hence, anything above (distance-based) or below (similarity-based) the threshold is identified as dissimilar. We selected three thresholds to represent a varying strictness of similarity. From strict (0.3), medium (0.5) and lenient (0.8). In this work, we applied commonly-used similarity matching techniques, n-gram Tversky Index and Levenshtein.

1) n-gram Tversky Index: n-gram is substrings of the length n. Typical gram sizes are 2 (bi-gram), 3 (tri-gram), 4 (four-gram) and 5 (five-gram). The underlying concept of using n-gram in similarity matching is if two strings X and Y are similar; therefore, there should be an overlap of n-grams between them [23]. The Tversky Index [24] measures the asymmetric similarity between a variant to a prototype. The use of $\alpha = \beta = 1$ will produce the Tanimoto coefficient, while $\alpha = \beta = 0.5$ will produce Dice's coefficient. α represents the weight of the prototype, and β corresponds to the variant's weight. Tanimoto coefficient looks at the intersection of two strings compared to the union of the strings. In comparison, Dice coefficient is the intersection of two strings over the average size of the strings.

$$S(X,Y) = \frac{|X \cap Y|}{|X \cap Y| + \alpha|X - Y| + \beta|Y - X|}, \tag{1}$$

2) Levenshtein Distance: Levenshtein [25] measures the distance between a pair of words. The idea is to use single-character edits, i.e., insertions, deletions, and substitutions, at its very minimum to change one word to another. The number of minimum edits required to transform describes the distance between the word pair.

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases} \tag{2}$$

Given two words *a* and *b*, the distance between the first *i* characters of *a* and the first *j* characters of *b*, $\mathrm{lev}_{a,b}(i,j)$, is 0 when there are no edits necessary. Distance is not equal to 0 when there exists either edits of deletion from *a* to *b* ($\mathrm{lev}_{a,b}(i-1, j) + 1$), insertion ($\mathrm{lev}_{a,b}(i, j-1) + 1$) or substitution ($\mathrm{lev}_{a,b}(i-1, j-1) + 1_{(ai \neq bj)}$).

E. Method and Measurement of Performance

Our focus is on parameter optimization to generate the highest performance. The parameters are distance threshold, gram size (nGram), number of neighbors, and the character length of edits (Levenshtein). To compare performance, we used a normalized Levenshtein edit. We have explained distance threshold, gram size, and character length of edits in Section III.*D*. Number of neighbors (nn) here refers to the total word pairs that should be included during a similarity search. For nn = 10, this means each word would be coupled

with 10 of its most similar neighbors. Allocating ten neighbors would produce 15,760 pairs of words with varying degrees of similarity. Optimizing nn is to seek enough neighbors to provide a constant number of true positives and false positives where they indicate the maximum performance a technique can generate. For our corpus, we found a neighbor size of 100 to be optimum. We set a distance threshold of 0.5 and tested on bi-gram, tri-gram, four-gram, and five-gram. Character lengths of 1 and 2 were tested. In the beginning, the subsets were parsed, and a BoW was formed. Any non-ASCII and diacritic characters were removed as they are meaningless. Stop words were retained, and lemmatization was not conducted to retain a word's original pattern. This is useful for any NLP tasks to be conducted later. Measurement of distance was then calculated for each word pair within the allocated neighbor size. Finally, each technique's performance is calculated using different parameters and compared.

We are interested to know how well the techniques correctly guess similar word pairs (true positive). The capability to identify word pairs that are not similar (true negative and false negative) is of no benefit. This is aligned with the actual effort of constructing a list of similar words. It is just more practical to collect similar word pairs than dissimilar ones, which are often substantially more in numbers. Hence, our ground truth set of human-annotated word pairs consists of only similar words. Besides these, there will be words that are incorrectly determined as similar, yet they are not (false positives). In this work, we employ Precision to measure how correctly identical words can be detected. Accuracy is not a suitable measurement since it assumes that the set of negatives is necessary to be known.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (3)$$

III. Results and Discussion

We obtained the following results:

- Precision across all of the chosen techniques is below 0.5. This indicates the spelling variant and vowel-less nature of SMML requires a more sophisticated approach than commonly-used similarity matching techniques to be better detected.
- Overall, a distance threshold of 0.3 presents the highest precision across all the techniques as compared to thresholds 0.5 and 0.8. This indicates that although a smaller number of data was chosen due to the strict threshold, it consists of a larger portion of true positives.
- The highest precision is 0.429 where the parameters are threshold = 0.3 and gram = 4. The lowest precision is 0.007 with threshold = 0.8, gram = 2 and Levenshtein edit = 1.
- Between grams, we see that the longer the gram, the better the precision. We observed that 2-gram has more related words because SMML is a largely short text; however, it contains a small number of true positives. Conversely, few SMML words reach up to 5 characters; hence, 5-gram has a lesser number of related words, yet contains more true positives. Precision can be improved if the number of false positives in 2-gram can be reduced.
- 2-gram and 4-gram produce stable precision for both Tanimoto and Dice coefficients with 4-gram producing better precision. 3-gram in Tanimoto does not produce an interesting result, while 5-gram risks producing 0 precision due to its unusual length in SMML.
- Tanimoto coefficient with gram = 5 and threshold = 0.3 yields 0 precision. This is due to a strict threshold until to the point where no data could meet the condition

In summary, from our study, we learned that similarity matching techniques could detect spelling variant and vowel-less SMML. However, a more sophisticated technique is necessary to attain an improved precision.

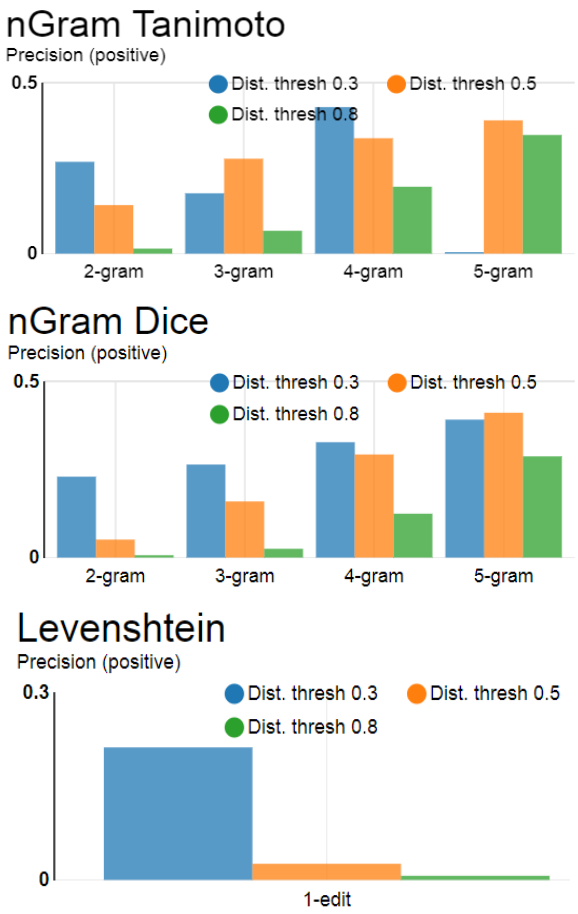

Fig. 2 Distance threshold

TABLE VIII
Human-annotated Ground Truth

nGram Tanimoto - Precision (positive)	2-gram	3-gram	4-gram	5-gram
0.3	0.269	0.176	0.429	0
0.5	0.142	0.278	0.196	0.39
0.8	0.015	0.067	0.39	0.347

nGram Dice - Precision	2-gram	3-gram	4-gram	5-gram

(positive)				
0.3	0.23	0.264	0.327	0.391
0.5	0.051	0.159	0.292	0.411
0.8	0.007	0.025	0.124	0.287

Levenshtein -Precision (positive)	1-edit
0.3	0.212
0.5	0.026
0.8	0.007

IV. CONCLUSION

In conclusion, we presented in this paper an inexhaustive formal categorization of SMML. We proposed the categories with reference to their inherent nature. The first category is of the use of various forms of spelling (spelling variant words). The second category is the mixing of Malay and English words in a sentence (Malay-English mix words). The third category is writing English words using Malay phonology (loan words/phrases). The fourth category is spelling Malay words based on regional slang (slang-based words). The final category is spelling words without the use of vowels (vowel-less words). We tested using commonly-used similarity-matching techniques to try to normalize spelling variant and vowel-less words. Results obtained showed that these techniques could produce good precision, but a more sophisticated technique is required. The normalization of the rest of the categories is open research work.

ACKNOWLEDGMENT

We are extremely grateful to the reviewers who have taken the time to give constructive comments and useful suggestions for the improvement of this article. The Malaysia Government supports this work under the Fundamental Research Grant Scheme (FRGS) at Universiti Teknologi MARA (UiTM) Shah Alam, Malaysia (FRGS/1/2015/ICT01/UiTM/03/01).

REFERENCES

[1] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," *Procedia Comput. Sci.*, vol. 17, pp. 26–32, 2013.

[2] L. Derczynski, D. Maynard, N. Aswani, and K. Bontcheva, "Microblog-genre noise and impact on semantic annotation accuracy," in *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, 2013, pp. 21–30.

[3] "Malay Language," *Encyclopedia Britannica*. [Online]. Available: https://www.britannica.com/topic/Malay-language.

[4] Statista, "Number of Facebook users in Malaysia from 2017 to 2023." [Online]. Available: https://www.statista.com/statistics/490484/number-of-malaysia-facebook-users/ .

[5] N. Elgendy and A. Elragal, "Big data analytics: a literature review paper," in *Industrial Conference on Data Mining*, 2014, pp. 214–227.

[6] R. Kitchin, "The real-time city? Big data and smart urbanism," *GeoJournal*, vol. 79, no. 1, pp. 1–14, 2014.

[7] X. Hu and H. Liu, "Text analytics in social media," in *Mining text data*, Springer, 2012, pp. 385–414.

[8] N. N. Yusof, A. Mohamed, and S. Abdul-Rahman, "Reviewing classification approaches in sentiment analysis," in *International conference on soft computing in data science*, 2015, pp. 43–53.

[9] S. Abdul-Rahman, A. A. Bakar, and Z.-A. Mohamed-Hussein, "An intelligent data pre-processing of complex datasets," *Intell. Data Anal.*, vol. 16, no. 2, pp. 305–325, 2012.

[10] S. B. Rodzman, M. F. I. A. Ronie, N. K. Ismail, N. A. Rahman, F. Ahmad, and Z. M. Nor, "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)*, 2018, pp. 1–6.

[11] I. Balazevic, M. Braun, and K.-R. Müller, "Language Detection For Short Text Messages In Social Media," *arXiv Prepr. arXiv1608.08515*, 2016.

[12] M. Lui and T. Baldwin, "Accurate language identification of twitter messages," in *Proceedings of the 5th workshop on language analysis for social media (LASM)*, 2014, pp. 17–25.

[13] "Loanword," *Lexico*. [Online]. Available: https://en.oxforddictionaries.com/definition/loanword.

[14] S. B. Basri, R. Alfred, and C. K. On, "Automatic spell checker for Malay blog," in *2012 IEEE International Conference on Control System, Computing and Engineering*, 2012, pp. 506–510.

[15] N. Samsudin, M. Puteh, A. R. Hamdan, and M. Z. A. Nazri, "Normalization of noisy texts in Malaysian online reviews," *J. ICT*, vol. 12, pp. 147–159, 2013.

[16] M. A. Saloot, N. Idris, and A. Aw, "Noisy text normalization using an enhanced language model," in *Proceedings of the International Conference on Artificial Intelligence and Pattern Recognition*, 2014, pp. 111–122.

[17] N. A. B. Muhamad, N. Idris, and M. A. Saloot, "Proposal: A Hybrid Dictionary Modelling Approach for Malay Tweet Normalization," in *Journal of Physics: Conference Series*, 2017, vol. 806, no. 1, p. 12008.

[18] M. A. Saloot, N. Idris, and R. Mahmud, "An architecture for Malay Tweet normalization," *Inf. Process. Manag.*, vol. 50, no. 5, pp. 621–633, 2014.

[19] "Panduan singkatan khidmat pesanan ringkas," *Dewan Bahasa dan Pustaka*. [Online]. Available: http://www.dbp.gov.my/khidmatsms.pdf.

[20] R.-M. Bali and N. P. Kuan, "Language Identifier for Bahasa Malaysia and Bahasa Indonesia."

[21] J. Williams and C. Dagli, "Twitter language identification of similar languages and dialects without ground truth," in *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 2017, pp. 73–83.

[22] M. Puteh, N. Isa, S. Puteh, and N. A. Redzuan, "Sentiment mining of Malay newspaper (SAMNews) using artificial immune system," in *Proceedings of the World Congress on Engineering*, 2013, vol. 3, pp. 1498–1503.

[23] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate record detection: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 1, pp. 1–16, 2006.

[24] A. Tversky, "Features of similarity.," *Psychol. Rev.*, vol. 84, no. 4, p. 327, 1977.

[25] L. Yujian and L. Bo, "A normalized Levenshtein distance metric," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1091–1095, 2007.