SOCIAL MEDIA ANALYTICS USING SENTIMENT AND CONTENT ANALYSES ON THE 2018 MALAYSIA'S GENERAL ELECTION

Vimala Balakrishnan[1*], Mohammed Kaity[2], Hajar Abdul Rahim[3] and Nazari Ismail[4]

[1,2]Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia

[3]School of Humanities, Universiti Sains Malaysia, Penang, Malaysia

[4]Faculty of Business and Accounting, Universiti Malaya, Kuala Lumpur, Malaysia

Email: vimala.balakrishnan@um.edu.my[1*](corresponding author), moh.kaity@gmail.com[2], hajar@usm.edu.my[3], nazari.ismail@gmail.com[4]

DOI: https://doi.org/10.22452/mjcs.vol34no2.3

ABSTRACT

This study analysed the political use of Twitter during the 2018 Malaysian General Election (GE14), using sentiment and content analyses to examine the patterns in online communication among urban Malaysians. Specifically, Naive Bayes, Support Vector Machine and Random Forest were used for sentiment analysis for the English tweets, with the results compared against two vectorization approaches. Content analysis involving human experts was used for the Malay tweets. Top trending hashtags were used to fetch tweets from April 15, 2018 to May 14, 2018, resulting in a curated corpus of 190 224 tweets. Naïve Bayes used along with Word2Vec outperformed all the other models with an accuracy of 63.7%, 66.8% and 64.9% for pre-GE14, GE14 and post-GE14, respectively. Generally, results indicate the majority of the sentiments to be positive in nature, followed by negative and neutral during pre-GE14, GE14 day and post-GE14 for the English speakers. Though similar sentiments were observed for the Malay speakers, the majority of their sentiments on election day were negative (i.e. 42%) as opposed to the English speakers (i.e. 31%).

Keywords: Social media analytics; Sentiment analysis; Content analysis; Election; Malaysia; Twitter

1.0 INTRODUCTION

Social media such as Facebook and Twitter play a pivotal role in electoral campaigns, reflecting information about policy preferences and opinions of politicians and their followers [1]. Twitter currently has over 500 million users, of which 328 million are monthly active users [2], reflecting its global outreach and potential impact. The micro-blogging platform enables opinion sharing using various content forms such as text, images, links etc., while allowing unidirectional and asymmetrical connections, where users can connect with other users without their approval. Therefore, a Twitter user's network contains not only family and friends, but also connections with thought leaders, influencers, celebrities and political leaders, among others [3].

Major events such as politics or elections attract a lot of social media attention, indicating that the platform can be an important game changer [1]. Some of the biggest and well-known case studies focusing on Twitter's role as a political driving factor are Barack Obama's and Donald Trump's presidential elections [4], [5], [6]. Other notable studies include those who investigated public perceptions of politicians [7], [8], [9], use of social media for political knowledge [10], [11], [12], presidential predictions [13], [14], [15], [16], [17], and impact of anonymity-preference and fear of isolation in South Korea [18], among others.

In recent years, studies have begun exploring social media analytics such as sentiment analysis to improve electoral predictions based on the number of seats [13], [9], [15], [16], [14]. For instance, [13] proposed a new seat forecasting method using social media analytics, and correctly predicted the winner of the 2017 Punjab assembly elections. Others such as [7] and [1] used sentiment analysis to gauge public perceptions of politicians, hence lending support that sentiment plays an important role in information diffusion on social media. For instance, [1] determined the sentiments and tweet topics by Donald Trump and Hillary Clinton during the 2016 US presidential campaign using SentiStrength (i.e. a free tool for text analytics), with results indicating the former tweeted more optimistically and positively than the latter. On the other hand, [7] identified sentiments of voters during the 2012 US presidential campaign using the NRC-Canada system (i.e. a system to detect sentiments of short informal texts), along with their emotions, styles and purposes.

Drawing inspiration from above, this study aims to analyse Malaysians' perceptions (i.e. sentiment) in the political discourse that took place on Twitter during the most recent General Election (GE14), which resulted in the long-governing Barisan Nasional (BN) government to be shockingly ousted by the opposition, a coalition of four political parties, Pakatan Harapan (Alliance of Hope). Specifically, the paper explores the role of Twitter based on textual communications using social media analytics, namely, sentiment analysis, content analysis and descriptive statistics based on tweets written in English and Bahasa Malaysia (i.e. the national language, herein referred to as the Malay language).

1.1. Research Gaps and Contributions

A vast majority of the previous studies have focused on specific political parties or politicians (e.g. [1] on Donald Trump and Hillary Clinton; [15] on Labour, Conservative, Sinn Fein etc.; [13] on Indian political parties such as Bharatiya Janata Party, Aam Aadmi Party and Indian National Congress), with recent studies focusing on the political participation and engagement among the users and voters [18], [9], [8]. As the active participation of the general public can potentially affect the favourability of political candidates, and thus election outcomes [19], the present study therefore focuses on both the Malaysian Twitter users and voters.

Additionally, most of the studies on the political uses of social media have examined English language tweets; except for a few that have considered multi-languages [20], [9], [13]. For instance, [13] gauged Twitter users sentiment during the 2017 Punjab assembly elections focusing on both English and Punjabi tweets; however, the latter were manually translated into English for sentiment analysis. Closer to home, [9] examined how the English and Malay speakers differ in the manner in which they used Twitter during the 2013 Malaysian General Election (GE13). The authors however, focused on the roles (i.e. public, media, journalist etc.) and functions (i.e. purpose of the tweet), with results indicating the Malay speakers to more likely use Twitter to seek political information and express their political opinions compared to the English speakers.

Finally, literature revealed most of the political Twitter sentiments were determined using existing tools and software such as SentiStrength [1], NRC-Canada system [7] and software [15], with limited studies on machine learning algorithms [13], [21]. For example, Support Vector Machine (SVM) with String2Vec was used to determine Twitter users sentiment polarity with a reported accuracy of 78.6% in [13] whereas [21] used Random Forest to investigate Italian Twitter users voting intentions in 2016. The authors found the majority of the users to be polarized toward no (48%), followed by neutral (27%) and yes (25%).

In light of the gaps above, the study aims to perform an in-depth analysis using social media analytics (i.e. Twitter analytics) in several ways, namely:

- The tweets investigated were of the two major languages used in the country, that is, English and Malay,
- The use of several machine learning algorithms (Naïve Bayes, SVM and Random Forest) with two word embedding approaches (Word2Vec and String2Vec) for the English tweets,
- The use of content analysis to determine the sentiments of the Malay tweets,
- The identification of top trending keywords for both the languages, analysed from an overall perspective and also based on their sentiments, and
- All the analyses were temporal-based, namely, pre-election (pre-GE14), election day (GE14) and post-election (post-GE14).

The rest of the paper is structured as follows: Section 2 provides the background of the study focusing on Malaysia, its political landscape and social media use. This is followed by the research methodology in Section 3. The findings are presented and discussed in the subsequent section, before concluding the paper in Section 5.

2.0 BACKGROUND

2.1. Malaysia

Malaysia is a multi-ethnic nation comprising of Malays, Chinese, Indians and others. Although Bahasa Malaysia is the official national language, English is dominantly used to communicate as well. The English speakers tend to be bi-lingual and they are not limited to an ethnic group. It is a common understanding that the English speakers belong to higher socioeconomic statuses (i.e. in terms of better education, income level etc.) compared to the Malay speakers [22]. As a matter of fact, English speaking Malaysians are often considered arrogant, boastful, too westernized and not part of the traditional community [23]. Over the past decade, there has been a notable increase in English communication (both offline and online) among Malaysians, specifically among the Millenials. Apart from English and Malay, other spoken and written languages include Mandarin, Cantonese, Tamil, Punjabi etc., though these are ethnic-specific.

2.2. Social Media Penetration

According to the most recent Malaysian Communications and Multimedia Commission (MCMC) survey (2018), the Internet penetration among Malaysians was determined to be 87.4% in 2018 compared to 76.9% in 2016 [24]. Not surprisingly, the young adults between 20 and 34 years old ranked as the highest users (53.6%), followed by those between 35 and 49 (24.7%). Although Facebook emerged as the most popular social media platform (97.3%), Twitter users ranked at 23.8%. A vast majority of the younger users (61.8%) were found to actively share content online, mostly via social media (73.8%) and group messaging platforms such as Whatsapp (70.6%). According to the survey, educational materials were the most shared content (71.3%), followed by entertainment and news (63.9%). In a similar vein, [25] found 61% of their university student participants used social media for sharing, with an average of 45% for political discussion.

The previous GE13 marked the beginning of the powerful influence of social media in political campaigns in Malaysia; with the then Prime Minister Najib Razak's Twitter account to be the most popular [26]. In fact, GE13 is interestingly known as the 'social media election' as the platform was crucial in mobilizing Malaysia's record turnout of 85% voters, and was actively used by both the ruling coalition and the opposition for campaigning and citizen outreach [26]. Thereafter, more politicians followed suit, and today almost all of the politicians representing the major parties, both old and young alike are on Twitter, with the popular ones including Najib Razak, Tun Mahathir Mohammed, Anwar Ibrahim and Khairy Jamaludin, among others.

3.0 METHODOLOGY

Fig. 1 depicts the overall methodology adopted in this study, beginning with the data collection up to the model evaluation. All the seven steps involved are elaborated in the subsequent sections.

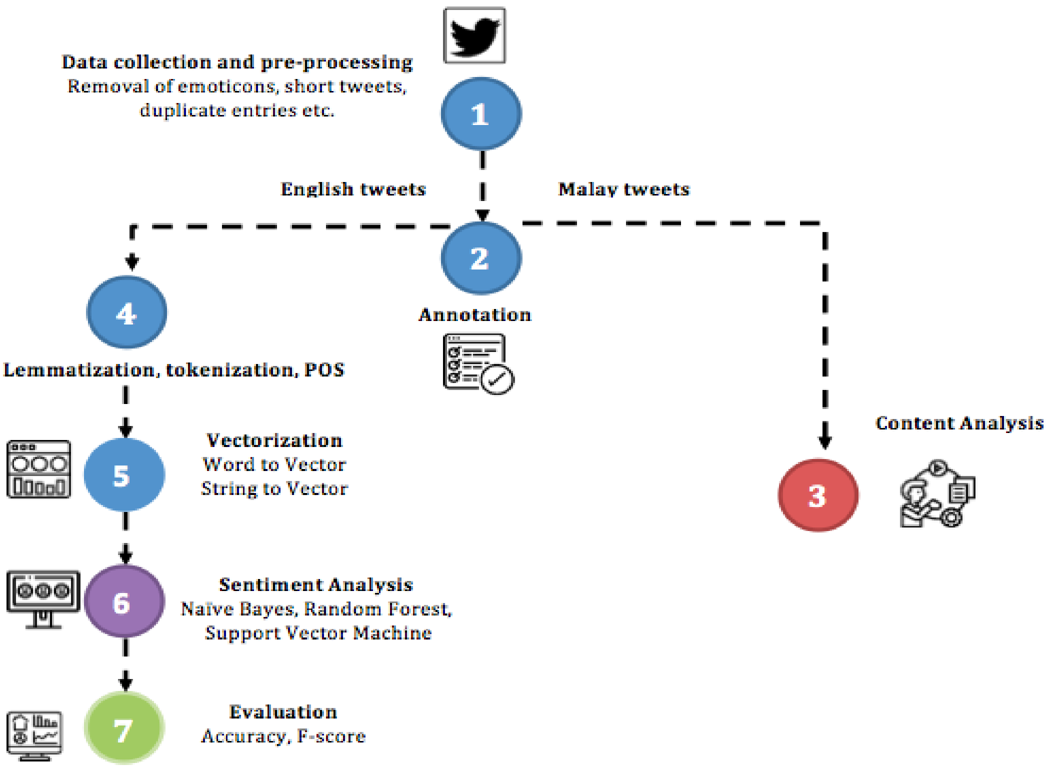


Fig 1: Overall methodology

3.1. Step 1 - Election Dataset

The dataset comprising of tweets were fetched between April 15, 2018 and May 14, 2018 beginning from the nomination date (i.e. April 15), up to the election date (i.e. May 9), and five days post-election. Approximately 311 307 (i.e. 55% English versus 45% Malay) tweets were gathered using top trending political hashtags that were monitored throughout the observed period (i.e. #GE14, #PRU14, #CarpoolGE14, #PulangMengundi, #UndiRabu, #Inikalilah, #MalaysiaMemilih).

A Python script was developed to fetch the Twitter data, consisting of tweets, retweets (i.e. a form of endorsement by sharing the tweet with others), mentions (e.g. specifying someone directly, @JohnDoe), responses to tweets, user locations, number of followers and following, language, date and time.

3.2. Step 1 - Pre-Processing and Tweet Statistics

Irrelevant details (i.e. noise) need to be removed from the dataset in order to increase the accuracy of the analysis [27]. Several criteria were used to filter and clean the dataset, namely:

- Tweets in English and Malay only (tweets containing fewer than three Malay words were considered as English, and vice-versa). Tweets in other languages were discarded.
- Removal of emojis, emoticons, special characters (@,!, &) (removed as the current study focused on textual analysis) and URLSs.
- Removal of short tweets (i.e. fewer than three words).
- Removal of tweets and retweets from the official news media (e.g. Astro Awani, Berita Harian etc.), as the majority of such tweets contained only official announcements.
- Removal of duplicate entries.

The data pre-processing resulted in a total of 190 324 tweets (i.e. 120 766 English; 69 558 Malay). Table 1 depicts the general breakdown of the tweets gathered. Generally, the number of retweets outnumbered the original tweets, however, this is not an uncommon phenomenon in Twitter where the majority of its users tend to repost tweets that they deem important or relevant [9], [13]. A sampling of approximately 15% tweets for further analyses resulted in a total of 27 839 (i.e. 19 699 English tweets versus 8140 Malay tweets).

Table 1: Datasets for English and Malay tweets

	Before cleaning		After cleaning		Final tweets (Sampling)	
Language	English	Malay	English	Malay	English	Malay
Tweets	21 831	63 763	33 549	28 029	7 966	4 839
Retweets	150 294	80 557	87 217	41 529	11 733	3 301
Total	172 125	144 320	120 766	69 558	19 699	8 140

3.3. Step 2 - Data annotation

The Malay tweets were then sent to human experts for annotation (further elaborated in Section 3.4). A total of 10K English tweets (i.e. approximately 3 300 for pre-GE14, GE14 and post-GE14) were randomly selected for human annotation and further pre-processing. Three linguistic experts were recruited to manually label the tweets as positive, neutral and negative. A sample annotation was also given to them as a guide. The annotation by the third expert was only referred to in conflicting cases (i.e. Expert 1 = positive; Expert 2 = negative). There were no cases with three different labels (i.e. Expert 1 = positive; Expert 2 = negative; Expert 3 = neutral). A small token of appreciation was provided to the experts.

3.4. Steps 4 and 5 - Additional pre-processing and vectorization

Textual data need to be further prepared for machine learning algorithms; therefore steps 3 and 4 (Fig. 1) were administered for the English tweets. This includes the removal of stop words (e.g. and, is, this etc.), followed by tokenization, which splits a given text into smaller fragments or tokens. For example, the text "BN Wins" is converted into "BN" and "Wins". Stemming (i.e. removing suffixes or prefixes: e.g. electing - elect) and lemmatization were also performed, followed by Part of Speech (POS) tagging, which tags each word to its grammatical category (e.g. noun, verb, adjectives etc.) [27].

Finally, in order to transform the text into its numerical representations, two word embedding approaches were examined, namely, Word2Vec and String2Vec. Word embedding represents words using vectors, by taking care of the semantic relationship between words, and ensures that a word that coexists more frequently is closer in the vector representation.

3.5. Step 6 - Sentiment analysis

Sentiment analysis, a discipline that extracts people's feelings, opinions, thoughts and behaviours from texts [28], is being widely used in various domains to gauge people's sentiment toward a topic or event, including politics and elections [15], [17], [7], [13]). In the present study, sentiment analysis was performed on the English tweets using several supervised machine learning algorithms (i.e. using labelled data), namely, Naïve Bayes, Support Vector Machine (SVM) and Random Forest. These algorithms are known to be popular in solving various classification problems [13], [29], [30]. For instance, SVM is deemed to be one of the most popular supervised machine learning algorithms in sentiment analysis due to its high accuracy and ability to handle large datasets compared to others [31]. All the algorithms were trained and tested using two word embedding approaches, namely, Word2Vec and String2Vec. The dataset was split into 90-10 (i.e. 90% - training versus 10% - testing). All the executions were accomplished using Python.

3.6. Step 7 - Evaluation

Finally, the performance of all the three algorithms were evaluated using two metrics, namely, accuracy (i.e. the number of instances correctly predicted) and F-score (i.e. harmonic mean of precision and recall), both of which are represented by Eq. (1) and (2) below, respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

$$F - score = \frac{2TP}{2TP + FP + FN} \qquad (2)$$

where FN, FP, TP and TN are the values of false negatives, false positives, true positives and true negatives, respectively.

As F-score takes both recall and precision into consideration, these metrics were not used to evaluate the algorithms separately [33]. F-score is deemed more appropriate to evaluate the performance of the classifications as it presents the weighted average of precision and recall [34], [35]. Higher accuracies and F-scores indicate better sentiment classifications.

Table 2: Sentiment analysis model evaluations

Models	Metrics	Pre-GE14		GE14		Post-GE14	
		Word2vec	String2Vec	Word2vec	String2Vec	Word2vec	String2Vec
Naïve Bayes	Accuracy	63.74	57.14	66.79	59.78	64.87	57.25
	F Score	61.93	53.18	60.68	52.31	59.85	50.52
		Word2vec	String2Vec	Word2vec	String2Vec	Word2vec	String2Vec
Support Vector Machine	Accuracy	58.52	52.34	58.30	51.75	56.88	51.08
	F Score	58.35	50.08	56.26	51.11	57.46	49.90
		Word2vec	String2Vec	Word2vec	String2Vec	Word2vec	String2Vec
Random Forest	Accuracy	58.52	54.27	59.04	54.52	58.55	53.88
	F Score	57.82	51.52	56.07	52.24	58.39	52.08

Table 2 depicts the results of the sentiment analysis for the English tweets. It can be observed that Naïve Bayes using Word2Vec outperformed the rest of the classification models in all three timelines, with an average accuracy of 65.13%. In fact, similar patterns were observed for all the models whereby the implementation of Word2Vec produced better accuracies and F-scores, regardless of the timelines. On the other hand, both Support Vector Machine and Random Forest did similarly well.

3.7. Step 3 - Content Analysis

In order to gauge the sentiment of the Malay speakers, a content analysis approach was adopted considering no sentiment analysis tool is available for the Malay language. In this approach, approximately 8 000 tweets (i.e. 2 600 randomly selected covering the three timelines) were provided to two graduate assistants fluent in the Malay language. Similar with the English annotation, a sample annotation was provided. Further, the assistants were also asked to identify emerging themes from the tweets, such as information, instigation, racism or sarcasm. The measure of agreement (i.e. inter coder's reliability) was determined using Krippendorff's alpha (i.e. $\alpha = 0.83$). As no machine learning algorithms were used for the Malay tweets, hence no accuracy measures are reported.

4.0 RESULTS AND DISCUSSION

The results of the sentiment and content analyses are presented in this section, beginning with the top topics identified, without taking sentiments into consideration, followed by Malaysians' perceptions based on their sentiments during pre-GE14, GE14 and post-GE14. Finally, top topics for each of these sentiments are presented. It is to note that all the results of the sentiment analysis are based on the best performing classification model, that is, Naïve Bayes with Word2Vec.

4.1. Top Trending Keywords for English and Malay Speakers

Table 3 and Table 4 show the top keywords for the English and Malay speakers, respectively, without taking their sentiments into consideration.

Table 3: Top keywords for the English speakers

Pre-GE14			GE14			Post-GE14		
Keyword	N	%	Keyword	N	%	Keyword	N	%
Malaysia	680	15	Vote	543	19	Malaysia	1198	21
Vote	650	14	Malaysia	361	13	PH	1051	19
BN	507	11	Results	357	12	Najib	528	9
Help	471	10	Election	266	9	New	498	9
Election	457	10	BN	258	9	BN	462	8
Infinity war	428	9	SPR	254	9	Mahathir	409	7
Need	410	9	Time	227	8	PM	398	7
Time	345	8	People	211	7	Time	366	7
Party	298	7	Hope	193	7	Tun	360	6
Sabah	269	6	Polling	187	7	Government	340	6

Topics related to BN, Vote, Election, Malaysia and Polling emerged mostly before and on the election day, whereas PH, Prime Minister (PM) and Tun Mahathir emerged the most after the election, mostly due to the opposition (PH) winning the GE14, with Tun being nominated as the new PM. Looking closer, an interesting keyword that emerged during pre-GE14 was Infinity War, and a further analysis of the sample tweets showed there was a trend among Malaysians making jokes about the contesting politicians based on the popular Marvel Avengers movie, Infinity War, which was playing in Malaysia during the time. A sample tweet includes "*My submission for the MBO #Avengers**InfinityWars** contest: "This movie is so amazing all the characters should be nominated for #GE14."* As for post-GE14, the keywords mainly focused on the winning party (i.e. PH), and discourses on what Malaysians referred to as "a new Malaysia", and the appointment of Tun Mahathir as the seventh Prime Minister.

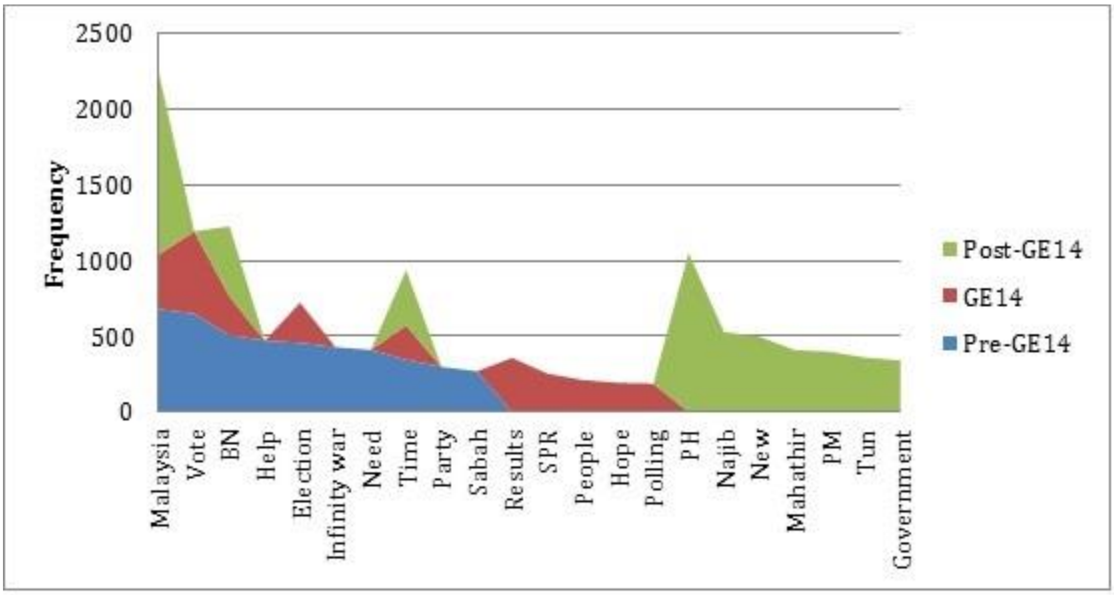


Fig 2: Temporal-based top keywords

Fig. 2 illustrates the keywords based on their timelines, with a clear trend notable for certain keywords. For example, though keywords such as Malaysia and vote appeared to be high throughout the period, there were also time-specific keywords such as SPR (Election Council), Hope and Polling, which appeared mostly during the election day, and keywords such as new (referring to new Malaysia), Mahathir and Government were noted after the election day.

Table 4: Top keywords for the Malay speakers

Pre-GE14			GE14			Post-GE14		
Keyword	N	%	Keyword	N	%	Keyword	N	%
Calon *(Candidate)*	609	20	Undi *(Vote)*	675	23	Menang *(Win)*	416	16
BN	488	16	BN	404	14	Rakyat *(Citizen)*	306	14
JomBN	366	12	Menang *(Win)*	396	14	Malaysia	305	14
Sabah	247	9	SPR	281	11	BN	304	14
Parlimen *(Parliament)*	205	8	Rasmi *(Official)*	272	10	Najib	264	9
Rakyat *(Citizen)*	194	8	Keputusan *(Result)*	256	7	Kerajaan *(Government)*	263	8
Undi *(Vote)*	179	7	Parlimen *(Parliament)*	191	5	Baru *(New)*	260	8
AnakKL	85	6	Tun	171	5	Tahniah *(Congratulations)*	147	6
Kerusi *(Chair)*	74	6	Malam *(Night)*	89	5	Hari *(Days)*	102	6
Bertanding *(Competing)*	62	6	Tunggu *(Wait)*	76	5	Orang *(People)*	91	5

Note: English translation in italic

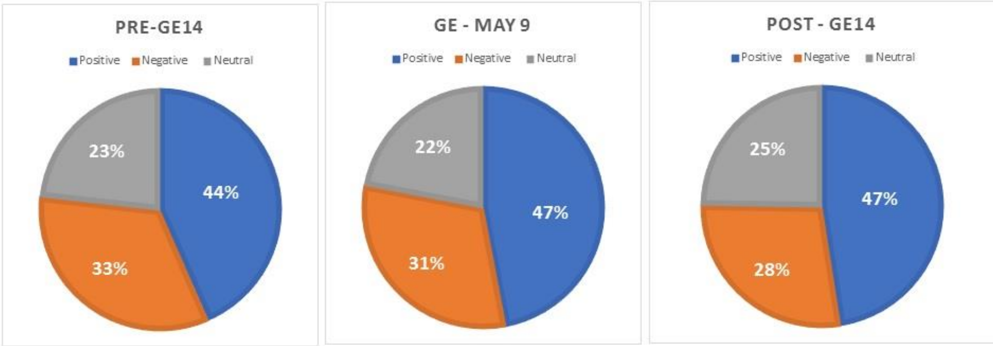
As for the Malay speakers (Table 4), the majority of the Twitter users seemed to be overwhelmingly BN supporters, as reflected by the high occurrences of JomBN (i.e. akin to let's support BN) during pre-GE14. Interestingly, the majority of the communications on GE14 (i.e. after 5pm) focused on 'malam' (night) and 'tunggu' (wait), indicating the long wait Malaysians endured for the GE14 results. In fact, many Twitter users joked about the long wait for the election results by comparing it to their experiences in waiting for their exam results.

The themes observed based on the content analysis showed that most of the communications among the Malay speakers were laden with jokes and sarcasm in relation to the electoral process, ink and the delay in result announcements (frequencies are not available as the experts were not required to label the tweets based on sarcasm/irony). One additional observation noted from analysing the tweets for both groups was the lack of overtly racist messages, regardless of the timelines, and thus indicating the majority of Malaysian Twitter users were not racist in their outlook and orientation. This, however, may not necessarily be sustainable in the long run should there be future economic and political turmoil in the country.

The keywords for pre-GE14 tended to be similar with the English speakers with the majority seemed to be PH supporters, as reflected by keywords such as 'menang' (win) and 'tahniah' (congrats). The emergence of 'baru' (new) is also in reference to the fact that Malaysia has now entered a new phase in its history, with the term new Malaysia being an indicative of the new sense of hope for the people of Malaysia. One topic that emerged for both the groups on GE14 day is SPR (i.e. Election Council). A glance at the sample tweets indicates Malaysians unhappiness in how SPR handled the electoral process.

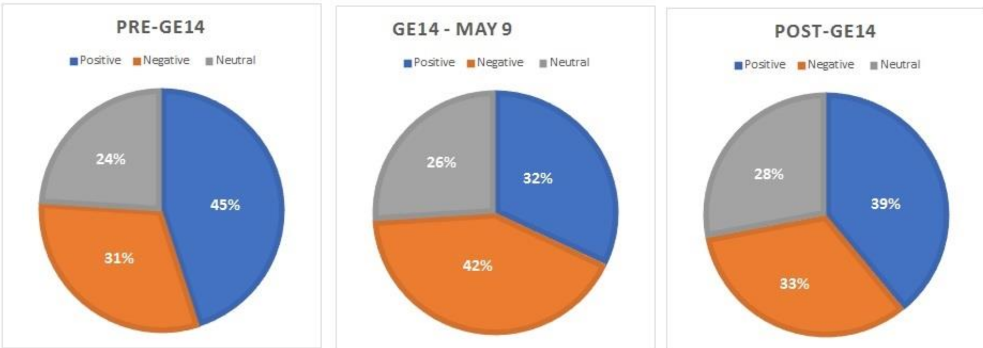
4.2. Public Sentiment Between Bi-Lingual and Malay Speakers

This section presents the results for the analyses performed to examine if the public sentiment differs between the English and Malay speakers, and to identify the top keywords based on their sentiments.



PRE-GE14						GE14 – MAY 9						POST-GE14					
Positive	%	Neutral	%	Negative	%	Positive	%	Neutral	%	Negative	%	Positive	%	Neutral	%	Negative	%
Vote	17	Malaysia	19	Help	15	Vote	17	Vote	17	Vote	22	Malaysia	24	Malaysia	21	PH	23
Malaysia	15	BN	14	Malaysia	15	Malaysia	14	Malaysia	13	SPR	12	PH	18	PH	19	Malaysia	17
BN	12	Vote	13	Vote	14	Results	14	Results	13	BN	10	New	13	Najib	11	Najib	13
Election	10	Election	9	BN	11	Best	11	BN	11	Results	10	Time	7	Mahathir	10	People	8
Thank	9	Candidates	9	Election	10	Election	9	Election	10	Malaysia	10	Better	7	BN	9	Mahathir	8
Help	9	GEanalyst	8	People	9	Time	8	SPR	9	Election	8	Najib	7	People	7	Want	7
Time	8	Help	7	Time	8	BN	8	Results	8	Box	8	BN	7	Tun	7	Tun	6
Good	7	State	7	Party	6	Better	7	Time	7	Polling	7	Government	6	Time	5	Years	6
Candidates	7	Time	7	Sabah	6	Hope	7	Waiting	7	Ballot	7	Tun	6	Results	5	Country	6
Like	6	Need	7	Ticket	6	Country	6	PH	6	Nervous	6	Mahathir	6	Election	5	Government	6

Fig 3: Sentiment analysis and top 10 keywords for the English speakers



PRE-GE14						GE14 – MAY 9						POST-GE14					
Positive	%	Neutral	%	Negative	%	Positive	%	Neutral	%	Negative	%	Positive	%	Neutral	%	Negative	%
Hebatkannegaraku	17	Calon	23	Calon	14	Undi	23	Undi	26	Undi	23	PH	23	PH	35	PH	22
Jombn	16	BN	16	BN	12	Menang	14	BN	11	SPR	15	Malaysia	14	Menang	13	BN	15
Dahulukanrakyat	15	Malaysiamemilih	11	Pilihbiruyanglebihmuda	12	BN	11	Parlimen	10	BN	11	Rakyat	14	BN	9	Najib	12
BN	11	Sabah	11	Malaysiamemilih	10	Langkawi	9	Keputusan	9	Keputusan	10	Menang	10	Malaysia	8	Menang	11
AnakKL	11	Parlimen	10	Orang	10	SPR	8	Calon	8	Tunggu	9	Tahniah	9	Najib	6	Boleh	10
Rakyat	8	Bertanding	6	Tolakbn	9	Tengok	8	SPR	8	Menang	8	BN	7	PM	6	Kerajaan	7
Pakatanharapan	6	Kerusi	6	Undi	9	Beratur	7	Dun	8	Blackout	7	Kerajaan	6	Rakyat	6	Rakyat	7
Parlimen	6	Parti	6	Rakyat	8	Keputusan	7	Rasmi	7	Saluran	6	Tun	6	Tun	6	Malaysia	6
Dun	5	Datuk	5	Malaysia	8	Malam	7	Datuk	6	Rasa	6	Baru	4	Kerja	6	Baru	5
Kerusi	5	Dun	5	1mdb	7	Malaysia	7	Menang	6	Rakyat	5	Negara	4	PAS	5	Orang	5

Fig 4: Sentiment analysis and top 10 keywords for the Malay speakers

Figures 3 and 4 illustrate the public sentiment for the English and Malay speakers, respectively, along with the top 10 keywords spread across the three timelines. In general, the sentiment seems to be vastly positive for both the languages; however, the majority of the Malay speakers' sentiments seemed to be negative on GE14 (i.e. 42%). Neutral sentiments were recorded the least, and the tweets were primarily official in nature (e.g. announcements). Keywords such as election, Malaysia, BN, PH and vote emerged throughout the communication for both the groups, and these were somewhat expected considering the nature of the event examined. For the English speakers, there seem to be a very strong feeling of nationalism regarding the fate of their country, judging from keywords such as vote (a possible huge reminder for Malaysians to vote for the sake of the country).

Looking at the negative sentiments for both the groups, it can be noted that SPR appeared to be frequently mentioned, in line with the results presented in Section 4.1. This is probably due to several factors such as the action of SPR in making it difficult for PH to contest during GE14, and their decision to hold GE14 on a working day (i.e. mid-week Wednesday) instead of during weekends, which was perceived to be unfair by many. Simply said, SPR was among the most disliked institution before and during the GE14, as indicated by some sample tweets below (in verbatim):

Bi-lingual speakers (GE14 day, negative sentiment) - *As Malaysians, irrespective of our political stance, I think we can all agree that **SPR** is a fucking disgrace to our country and rights.*

Malay speakers (GE14 day, negative sentiment) - *Please jadi demokrasi dan telus weh **SPR**, apa function kalau semua benda memang tak masuk akal.*

(Please be democratic and transparent SPR, what is the function if everything done is senseless)

Comparatively, the higher negative sentiments among the Malay speakers can be attributed to topics related to blackouts as well. This is because during the previous GE13, there were multiple occurrences of power outages at counting stations resulting in changes in results in very slim victories to the BN candidates. The Malay speakers have clearly not forgotten the incident, with many cracking jokes related to blackouts. A sample of translated tweet would be "*Guys get ready with candles, might suddenly blackout*".

Based on the top keywords identified, we looked at the sample tweets and it can be generally surmised that despite the slight differences between the sentiments of the English and Malay speakers, they nevertheless showed patriotisms toward the country by mostly communicating about providing financial support to those who need to travel to their home-states to vote (especially flights to Sabah and Sarawak) and arranging carpool sessions prior to GE14. As for GE14 day itself, the majority of the discourse revolved around encouraging others to vote, making fun of the electoral process and finally, communications geared towards the results after GE14. Other major communications post-GE14 focused on the victory of PH and Malaysians frustration in the delay of Tun Mahathir being appointed as the Prime Minister.

5.0 CONCLUSION, LIMITATION AND FUTURE DIRECTION

The study examined the Twitter communication during the most recent general election in Malaysia using sentiment and content analyses. Naïve Bayes with Word2Vec outperformed the rest of the classification models with an accuracy of 63.7%, 66.8% and 64.9% for pre-GE14, GE14 day and post-GE14, respectively. In general, findings indicate the overall sentiments to be positive throughout the period investigated, although negative sentiments seemed to be slightly higher among the Malay speakers on GE14 day compared to the English speakers.

Top keywords observed were in line with the timelines, with the majority of them related to voting (i.e. more of encouragement), carpooling and congratulatory remarks once the opposition wins. As for the negative sentiments, topics were mainly on SPR and their incompetency in handling the electoral process. In conclusion, our findings shed light on a novel aspect of the recent GE14, providing important insights and directions for research work on the political use of social media, especially in a developing country in South East Asia.

The study however, is not without its limitations. First, the tweets were gathered based on the trending hashtags. Though this eliminated biases compared to using single hashtags (such as the work of [9]), it may also be fruitful to explore the themes and discourses pertaining to specific hashtags. For example, a hashtag related to a single politician such as Tun or political party such as BN, may provide useful insights of Malaysian Twitter users perceptions and sentiments as well. Second, existing studies have showed Twitter features such as number of followers (and thus popularity), number of retweets and user activities to be positively correlated to positive sentiments [32]. This was not investigated in the present study; hence future studies could explore these relationships.

Third, social media data contain a lot of noise including spams and other bot activities, hence the importance in pre-processing them prior to any classifications or regressions. Although common precautions have been duly made in this study, no specific tasks were administered to check for spam activities, for example. Future studies could address this issue. It is also to note that the current study aimed to gauge Malaysians' sentiments during GE14, hence the models were not evaluated in terms of future predictions of sentiments. This is also an interesting avenue for future studies to explore.

Finally, it is to note that the study mainly focused on Twitter users (i.e. urbanized), and thus the findings related to English and Malay tweets should not be generalized across all Malaysians.

6.0 ACKNOWLEDGMENT

The authors would like to extend their heartfelt gratitude to the human experts who assisted with the data annotation. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] Yaqub, U., Chun, S.A., Atluri, V. & Vaidya, J. (2017a) "Analysis of political discourse on twitter in the context of the 2016 US presidential elections", *Government Information Quarterly 34 (2017)* 613–626.

[2] Statista. (2018). "Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2018 (in millions)". Retrieved from *https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.*

[3] Valenzuela, S., Correa, T., & Gil de Zúniga, H. (2017). "Ties, likes, and Tweets: Using strong and weak ties to Explain differences in protest participation across Facebook and twitter use". *Political Communication*, 35(1), 117-134.

[4] Enli, G. (2017). "Twitter as arena for the authentic outsider: Exploring the social media campaigns of Trump and Clinton in the 2016 US Presidential election". *European Journal of Communication*, 32(1), 50-61.

[5] Kreiss, D. (2016). "Seizing the moment: The presidential campaigns' use of Twitter during the 2012 electoral cycle". *New Media & Society*, 18(8), 1473-1490.

[6] Pramuk, J. (2016). "Trump spent about half of what Clinton did on his way to the presidency". CNBC. Retrieved from *http://www.cnbc.com/2016/11/09/trump-spent-about-half-of-what-clinton-did-on-his-way-to-the-presidency.html*.

[7] Mohammed, S. M., Zhu, X., Kiritchenko, S., & Martin, J. (2015). "Sentiment, emotion, purpose, and style in electoral tweets". *Information Processing & Management*, 51(4), 480–499.

[8] Chen, C. Y., & Chang, S. L. (2017). "User-oriented perspective of social media used by campaigns". *Telematics and Informatics*, 34(3), 811-820.

[9] Ahmed, S., Jaidka, K. & Cho, J. (2018) "Do birds of different feather flock together? Analyzing the political use of social media through a language-based approach in a multilingual context", *Computers in Human Behavior* 86 (2018) 299-310.

[10] Shehata, A., & Strömbäck, J. (2018). "Learning political news from social media: Network media logic and current affairs news learning in a high-choice media environment". *Communication Research*, https://doi.org/10.1177/0093650217749354.

[11] Xenos, M. A., Scheufele, D. A., Brossard, D., Choi, D. H., Cacciatore, M., Yeo, S., et al. (2018). "News media use and the informed public in the digital age." *International Journal of Communication*, 12, 706–724.

[12] Lee. S., & Xenos, M. (2019) "Social distraction? Social media use and political knowledge in two U.S. Presidential elections", *Computers in Human Behavior*, 90 (2019) 18–25.

[13] Singh, P., Dwivedi, Y. K., Kahlon, K. S., Pathania, A. & Sawhney, R. S. (in press) "Can twitter analytics predict election outcome? An insight from 2017 Punjab assembly elections", *Government Information Quarterly,* https://doi.org/10.1016/j.giq.2019.101444.

[14] Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P. G., & Cranor, L. F. (2011, July). "I regretted the minute I pressed share: A qualitative study of regrets on Face- book". *In Proceedings of the 7th Symposium on Usable Privacy and Security* (p. 10). ACM.

[15] Burnap, P., Gibson, R., Sloan, L., Southern, R. & Williams, M. (2016) "140 characters to victory?: Using Twitter to predict the UK 2015 General Election Electoral Studies" 41, 230-233.

[16] Yaqub, U., Chun, S., Atluri, V., & Vaidya, J. (2017b). "Sentiment based analysis of tweets during the US presidential elections". *Proceedings of the 18th annual international conference on digital government research* (pp. 1–10). ACM.

[17] Rill, S., Reinel, D., Scheidt, J. & Zicari, R. V. (2014) "PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis", *Knowledge-Based Systems* 69, 24–33.

[18] Shim, K. J. & Oh, S. K. (2017) "Who creates the bandwagon? The dynamics of fear of isolation, opinion congruency and anonymity-preference on social media in the 2017 South Korean presidential election Computers in Human Behavior" 86, 181-189.

[19] Wang, D., & Yue, C. (2016). "Research on the usage of social media in Taiwan region's political campaigns and elections". *Intercultural Communication Studies,* 25(3),232-247.

[20] Nulty, P., Theocharis, Y., Popa, S. A., Parnet, O., & Benoit, K. (2016) "Social media and political communication in the 2014 elections to the European Parliament Electoral Studies" 44, 429-444.

[21] Marozzo F, Bessi A. "Analyzing polarization of social media users and news sites during political campaigns". *Social Network Analysis and Mining,* 8, 1.

[22] Huat, W. C. (2013). "Malaysia's education divide. Penang monthly". Retrieved from *https://penangmonthly.com/article.aspx?pageid=7026&name=malaysians_education_divide*.

[23] Lee, S. K., Lee, K. S., Wong, F. F., & Azizah, Y. A. (2010). "The English language and its impact on identities of multilingual Malaysian undergraduates". *GEMA Online Journal of Language Studies*, 10(1), 87-101.

[24] "Malaysian Communications and Multimedia Commission (2018) Internet Users Survey 2018", Statistical Brief Number 23, pp. 1 – 52. Retrived from *https://www.mcmc.gov.my/skmmgovmy/media/General/pdf/Internet-Users-Survey-2018.pdf.*

[25] Yusop, F. D. & Sumari, M. (2013) "The use of social media technologies among Malaysian youth Procedia" - *Social and Behavioral Sciences* 103, 1204 – 1209.

[26] Gomez, J. (2014). "Social media impact on Malaysia's 13th general election". *Asia Pacific Media Educator*, 24(1), 95-105.

[27] Sun, S., Luo, C., & Chen, J. (2017). "A review of natural language processing techniques for opinion mining systems". *Information Fusion,* 36, 10-25.

[28] Song, M., Park, H., & Shin, K. S. (2019). Attention-based long short-term memory network using sentiment lexicon embedding for aspect-level sentiment analysis in Korean. Information Processing & Management, 56(3), 637-653.

[29] Kaur, W. & Balakrishnan, V. (2018) "Improving sentiment scoring mechanism: a case study on airline services", *Industrial Management & Data Systems*, 118, 1578-1596.

[30] Yordanova, S., & Kabakchieva, D. (2017). "Sentiment Classification of Hotel Reviews in Social Media with Decision Tree Learning". *International Journal of Computer Applications*, 158, 1-7.

[31] Huq, M. R., Ali, A., & Rahman, A. (2017). "Sentiment analysis on Twitter data using KNN and SVM". *International Journal of Advanced Computing Science and Application*, 8(6), 19-25.

[32] Chu, K. H., Majmundar, A., Allem, J. P., Soto, D. W., Cruz, T. B., & Unger, J. B. (2019). Tobacco use behaviors, attitudes, and demographic characteristics of tobacco opinion leaders and their followers: Twitter analysis. Journal of medical Internet research, 21(6), e12676.

[33] Shaojun, H., Jin, C., Ruixu, G., & Guijun, W. (2012, December). The Capability Analysis on the Characteristic Selection Algorithm of Text Categorization Based on F1 Measure Value. In *2012 Second International Conference on Instrumentation, Measurement, Computer, Communication and Control* (pp. 742-746). IEEE.

[34] H.-T. Nguyen, L.-M. Nguyen, (2020) "ILWAANet: An Interactive Lexicon-Aware Word-Aspect Attention Network for aspect-level sentiment classification on social networking", *Expert Syst. Appl*., 146, 113065.

[35] A. Giachanou, F. Crestani, (2016) "Like it or not: A survey of twitter sentiment analysis methods", *ACM Computing Surveys (CSUR)*, 49, 28