Int. J. Advance Soft Compu. Appl, Vol. 11, No. 1, March 2019
ISSN 2074-8523

Building a Malay-English Code-Switching Subjectivity Corpus for Sentiment Analysis

Emaliana Kasmuri, Halizah Basiron

Fakulti Teknologi Maklumat dan Komunkasi,
Universiti Teknikal Malaysia Melaka, 75450 Durian Tunggal Melaka
email: emaliana@utem.edu.my

Fakulti Teknologi Maklumat dan Komunkasi,
Universiti Teknikal Malaysia Melaka, 75450 Durian Tunggal Melaka
email: halizah@utem.edu.my

Abstract

Combining of local and foreign language in single utterance has become a norm in multi-ethnic region. This phenomenon is known as code-switching. Code-switching has become a new challenge in sentiment analysis when the Internet users express their opinion in blogs, reviews and social network sites. The resources to process code-switching text in sentiment analysis is scarce especially annotated corpus. This paper develops a guideline to build a code-switching subjectivity corpus for a mix of Malay and English language known as MY-EN-CS. The guideline is suitable for any code-switching textual document. This paper built a new MY-EN-CS to demonstrate the guideline. The corpus consists of opinionated and factual sentences that are constructed from combination of words from these the languages. The sentences were retrieved from blogs and MY-EN-CS sentences are identified and annotated either as opinionated or factual. The annotated task yields 0.83 Kappa value rate that indicates the reliability of this corpus.

Keywords: *Annotation guideline, code-switching corpus, sentiment analysis, subjectivity corpus*

1 Introduction

Combining local and foreign languages in verbal and textual communication has become a norm for multi-ethnic community. Foreign words are used as substitute to some of the words in the local language. Foreign words are also used when there are no equivalent words that have the same meaning in the local language. Thus, it is convenient for the speaker to used foreign words to convey his message. Combining languages in verbal and textual communication happens within a sentence (inter-sentential) or in between two sentences (intra-sentential). Inter-sentential combination occurs when foreign words are combined with local words within a sentence. As an example, in the sentence "*Kita kena tailor what kind of education, technology, expertise yang diperlukan pada masa 30 tahun akan datang*" (English translation: We have to tailor the education, technology and expertise to suit the need for the next 30 years). In the example, the foreign words (the underlined words) are combined with the local words in the sentence. Intra-sentential is a change of language from one sentence to another. As an example, in the sentence "*Anak muda Malaysia mahu negara yang bahagia. We want a country with people that are fulfilled with their lives*" (English translation: The Malaysia youth wants a peaceful country. We want a country with people that are fulfilled with their lives). In the example, the language in the first sentence is Malay and the language in the second sentence is English. In linguistic study, the first sentence is known as code-mixing and the second sentence is known as code-switching [1]. This paper is using the term code-switching referring to all types of language mixing and switching.

The trends of social media communication have amplified the phenomenon of code-switching in multi-ethnic community. It is common for many multi-lingual social network users to exchange factual and opinionated information using code-switching. Recent discovery reveals that code-switching was used by prominent and influential people in their social network accounts. This shows that code-switching has become more acceptable to many social network users regardless of their status. This scenario motivates this paper to create a new kind of corpus as one of the important resources for sentiment analysis.

The aim of this paper is to establish a new subjectivity annotation guideline for code-switching sentences. The guideline is independent from any mixture of languages. The expected outcome from this guideline is a collection of annotated subjective code-switching sentences usable for subjectivity classification and sentiment analysis. This paper has selected Malay and English code-switching language to demonstrate the guideline.

This paper is organized in the following sections. Section 2 describes previous studies that have developed corpus for sentiment analysis. Section 3 describes the guideline and procedure to build code-switching subjectivity corpus.

Section 4 reports and discusses the result of this paper and finally Section 5 concludes this paper.

2 Related Work

2.1. Type of sentiment information

Information in a text consists of factual information and opinionated information. The factual information describes the property or attribute of the discussed subject matter such things, people, events and organizations. As an example, in the sentence "*The opening ceremony of the exhibition will be held in Hall A*", describes the place for the opening ceremony. The opinionated information describes the writer's personal experience, evaluation, opinion, judgement, view or feeling towards the subject matter. As an example, "*iphoneX is quite expensive for middle income earner like me*", describes the author's evaluation for the price of iPhone X. The expression that contains personal experience, evaluation, opinion, judgement, view or feeling is known as private states [2]. The study of sentiment analysis interprets factual information as objective information and opinionated information as subjective information [2]. This paper uses the term objective and subjective to refer to such information.

Subjective information can be categorized as either positive or negative information. This paper labelled positive or negative information as polarized information. Positive information indicates preference of the opinion expressed towards the subject matter. For example, "*Money well spent on the new iPhone*" indicates the author's positive satisfaction on iPhone. Negative information indicates criticality of the expressed opinion. For example, "*That is the two hours that I will never get back after watching The Peace Maker*". This sentence shows the author's negative sentiment on The Peace Maker movie. Polarized information can be further categorized based on the strength of the information such as strongly positive-positive-weakly positive or specific emotion such as happy, sad, jealous or disgusted. This paper discovers there is a hierarchical relation between subjective-objective, positive-negative, strength and classes of emotions. The relation is shown in Fig. 1. The shaded box in Fig. 1 shows the type of information deals in this paper.

2.2. Annotated sentiment corpus

Sentiment analysis consists a series of processes that determined the types of sentiment information as described in Fig. 1. Compilation of several literature concludes that machine learning technique is the most preferable technique used by many studies of sentiment analysis [3][4][5]. Machine learning technique learns effectively from annotated sentiment corpus. Therefore, annotated sentiment corpus is the most essential element in the study of sentiment analysis.

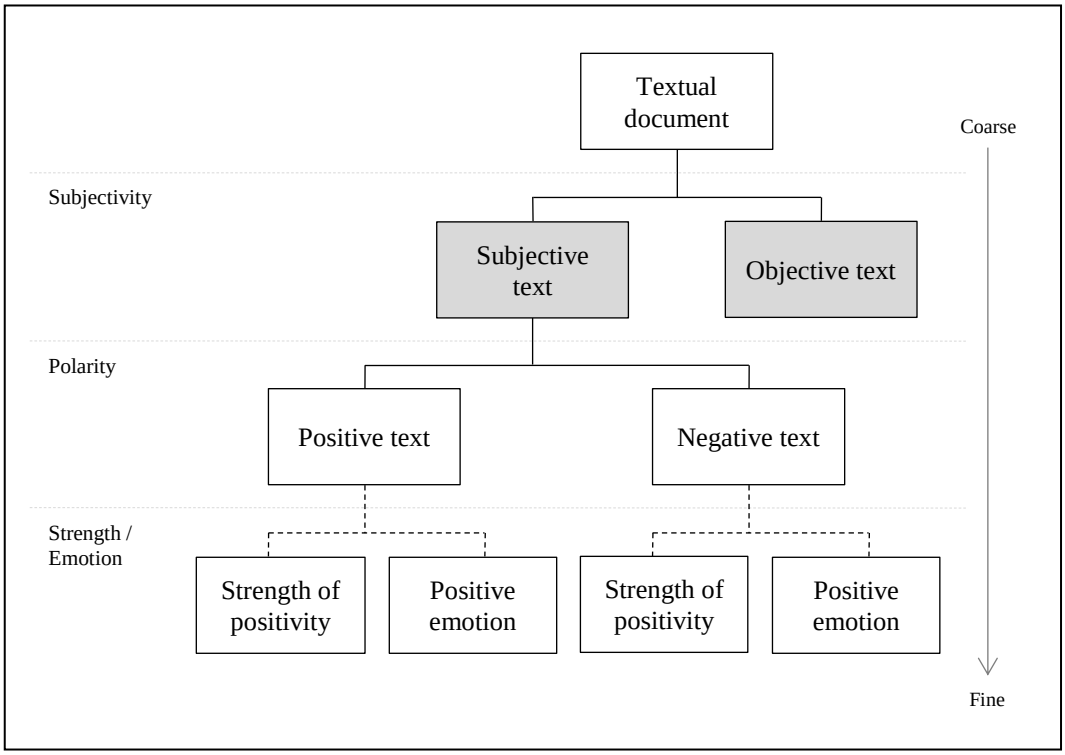

Fig. 1: Hierarchical relationship of information in sentiment analysis

The Wall Street Journal (WSJ) Corpus [6], the Multi-Perspective Question Answering (MPQA) corpus [7] and the Cornell Movie Review (CMR) corpus [8] are among the early corpora built for the study of sentiment analysis. The WSJ corpus is a huge collection of news articles. The studies that used this corpus have extracted some of the articles and annotated them with subjective-objective annotation [10] and positive-negative annotation [11]. The data MPQA and CMR corpora are annotated with subjective-objective annotation. These corpora are the most popular corpora that are being used by many studies of sentiment analysis to date [5].

Starting from 2013, text from social network sites such as Twitter have become the most preferable analyzed text in the study of sentiment analysis [5]. With the advance of mobile technology, the information is rapidly created by its users. The content created by the users of social network sites consist of objective and subjective information. This situation is an advantage to the studies of natural language processing especially sentiment analysis, where huge amount of data has become accessible to the study. Since then, the need to study sentiment expressed on social network sites has become critical. Therefore, many studies have created their own corpus using text from social network sites. In addition to that, the characteristics of texts from social network sites are different from the established corpus such as from the WSJ, the CMR and the MPQA. Therefore, the need for

some of the studies of sentiment analysis to create a new corpus that is sampled from social network sites are justified [12][13].

A sentiment corpus is created either using manual or automated approach. Manual approach has been the prime chosen approach by many studies because the expertise of the annotators. Therefore, manual annotated sentiment corpus has been considered as a gold standard corpus [7]. The annotated corpus has been used in many experiments to validate the prediction and classification model built for sentiment analysis. The annotated corpus represents snapshots of real data from the represented domain. The corpus consists of a collection of textual documents annotated with labels such as objective/subjective, positive/negative and positive/negative/neutral [7][14].

Generating sentiment corpora that are manually built and annotated is time consuming and labor intensive [15]. Despite these factors, having manually annotated corpus is still necessary for machine learning studies because machine learning techniques learned effectively from annotated corpora. However, annotating the corpus is a difficult task [15][16]. This is evident with the low reliability score achieved in the inter-annotator agreement [15]. The score is low because sentiment texts are subjective texts. Therefore, the evaluation of the text may differ from one annotator to another annotator. The annotation was highly influenced by the annotator's background. Furthermore, some of the texts used ambiguous subjective words. These factors affect the annotator's interpretation of the analyzed text. The construction of compounded sentences with many conjunctions used to connect the opinion from various aspects of the discussed subject matter adds more complication to the annotation task. This paper assumes that these challenges perhaps were not anticipated prior to the annotation tasks. It was also assumed that the absence of annotation guidelines contributed to the difficulty.

The studies of sentiment analysis in other languages had created their own corpora such as Arabic [17], French [18], Spanish [19], Italian [20], Greek [14], Portuguese [21] and Hindi [22]. Majority of the study used manual approach to annotate the corpora. Machine translation service was used to overcome the problem of limited sentiment corpus in Portuguese [21].

2.3. Code-switching sentiment corpus

The number of studies in sentiment analysis that involve code-switching text is increasing even though it is not as rapid as the number of mono-language studies. Nevertheless, the growing number of studies that used code-switching text indicates its importance in sentiment analysis system. Ignoring code-switching text in sentiment analysis will lead to inaccurate analysis result. Code-switching is a language that evolved from combination of local and foreign languages. The characteristics and the construction of code-switching text are different from mono-lingual text. The unavailability of sentiment corpus for languages other than

English can be overcome by translating English corpus into the target language using automatic machine translation service. However, this is not a feasible approach for the code-switching text because the service of machine translation cannot choose at random the words or parts of the sentence to be changed or replaced with the words from the foreign language. Even if the service of automated machine translation is able to do so, the result will not be as authentic as code-switching text composed by the multi-lingual speakers.

There are a few studies that had created code-switching corpus for sentiment analysis as shown in Table 1. In all of the studies in Table 1, English is the foreign language combined with the local languages such as Chinese, Malay, Spanish and Hindi. Tweets are the most selected type of text used to build the corpus because the accessibility of the data on the open platform. The tweets are rapidly generated as compared to other types of data such as blog postings, feedbacks and comments. The domain represented by the corpus are local domain, that is the subject matter that received the most attention from the local people of the specific location. The majority of the studies in Table 1 annotated the text in their corpus either as positive or negative. These corpora are suitable for polarity classification in sentiment analysis. Only one study described in Table 1 used multiple label of emotion. This corpus is used in emotion classification of sentiment analysis. The reliability of the annotated corpus is measured using Kappa score. The reliability score of the studies described in Table 1 is between 0.5 to 0.9 Kappa.

Table 1: Code-switching corpus for sentiment analysis

Study	[23]	[24]	[25]	[26]
Language	Singlish	English-Spanish	English-Hindi	English-Chinese
Type of Text	Tweets	Tweets	Comments	Post
Domain	Local	General	Government	Unspecified
Method of text selection	Automated language identification	Existing corpus from	Manual	Manual
Size	215 positives, 459 negatives	3,062 positive and negative	1,011 positive and negative	4,195 multiple emotions[1]

[1] Multiple emotions: Happiness, Sadness, Fear, Anger, Surprise

Label	Positive-Negative	Positive-Negative	Positive-Negative	Multiple emotions[2]
Number of annotators	3	3	2	2
Reliability score	74.00% positive 79.00% negative[3]	50.00% – 69.30% positive 62.90% – 66.40% negative[4]	95.32% Hindi, 96.82% English[5]	69.20%[6]

In general, the annotation schemes used to annotate the code-switching sentiment corpus described in Table 1 are similar. Sentiments expressed in local and foreign languages were distinctively annotated [23][24][25]. The distinctive annotation was necessary because the polarity of the sentiment could change between local and foreign languages. This annotation showed richness and complexity of annotated code-switching sentiment corpus. The performance of polarity classification for code-switching text could be improved with eliminating non-polarity or objective text [26]. Majority of the existing code-switching sentiment corpus exclude non-opinionated or objective annotation. Therefore, this paper designs a new guideline to annotate code-switching subjectivity corpus to fill the gap. This paper used Malay-English code-switching (MY-EN-CS) as subject to demonstrate the proposed guideline.

3 The Annotation Guidelines

The procedure to build MY-EN-CS corpus for sentiment analysis is adapted from [27]. The process consists of seven steps as shown in Fig. 2. The process starts with selecting the type of the text that will represent the corpus. The task for this step consists of the method to gather, extract, catalogue and index the raw text. This paper found that this step is similar to the pre-processing step in text analysis process. After that, a list of annotation label is defined based on how the gathered text should be categorized. The annotation label is determined prior the process. A clear and comprehensive description for the annotation labels shall be included. In the next step, the annotators are appointed for the annotation tasks. The annotation

[2] Multiple emotions: Happiness, Sadness, Fear, Anger, Surprise
[3] Fleiss Kappa
[4] Krippendorf Alpha
[5] Cohen Kappa
[6] Cohen Kappa

training should be provided to the appointed annotators in order to impart the knowledge and skills required for the task.

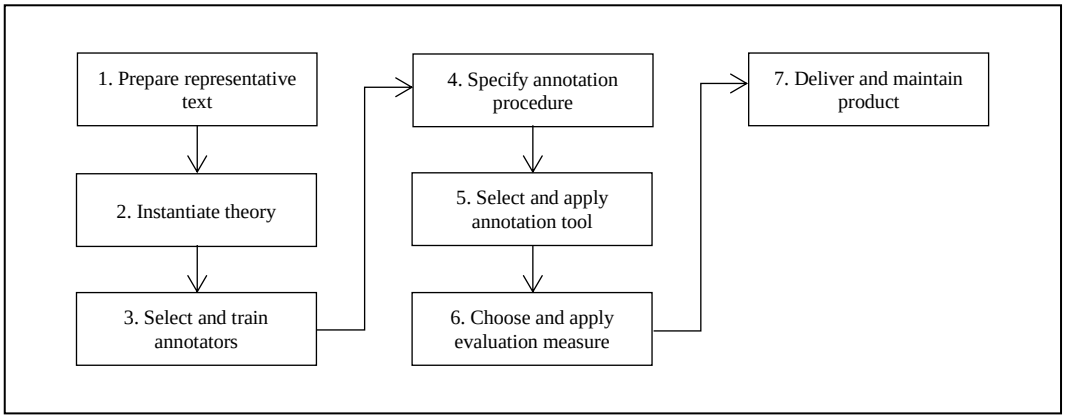

Fig. 2: General annotation process to build a corpus

The annotation guideline should be established before the annotation task begin. The guideline consists of description of task to manage and monitor the progress of the annotation process. The monitoring task shall include steps to solve annotation issues which will hampering the progress of the annotation tasks. In the next step, an annotation tool will be selected for the annotation task such as Amazon Mechanical Turk, Crowdflower or WebAnno [28]. The selected tool shall support the procedure established for the annotation task for ease of task management and monitoring. After the annotation task is completed, the reliability of the annotated corpus is measured. The measurement determines the quality of the annotated corpus, the correctness of the underlying theory used in the annotation and the effectiveness of the established procedure. The process continues with the selection and compilation of annotated data for the corpus. Then, finally the annotated corpus is disseminated for ease of access.

3.1. Selecting and preparing representative text

Finding raw corpora for MY-EN-CS is a challenging task. The result produces by crawlers to find code-switching data are often inaccurate because only one language can be specified to the crawlers. Therefore, this paper has to resort to manual code-switching identification instead of the automated approach. The process of preparing the representative text for this paper is described in Fig. 3.

This paper has identified 23 personal blogs to build the annotated MY-EN-CS corpus. These blogs contain entries with great description about the blogger personal experience on various events and products using mixture of Malay and English languages. Therefore, it has a high potential of having huge number of opinionated sentences that is qualified for MY-EN-CS corpus. Though blogging activities are not as active as in the early of 2000s, blogs are still presently relevant because of the detailed description of the bloggers' personal experience. The blogs

did not limit the number of characters per blog post, therefore the blogger fully utilized this capacity to elaborate and provide as many information as possible to the reader. Furthermore, the blogs are public blogs. Therefore, the accessibility to the blogs are not limited to the member of the blogs. Hence, this paper deemed the blogs as viable representation for the MY-EN-CS corpus. In addition, there are two similar characteristics of the sentences constructed in the selected blogs in this paper with the ones used in social network sites. The similarities are usage of creative spellings, smileys, emoticons, emojis and the excessive use of punctuations and abbreviations. With the recent development in Twitter the number of characters for posting has increased from 140 to 280 characters. Thus, blogs are deemed as comparable to Twitter.

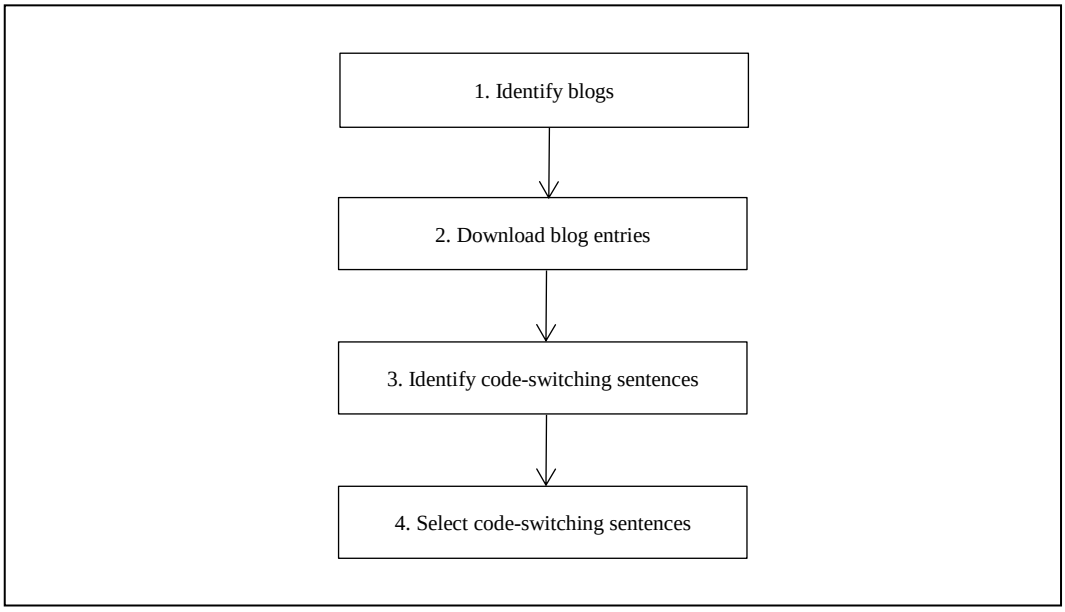

Fig. 3: The process of preparing representation text for MY-EN-CS

The first process in preparing the representative text is identifying the blogs. The process continues with blogs' entries collection. This task is executed by downloading the entries of the blogs using Python program. The next process is identifying mono-lingual and code-switching sentences. A rule-based method combined with dictionary look-up technique is used to determine the sentence either as mono-lingual or code-switching. 1-gram technique is used to extract each word in the processed sentences. A Java program interfacing with two lexicons, WordNet [29] and WordNet Bahasa [30], are used to implement this task. The program collects and counts words that belongs to English words, Malay words, shared words and out-of-vocabulary (OOV) words. Shared words are words that have entries in both lexicons while OOV words are words that do not have entries in both lexicons.

In the final process mono-lingual and code-switching sentences are identified and selected. In this paper, a sentence is assumed to be a MY-EN-CS

sentence with at least the presence of one functional Malay and one functional English word. Functional words are words that does not belongs to stop word list. In addition to that, a sentence that have between three (3) to 20 words of length is selected for the annotation task. Sentences that has less than three words are not informative to be determined as either subjective or objective sentences. Sentence that has more than 20 words adds more complication to the annotation task. Often these kinds of sentences have an overwhelming information that leads to difficulties in determining the sentence as either subjective or objective. Furthermore, these kinds of sentences are poorly constructed by the bloggers with improper or non-existent punctuation and segmentation of subject matter narrated in their blog posts.

3.2. Instantiating the theory

This paper adopted the annotation scheme described in [24]. However, the work was designed for mono-lingual sentences. This paper extends the previous work to accommodate code-switching sentences. Thus, the need to have the MY-EN-CS corpus. The MY-EN-CS contains a collection of subjective and objective sentences constructed using Malay and English words.

In the selection process the entries of the blogs used three types of languages which are Malay (MY), English (EN) and Malay-English code-switching (MY-EN-CS). Each language consists of subjective (OPI) and factual (FAC) sentences. However, during the selection process has been executed, English and Malay mono-lingual sentences were found in the selected sentences due to shared words. This paper did not view this as a methodological problem rather as controlled variable for the task. In consideration to this controlled variable, the annotation scheme for this paper is defined in the following: -

1. An English sentence that contains opinion expression is labelled as EN-OPI. As an example, "*Maybe because Im pretyy hihihihihi bimbo laugh*". The word "*pretty*" shows an evaluative expression towards the blogger and "*hihihihihi bimbo laugh*" indicate the emotion of the blogger claiming on herself. Therefore, this sentence is categorized as EN-OPI.
2. An English sentence that describes facts of an entity is labelled as EN-FAC. As an example, "*The owner of RedGlow Najwa Arlina*". This sentence described the name of an entity. Therefore, this sentence is categorized as EN-FAC.
3. A Malay sentence that contains opinion expression is labelled as MY-OPI. As an example, "*aku jarang dengar radio*" (English translation: I rarely listened to the radio). The word "*jarang*" is an estimation of time established by the blogger on himself and it varies from one person to another. Thus, this sentence is categorized as MY-OPI.
4. A Malay sentence that describes facts of an entity is labelled as MY-FAC. As an example, "*Depa ckp kt blok pompuan ada pocong kt tingkat 5*" (English translation: They said there is a ghost at the fifth floor at the ladies' block). The

sentence described the existence of ghost at the ladies' block. Therefore, this sentence is categorized as MY-FAC.

5. A MY-EN-CS sentence that contains opinion expression and using English word or phrases to describe the expression is labelled as CS-EN-OPI. As an example, "*And I don't understand why it has to be me yang kena benda benda macam ni*" (English translation: And I don't understand why these things are happening to me). The phrase "*don't understand why*" shows the frustration of the blogger. Therefore, this sentence is categorized as CS-EN-OPI.
6. A MY-EN-CS sentence that contains opinion expression using Malay word or phrases to describe the expression is labelled as CS-MY-OPI. As an example, "*Biarlah apa orang nak kata apa I have to put myself first even it means mengenepikan orang lain*" (English translation: I don't care what people are going to say, I have to put myself first before others). This sentence shows the emotion of the blogger with the usage of "*Biarlah orang nak kata apa*". Thus, this sentence is categorized as CS-MY-OPI.
7. A MY-EN-CS sentence that describes fact of an entity using either Malay or English words or phrases is labelled as CS-FAC. As an example," *Kakak mai sini on Monday evening*" (English translation: My sister came on Monday evening). This sentence states the arrival time of the blogger's sister. Therefore, this sentence is categorized as CS-FAC.

3.3. Selecting and training the annotators

In order to limit the effect of bias opinion, two annotators are required for the annotation process. For this paper, two annotators were selected as candidate annotators. These candidates were undergraduate students who were proficient in written and spoken Malay and English. The candidates were briefed on the annotation workflow, the annotation scheme and the annotation task before the annotation process commenced.

The candidates were given a set of sample subjective and objective text to be annotated. The sample annotation served as an evaluation of the candidates' comprehension of the annotation process. The sample annotations were then verified by the first author. Both annotators achieved excellent result. Therefore, both were appointed for the task.

3.4. Specifying the annotation procedure

Annotating sentiment for monolingual corpus is difficult and can lead to poor inter-annotator agreement score [31]. For this paper, a monitoring procedure was established to ascertain the quality of the annotated corpus and to clarify ambiguous sentences. The procedure was setup to resolve problematic sentences and completeness of labelling.

A bi-weekly meeting was held to discuss the issues and the problems related to the annotation task. The annotators would report the problematic sentences and the team would discuss about them. Then, the annotators will mark a confident score with either 0 as non-confident or 1 as confident to the deliberated sentences. This process was repeated for all problematic sentences until the confident score of 1 was achieved. The non-reported sentences were assumed as confidently annotated sentences in this paper.

The curator of the annotation task (the first author) monitors the project closely using annotation tool, WebAnno. The files which were marked as finished by both annotators were nonetheless verified by the curator to ensure no sentence was missed by the annotators.

3.5. Choosing annotation tool

Handling voluminous texts for annotation is a challenging task. Managing these texts manually is an expensive effort and inefficient. This paper used WebAnno to manage the annotation task. WebAnno is a web based annotation tool that allow the user to create custom annotation scheme (apart from the preloaded annotation scheme), to monitor the progress of the annotation project, to curate the annotated data and to produce annotation result [28]. WebAnno was selected because the features provided by the tool met the requirements of this paper in terms of managing and monitoring the annotation task.

3.6. Choosing and applying evaluation measure

The reliability of the annotated sentences was measured using Kappa value and inter-annotator agreement. Majority of the studies that built their own annotated sentiment corpus used Cohen Kappa to measure the reliability of the corpus [32][33][16]. On the other hand, some studies used Fleiss Kappa [23][14] and Krippendorf Alpha [34]. WebAnno provides all three measurements. Kappa value was used to interpret the agreement between the annotators. The value of Kappa was shown in Table 2.

Table 2: Kappa value interpretation

Kappa Value	Less than zero	0.01 – 0.02	0.21 – 0.40	0.41 – 0.60	0.61 – 0.80	0.81 – 1.00
Agreement	Poor	Slight	Fair	Moderate	Substantial	Almost perfect

3.7. Delivery and maintaining the product

The final product of the annotation process contains the sentences annotated as CS-EN-OPI, CS-MY-OPI and CS-FAC. These annotated sentences

were selected for the corpus. This paper did not put the corpus for public dissemination. However, the corpus can be obtained from the first author. The maintenance plan for this corpus is not defined.

4 Results

4.1. Result from pre-selection

This paper has downloaded blog entries posted between 1 January 2011 and 31 December 2017 from 23 personal blogs. The entries were separated into individual sentences for code-switching sentence identification and annotation task. This paper processed 5,849 sentences from 1,091 entries of various length. These sentences consisted of Malay, English and MY-EN-CS sentences. There were duplicated sentences found from the processed sentences. These duplicated sentences described mailing addresses, advertisements and promotions, empty sentences, republished of previous entry (instead of specifying hyperlinks) and voting requests. The duplicated sentences were removed. Basic statistics concerning the downloaded blog postings after the removal of duplicated sentences is shown in Table 3.

Table 3: Basic statistics concerning the downloaded blog posting

Statistical Information	Number of blogs	Number of entries	Number of sentences	Number of words
Statistical value	23	1,091	5,090	59,334

The sentences processed in this paper varied in length. The length referred to the number of words used in a sentence. This paper found the minimum length of a sentence was one (1) word and the maximum length was 340 words. Majority of the processed sentences were between one (1) word to 10 words. This paper found that sentences with less than three (3) words were not sufficiently informative to be determined as either subjective or objective sentence. Therefore, these sentences were discarded at this phase. This paper also found sentences with length of more than 20 words were too overwhelming in the sense that the discussed matter was not properly segmented and poorly punctuated. This paper deemed sentences with length of more than 20 words as vague sentences. Therefore, these sentences are also discarded at this phase. Fig. 4 shows distribution of sentences based on length.

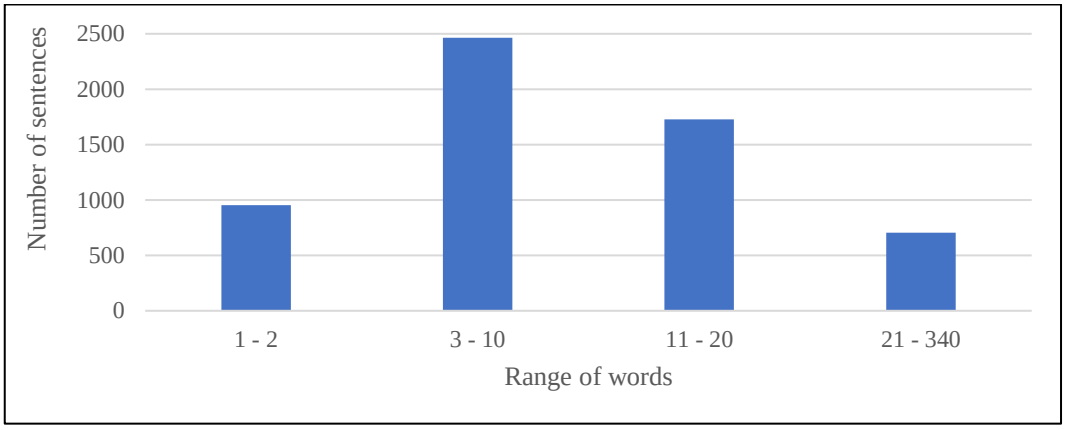

Fig. 4: Distribution of sentences based on range of words

4.2. Results from post-selection and pre-annotation

A total of 4,191 sentences from 5,849 sentences were selected for the annotation task. The distribution of the selected sentences is shown in Fig. 5. The histogram shows the length of the sentences selected for the annotation task is well distributed.

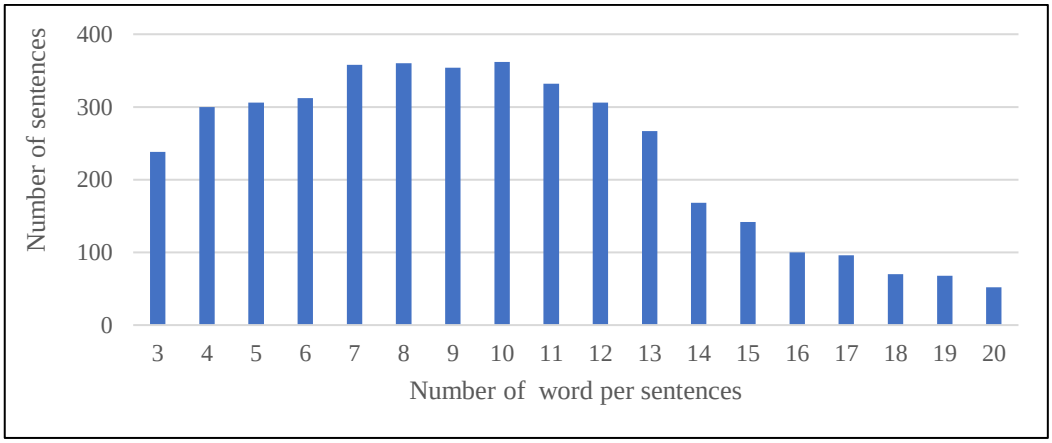

Fig. 5: Distribution of length of selected sentences for the annotation

This paper processed total of 21,851 words from 4,191 sentences. These 21,851 were all the words that were counted regardless of repetition. For example, the word *pen* was found three times in the selected sentences and was counted three times (was regarded as three words) and included in the total of 21,851 words. These 21,851 words were regarded as non-distinct words. From these 21,851 words, 10,376 were distinct words. Distinct words were words that were used only once in the selected sentences. Table 4 shows details of words processed from the selected sentences for the annotation task.

Table 4: Basic statistic of word collected from the selected sentences

Language	English	Malay	OOV	Total[7]	Shared words
Non-distinct	7,793	10,645	3,412	21,851	4,104
Distinct	3,889	3,420	3,067	10,376	768

4.3. Post annotation

This paper annotated 4,191 sentences using the annotation scheme described in section 3.2. In term of language distribution, 60.00% of the annotated sentences was labelled as MY-EN-CS, 35.00% as Malay and 5.00% as. In terms of subjectivity distribution, 52.00% of the sentences were subjective and 48% of the sentences were objective. In term of language distribution, the results shown that majority of the annotated sentences were MY-EN-CS. In terms of subjectivity distribution, the result shown that the selected sentences were subjective.

The result of the annotation process is shown in Fig. 6. The labels designated in the histogram were the labels agreed by both annotators. The label FALSE indicate the sentences that were labelled differently by the annotators. Most of MY-EN-CS sentences were opinionated sentences (referring to CS-EN-OPI and CS-MY-OPI). Majority of MY-EN-CS sentences used opinionated Malay words. The difference between CS-EN-OPI and CS-MY-OPI is substantial due to the background of the bloggers.

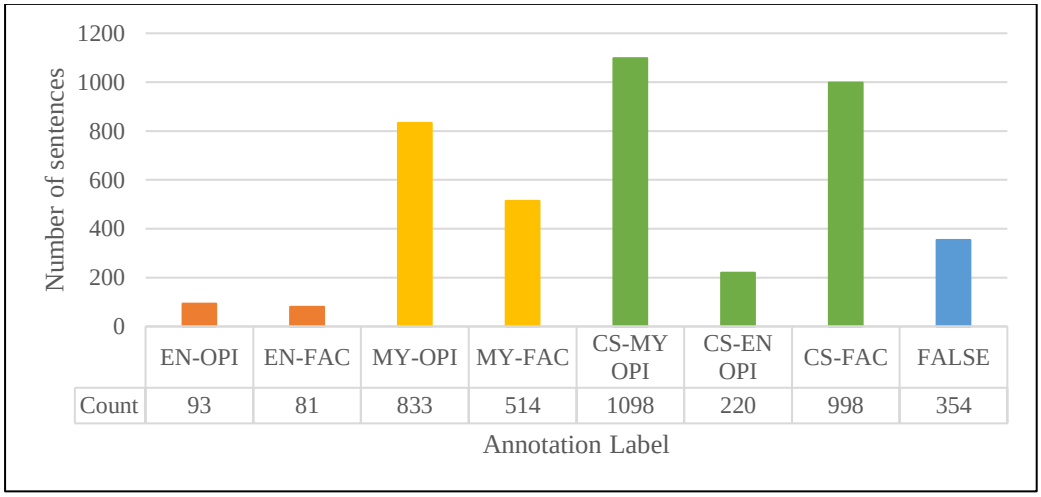

Fig. 6: Histogram of annotated sentences

[7] Total is exclusive from shared words

For this paper the result of the annotation achieved Kappa value of 0.83. According Table 2 this value signifies the annotated sentences are highly reliable. Based on this result, a total of 2,316 code-switching sentences were selected into the corpus. Other sentences are discarded from this corpus.

The result in Fig. 6 shows the difference in number of sentences between MY-OPI and CS-MY-OPI is not significant. This is due to the considerations of OOV words as Malay words. These OOV words were creatively spelled in the posting. Table 5 shows example of OOV words and respective correct form in Malay.

Table 5: List of some OOV word with creative spelling

OOV Words	**Correct Form**	**OOV Words**	**Correct Form**
sepatutnye	sepatutnya	paham	faham
tamau	tidak mahu	ilangkan	hilangkan
arituh	hari itu	bleh	boleh
xde	tidak ada	muke	muka

Another example of OOV words were spelled creatively according to how is sounds to the Malay speaker such as *kompem* (confirmed), *rumate* (roommate), *hensem* (handsome) and *saikosis* (psychotic). This paper categorized these words as Malayanization words. These words mark the importance of English words in MY-EN-CS sentences. Analysis of OOV words and shared words shows the influence of these words in the sentence selection. Ignoring both type of words at the selection phase will cause the system to miss out on important information. Therefore, these words need to be process systematically.

5 Conclusion

Building a code-switching sentiment corpus is a challenging task. The results from automatic language detection systems are often inaccurate and inconclusive. Consequently, the potential MY-EN-CS blogs were identified manually. The contents from 23 personal blogs were downloaded and pre-processed for MY-EN-CS identification. This paper used shallow lexical based approach to identify MY-EN-CS sentences. A total of 4,191 sentences were selected for the annotation task. As a result, 2,316 annotated MY-EN-CS sentences were selected for the corpus. The annotated task yields 0.83 Kappa value rate that indicate the high reliability of this corpus. This corpus contains subjective and objective MY-EN-CS sentences. The MY-EN-CS subjective sentences consist of words that are used to express personal opinion or emotion using Malay and English words. The result has shown imbalance distribution of the sentences in MY-EN-

CS corpus. Therefore, the corpus needs to be improved. Even though the corpus is imbalanced, the initial MY-EN-CS corpus provide a reliable platform to start with subjectivity and polarity classification for MY-EN-CS. The imbalanced annotated data in MY-EN-CS corpus reflects the actual situation of subjective and objective expression for English and Malay bilingual speakers. Consequently, this paper does not regard this issue as a methodological problem. For future work, this paper will use the corpus to identify and classify subjective and polarity sentences automatically.

References

[1] Muysken P. (2000). Bilingual Speech: A Typology of Code-Mixing. *Cambridge University Press*.

[2] Liu, B. (2015). Sentiment Analysis: Mining Opinions, Sentiments and Emotions. *Cambridge University Press*.

[3] Tang, H., Tan, S., & Cheng, X. (2009). A Survey on Sentiment Detection of Reviews. *Expert Syst. Appl.*, *36*(7), 10760–10773.

[4] Ravi, K., & Ravi, V. (2015). A Survey on Opinion Mining and Sentiment Analysis: Tasks, Approaches and Applications. *Knowledge-Based Syst.*, *89*, 14–46.

[5] Kasmuri, E., & Basiron, H. (2017). Subjectivity Analysis in Opinion Mining-A systematic Literature Review. *International Journal of. Advance Soft Computing and Its Application*, *9*(3).

[6] Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the Semantic Orientation of Adjectives. In *Proceedings of the 35th Annual Meeting on Association for Computational Linguistics*, (pp. 174–181).

[7] Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*, (pp. 347–354).

[8] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10*, (pp. 79–86).

[9] Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning Subjective Language. *Comput. Linguist.*, *30*(3), 277–308.

[10] Li, Y.M., & Li, T. Y., (2013). Deriving Market Intelligence from Microblogs, *Decis. Support Syst.*, *55*(1), 206–217.

[11] Esuli. A., & Sebastiani, F. (2005). Determining the Semantic Orientation of Terms Through Gloss Classification. In *Proceedings of the 14th ACM*

International Conference on Information and Knowledge Management - CIKM '05, (pp. 617).

[12] Thakor, P., & Sasi, S. (2015). Ontology-Based Sentiment Analysis Process for Social Media Content. In *Procedia Computer Science*, *53*, 199–207.

[13] Missen, M. M. S., Boughanem, M., & Cabanac, G. (2013). Opinion mining: Reviewed from Word to Document Level. *Social Network Analysis Mining*, *3*(1), 107–125.

[14] Makrynioti, N., & Vassalos, V. (2015). Sentiment Extraction from Tweets: Multilingual Challenges. In *Big Data Analytics and Knowledge Discovery: 17th International Conference*, (pp. 136–148). Springer International Publishing.

[15] Missen, M. M. S., Boughanem, M., & Cabanac, G. (2009). Challenges for Sentence Level Opinion Detection in Blogs. In *Proceedings of the 2009 8th IEEE/ACIS International Conference on Computer and Information Science*, (pp. 347–351).

[16] Van de Kauter, M., Breesch, D., & Hoste, V. (2015). Fine-Grained Analysis of Explicit and Implicit Sentiment in Financial News Articles. *Expert Syst. Appl.*, *42*(11), 4999–5010.

[17] Shoukry, A., & Rafea, A. (2012) Sentence-level Arabic Sentiment Analysis. In *International Conference on Collaboration Technologies and Systems*, (pp. 546–550).

[18] Ghorbel, H., & Jacot, D. (2011). Sentiment Analysis of French Movie Reviews. *Advances in Distributed Agent-Based Retrieval Tools SE - 7*, *361*, 97–108.

[19] Martín-Valdivia, M.T., Martínez-Cámara, E., Perea-Ortega, J. M. & Ureña-López, L. A. (2013). Sentiment Polarity Detection in Spanish Reviews combining supervised and unsupervised approaches. *Expert Syst. Appl.*, *40*(10), 3934–394.

[20] Amelio, A., & Pizzuti C. (2015). TASS: A Naive-Bayes Strategy for Sentiment Analysis on Spanish Tweets. In *2015 IEEE 27th International Conference on Tools with Artificial Intelligence*, (pp. 713–720).

[21] Becker, K., Moreira, V. P., & dos Santos, A. G. L. (2017). Multilingual Emotion Classification using Supervised Learning: Comparative experiments. *Information Processing & Management*, *53*(3), 684–704.

[22] Shenoy, P. D. (2015). HSAS : Hindi Subjectivity Analysis System. In *2015 Annual IEEE India Conference (INDICON)*, (pp. 1–6). IEEE

[23] Lo, S. L., Cambria, E., Chiong, R. & Cornforth, D. (2016). A Multilingual Semi-Supervised Approach in Deriving Singlish Sentic Patterns for Polarity Detection. *Knowledge-Based System*, *105*, 236–247.

[24] Vilares, D., Alonso, M. A. & Gómez-Rodríguez, C. (2017). Supervised Sentiment Analysis in Multilingual Environments. *Information Processing & Management*, *53*(3), 595–607.

[25] Gupta, D., Lamba, A., Ekbal, A. & Bhattacharyya P. (2016). Opinion Mining in a Code-Mixed Environment: A Case Study with Government Portals. In *Proc. of the 13th Intl. Conference on Natural Language Processing* (pp. 249–258).

[26] Wang, Z., Lee, S. Y. M., Li, S., & Zhou G. (2017). Emotion analysis in Code-Switching Text with Joint Factor Graph Model. *IEEE/ACM Trans. Audio, Speech, Languange Process*, *25*(3), 469–480.

[27] Hovy, E. & Lavid, J. (2010), Towards a 'Science ' of Corpus Annotation : A New Methodological Challenge for Corpus Linguistics," *International Journal of Translation*, *22*, 25.

[28] de Castilho., R. E., Mujdricza-Maydt. E., Yimam, S.M., Hartmann, S., Gurevych, I., Frank, A., Biemann, C. (2016). A Web-Based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities*, (pp. 76–84).

[29] Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communication of the ACM*, *38*(11), 39–41.

[30] Noor, N. H. B. M., Sapuan, S., & Bond, F. (2011). Creating the Open Wordnet Bahasa. In *25th Pacific Asia Conference of Language and Computation*, (pp. 255–264).

[31] Missen, M. M. S. & Boughanem, M. (2009). Sentence-Level Opinion-Topic Association for Opinion Detection in Blogs. In *Proceedings International Conference on Advanced Information Networking and Applications*, (pp. 733–737).

[32] Jain, V. K., Kumar, S., & Fernandes, S. L. (2017). Extraction of Emotions from Multilingual Text Using Intelligent Text Processing and Computational Linguistics. *Journal of Computational Science*, 1–11.

[33] Liu, Z., & Jansen, B. J. (2016). Understanding and Predicting Question Subjectivity in Social Question and Answering. *IEEE Transactions on Computational Social Systems*, *3*(1), 32–41.

[34] Vilares, D., Alonso, M. A., & Gómez-Rodrıguez, C. (2016). EN-ES-CS: An English-Spanish Code-Switching Twitter Corpus for Multilingual Sentiment Analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* , (pp. 4149–4153) .