






Singlish Checker: A Tool for Understanding and Analysing an English Creole Language

Lee-Hsun Hsieh , Nam-Chew Chua, Agus Trisnajaya Kwee, Pei-Chi Lo,
Yang-Yin Lee, and Ee-Peng Lim  

Living Analytics Research Centre, Singapore Management University,
Singapore, Singapore
eplim@smu.edu.sg

Abstract. As English is a widely used language in many countries of different cultures, variants of English also known as English creoles have also been created. Singlish is one such English creole used by people in Singapore. Nevertheless, unlike English, Singlish is not taught in schools nor encouraged to be used in formal communications. Hence, it remains to be a low resource language with a lack of up-to-date Singlish word dictionary and computational tools to analyse the language. In this paper, we therefore propose **Singlish Checker**, a tool that is able to help detecting Singlish text, Singlish words and phrases. To develop this tool, we first construct a large set of Singlish words and phrases by identifying different sources of Singlish words and their definitions and integrating them. We later propose a Singlish classifier model based on a BERT model fine-tuned with a large number of classified Singlish sentences. Our experiment show that the BERT-based classifier can achieved very high F1 performance, outperforming the baseline.

Keywords: Singlish · Singlish classification · Singlish dictionary

1 Introduction

Motivation. As English is a widely used language in many countries of different cultures, variants of English also known as English creoles have also been created. Singlish is one such English creole used by people in Singapore, a country of slightly more than 5.5 millions. Singlish has been influenced by other major non-English languages used in the country and they include, in decreasing order of popularity, Chinese (and its dialects such as Hokkien, Teochew, and Cantonese), Malay, and Tamil.

Consider the sentence example: “The weather is so hot, I buay tahan.” The word phrase “buay tahan” is Singlish and it combines a Hokkien (a southern Chinese dialect) word “buay” and another Malay word “tahan” carrying the meanings of “cannot” and “tolerate” respectively. The above sentence therefore says that a person cannot tolerate the hot weather. Without a good vocabulary of Singlish words or phrases, one will have difficulty understand the sentence. To

make understanding Singlish harder, an existing English word could acquire a new meaning when it appears in Singlish text. For example, the word “uncle” in Singlish often refers to some unrelated older man such as an old taxi male driver (“taxi uncle” in Singlish) and policeman (“police uncle” in Singlish). Singlish sentences may also have grammatical structures different from that of English.

Unlike English, Singlish is not taught in schools nor encouraged to be used in formal communications. Nevertheless, Singlish is still highly popular in informal conversations and online communications as the language has become an integral part of the Singapore identity. Due to its informal nature and a lack of formal research, Singlish is a low resource language with a lack of up-to-date Singlish word dictionary and computational tools to analyse the language. Several individuals have in the past tried to construct dictionaries of Singlish words and phrases. These dictionaries are however not integrated with one another and are not well utilized in any tools for analysing Singlish content.

Objective. In this paper, we therefore propose Singlish Checker, a tool that is able to help detecting Singlish text, Singlish words and phrases. Our long term goal is to develop a set of NLP tools for helping non-Singlish users understand the language and researchers analyse Singlish content for studying the language trends and the underlying content topics and sentiments. To develop the Singlish Checker tool, we first aim to construct a large pool of Singlish words and phrases by identifying and integrating different sources of Singlish words and their definitions. We also want to build a classifier to determine Singlish content. This classifier will be very useful to the construction of a large corpus of Singlish text for future research purposes.

Contributions. In the following, we summarize our research contributions:

- We have found six disparate sources of Singlish words/phrases and their definitions and integrated them into a combined list. This combined list covers more than 1600 distinct Singlish words/phrases, which to our best knowledge is the largest so far.
- We propose a Singlish classification model based on a BERT model fine-tuned with many likely Singlish sentences. Our experiment show that the classification model can achieved F1 performance of more than 0.9.
- We have developed a working Singlish Checker system which has a web-based demo interface to showcase the classification and Singlish word extraction functions.

Outline of Paper. We organize the rest of the paper as follows. Section 2 covers the relevant previous research. Section 3 outlines the construction of Singlish dictionary. We then propose our Singlish classification model and conduct experiment evaluation of the model in Sect. 4. The Singlish checker system is then presented in Sect. 5. Section 6 concludes the paper.

2 Related Works

Linguistic studies on Singlish. While Singlish is largely used in the informal settings, linguists have shown great interests in the analysis of this creole language,

finding the origins of Singlish words, and linking them with the user culture. For example, Wong studied the use of “one”, an English particle, in Singlish sentences and postulated that the frequent use of “one” can be attributed to Singlish users wanting to speak more definitively than the Anglo English speakers [14]. Gupta studied how Singlish is used in different settings on the web and suggested the use of Singlish is related to establishing the Singapore identity [6]. In [2], the use of Singlish particles such as “ah”, “lah”, “leh”, etc., in Singlish users’ social networks was analysed. The work concluded that some particles are more frequently used in personal topics than in non-personal topics. The use of particles is also influenced by the user’s ethnicity.

NLP Works on Singlish. As a creole language, there are limited NLP tools for analysing Singlish. Wang et al. was the first to construct Singlish dependency treebank using the Universal Dependencies scheme. They further proposed a neural stacking model-based Singlish parser integrated with English syntactic knowledge [13]. In a subsequent work, they further extended the treebank, and introduced a neural multi-task model-based parser which outperforms the previously proposed stacking model [12]. Both [12, 13] develop their parser models based on manually labelled Singlish sentences, and they do not address the recognition of Singlish words. To aid future Singlish NLP studies, Chow and Bond constructed computational grammar for Singlish, which was built upon standard English Head-driven Phrase Structure Grammar (HPSG) [3]. They included newly defined lexical types and rules to cover Singlish use-of-words. This work also does not address the recognition of Singlish words.

One popular task in NLP is sentiment polarity identification. Early works in Singlish sentiment polarity identification focused on constructing sentic dictionaries. For instance, Lo et al. employed a semi-supervised approach to derive Singlish sentic patterns [9]. Experiment results suggested that considering both English and Singlish sentic patterns lead to better sentiment prediction accuracy compared to English-only models. Ho et al. focused on concept-level sentiment knowledge base and proposed SenticNet, which consisted of semantics and sentics associated with Singlish words and phrases [7]. In a more recent study, Leow and Lo developed deep neural networks (DNNs) models for Singlish polarity detection [8]. They found that proper pre-processing is the key to high prediction accuracy. In addition, RNN was reported to outperform CNN in Singlish polarity classification task. In the above works, the Singlish sentences and words were manually determined.

Finally, there are also previous works that investigated the analysis of Singapore user generated content. Silva et al. predicted Singapore users’ personal value by analysing their social network content with Singapore-Linguistic Inquiry and Word Count (S-LIWC) [11]. S-LIWC is a text analytic toolkit which counts the use of words and captures writers’ psychometric properties specifically developed for Singlish content. [4] explored the linguistic factors that affect the popularity of Reddit posts in r/Singapore, where both Singlish (basilect) and standard English (acrolect) co-exist. Results showed that popular posts drawn on basilect-induced features as they connected better to the local audience.

3 Singlish Dictionary Construction

As we need a comprehensive and up-to-date set of Singlish words and phrases to determine known Singlish words and phrases in a sentence, we first search for existing online sources of Singlish words and phrases. With some efforts, we found the following six sources:

- **Coxford Singlish Dictionary (Coxford)**: This dictionary is located at <http://colingoh.com/project/the-coxford-singlish-dictionary/> and it covers 809 Singlish words and phrases. The dictionary has no longer been maintained nor updated since 2001.
- **Eat Drink Men and Women forum (EDMW)**: This is a forum at [Hardwarezone.com.sg](http://hardwarezone.com.sg) and the first post in this forum defines about 277 Singapore acronyms and Singlish words commonly used by the forum users. The forum can be found at <https://forums.hardwarezone.com.sg/eat-drink-man-woman-16/acronyms-lingo-peculiar-edmw-1738415.html>. We manually determine the Singlish words to be included in our research.
- **SinglishWiki Singapore Internet Lingo (SinglishWiki)**: This list of 103 Singlish words can be found at https://singternet.fandom.com/wiki/Singapore_Internet_Lingo_Wiki. It was last updated in June 2011.
- **Singaporelang (Singaporelang)**: This list of 113 Singlish words was compiled by Zinkie Aw in 2015. It is no longer available online.
- **Singlish Vocabulary Wikipedia pages (Wiki_Singlish)**: This list of 285 Singlish words appeared at https://en.wikipedia.org/wiki/Singlish_vocabulary
- **Singlishdictionary.com (SinglishDict)**: This list was compiled by Jack Tsen-Ta Lee and it covers 1,193 Singlish words and phrases (as at May 2016). This list can be found at <http://www.mysmu.edu/faculty/jacklee/>.

For data sources that involve webpages, we first develop separate crawlers to collect the Singlish words/phrases along with their definitions. Finally, we merge the words/phrases together, removing the duplicates (as they appears in more than one source), and removing the well known organization and location entities commonly mentioned in Singapore’s content. We finally obtain the integrated set of 1628 distinct Singlish words and phrases known as **Merged_SinglishDict**.

4 Singlish Classification

4.1 BERT-Based Singlish Classification Model

We adopt a classification model that uses a specially trained BERT model known as SinglishBERT as shown in Fig. 1. As SinglishBERT is a transformer-encoder trained with large amount of Singlish text, it is able to generate contextualized word embeddings for each word and special token of the input sentence. We shall elaborate the training of SinglishBERT in Sect. 4.2. The classification model takes the input sentence of n word tokens w_1, w_2, \dots, w_n , adds the special tokens [CLS] and [SEP] to be beginning and end of sequence, respectively, before

passing the token sequence to the SinglishBERT. The output embeddings of the [CLS] token generated by SinglishBERT is used as the representation of the sentence and fed to a feed forward neural network before the output is passed to a sigmoid cross entropy loss function for optimization so as to generate a prediction of Singlish label with a score between 0 and 1.

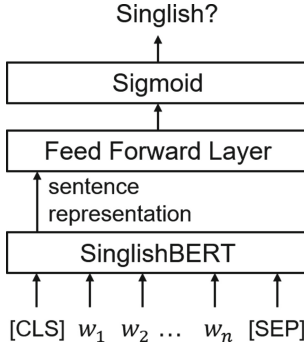


Fig. 1. Singlish Classification Model

4.2 SinglishBERT: A BERT-Based Classifier

SinglishBERT is an important module in our classification model responsible for generating contextualized token embeddings for every word and special token. Each token embedding is a vector representation that allows the token to be compared with other token. When two tokens from different sentences share the same sense or meaning, SinglishBERT is expected to return their embeddings to be close to each other in the representation space.

As the name suggests, SinglishBERT is a version of BERT which has been trained with large amount of English text to generate token embeddings [5]. Nevertheless, SinglishBERT is not developed by directly fine-tuning BERT. Instead, we fine-tune another BERT variant called SingBERT, which has been fine-tuned with a text corpus collected from subreddits *r/singapore* and *r/malaysia* (due to the similar way of speaking English in Singapore and Malaysia), and online forums popular among Singapore users [1]. As this corpus may still carry substantial amount of English text, we construct another text corpus to further fine-tune SingBERT to become SinglishBERT.

This new large text corpus consisting of 673,205 classified public Singlish tweet sentences generated by more than 150,000 Singapore users in 2020. These sentences are selected based on the following selection criteria:

- They have been assigned “en” language suggesting that the language used is likely to be English (or some English variants);
- Each sentence contains at least 5 and at most 30 words as an overly short or long sentence may be spam;

- Less than $\frac{1}{4}$ of the letters or characters in the sentence are in uppercase or non-English as we want to rule out spam and advertisement content;
- Less than a quarter of the words in the sentence are Malay words with the help of an online Malay dictionary; and
- Every sentence has a confidence score higher 0.999 assigned by our baseline classifier which will be introduced in the experiment section (see Sect. 4.3). We have intentionally selected a very confidence score threshold to ensure that the selected sentences are likely to be classified Singlish correctly by the baseline classifier.

With the above Singlish corpus, we fine-tune SingBERT to obtain Singlish-BERT.

4.3 Experiments

In this section, we conduct experiments to evaluate the accuracy of our BERT-based Singlish classifier and compare it with some baseline classifier.

Baseline Singlish Classifier. To compare against the above BERT-based classifier, we introduce a strong baseline Singlish classification model as shown in Fig. 2. The baseline classifier utilizes three distilled-GPT2 models: DistilGPT2(Singlish), DistilGPT2(English) and DistilGPT2(Malay). DistilGPT2 is a lightweight version of Generative Pre-trained Transformer 2 (GPT-2) [10]. It is not only a decoder that can generate a sentence, but also a language model that returns a perplexity defined as the exponentiated average negative log-likelihood of the input sentence. A small perplexity suggests that the distilled-GPT2 is likely to be able to generate the input sentence. DistilGPT2(Singlish) is a distilled-GPT2 pre-trained with English corpus consisting of about 100,000 standard English sentences generated by Singapore users. We also pre-train distilled-GPT2 with Singlish and Malay corpuses to obtain DistilGPT2(Singlish) and DistilGPT2(Malay) respectively. Our Singlish and Malay corpuses consists of 10,000 and 120,000 Singlish and Malay sentences respectively. When the same input sentence is fed to the three DistilGPT2 models, we obtain three perplexity scores corresponding to the three models. We concatenate these three scores into a feature set and pass to an SVM classifier with a Radial Basis Function (RBF) kernel.

Labeled Datasets. While the different classifiers are built upon different pre-trained models which use different pre-training datasets, we construct a common labeled datasets for training and testing the classification models. We first gather a set of 9060 manually labeled Singlish sentences from Reddit and Twitter as positive sentences. These sentences are authored by Singapore users who are active in subreddits covering Singapore matters or have indicated Singapore location in their Twitter profiles. The negative sentences consists of 20,000 manually labeled English sentences and another 20,000 manually labeled Malay sentences which are authored by Singapore users on Twitter. Malay sentences are chosen largely because Malay language is commonly used in Singapore and the language adopts the same alphabet as English.

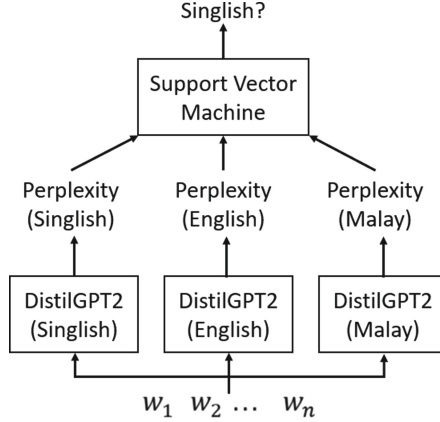


Fig. 2. Baseline Singlish Classifier

Accuracy Metrics. To report the model accuracy, we use the 5-fold cross validation approach dividing the labeled dataset into 5 disjoint subsets or folds of the same size such that each subset has the same proportions of positive (or Singlish) and negative (or English/Malay) sentences. We then use one of the folds for testing and remaining four folds for training the model. This evaluation is repeated for every fold used for testing. The final accuracy of prediction results over the five folds is obtained by averaging the fold-specific accuracy results. In our experiment, we define Singlish to be our target class and adopt the performance metrics: Precision, Recall and F1. Let S be a set of sentences. For a sentence $s \in S$, we use $g(s)$ and $p(s)$ denote the ground truth and predicted labels of the sentence s respectively. Precision (Pr), recall (Re) and $F1$ are defined as follows:

$$Pr = \frac{|\{s \in S | g(s) = p(s) = \text{“Singlish”}\}|}{|\{s \in S | p(s) = \text{“Singlish”}\}|}$$

$$Re = \frac{|\{s \in S | g(s) = p(s) = \text{“Singlish”}\}|}{|\{s \in S | g(s) = \text{“Singlish”}\}|}$$

$$F1 = \frac{2 \cdot Pr \cdot Re}{Pr + Re}$$

Results. Table 1 shows the accuracy results of our proposed BERT-based classifier and the baseline classifier. While the baseline classifier yields fairly good performance with an F1 of 0.842, BERT-based classifier still outperforms it by a substantial margin. This could be attributed to larger set of features (768 of them) provided by the BERT-based classifier compared with the three-only features used by the baseline classifier. The precision and recall results of BERT-based classifier are surprisingly very high contributing to the F1 of 0.944. This accuracy level also suggests that the model is ready to be deployed in real world applications.

Table 1. Classification results

	Precision	Recall	F1
Baseline Classifier	0.897	0.793	0.842
BERT-based Classifier	0.957	0.932	0.944

5 Singlish Checker: System Features and User Interface

Fig. 3. Singlish Checker UI: Input sentence

In this section, we describe the web-based Singlish Checker application that utilizes both the `MergedSinglistDict` and Singlish classification model. Functionally, this application takes a sentence as input and returns the a score between 0 and 1. When the score is closer to 1 when the sentence is predicted to be Singlish, and 0 otherwise. Figure 3 depicts the user interface of Singlish Checker which contains a text box for one or more input sentence. In this figure, we show the sentence “The weather is so hot, I buay tahan.”. When the input consists of multiple sentences, the Singlish Checker will return the average of sentence scores instead. To illustrate the use of Singlish Checker, the user interface includes several example sentences for the user to select. Once the input sentence(s) is entered, the user should click on the “Submit sentence” button.

Figure 4 depicts the results after the user submit the input sentence. The results include a sentence score, and the list of Singlish words or phrases detected

Sentence analysis result:

Input sentence: *The weather is so hot, I buay tahan.*

✓ Sentence score (1: Singlish, 0: Non-Singlish): **0.9994**

✓ Singlish word(s) detected from input sentence.

- **buay tahan** (confidence score: 0.7948)

1. *Combination of the Hokkien term buay and Malay term tahan. Means unable to withstand or colloquially cannot stand it i.e intolerable. (From [Singlish Vocabulary Wiki](#))*
2. *bœ ta:ha:n | [buay-tar-han] slang phrase. Cannot stand it anymore. Also see 'tak boleh tahan'. Example: 'I buay tahan already, next time he bullies another kid I will scold him!' From Hokkien. There is a variant from Malay too - 'Tak Boleh Tahan'. 本地俚语: 再也无法忍受。 口语例句: (源自福建话) '我buay tahan, 为什么他总爱欺负小孩子?' 与马来语'tak boleh tahan'有一些相似。 frasa slang. Tidak boleh tertahan lagi. Contoh bahasa percakapan: 'Aku dah buay tahan, kenapa dia suka sangat buli budak itu?' Daripada bahasa Hokkien dan Melayu, "tahan" ialah perkataan Melayu. Ada juga variasi frasa ini dalam bahasa Melayu - 'Tak boleh tahan'. *புவே தாஹான் • கொச்சை வழக்கு சொற்றொடர். இனிமேலும் தாங்க முடியலை. பேச்சு வழக்கு உதாரணம்: அவன் ஏன் அந்த சின்ன படயனை கொடுமைபடுத்திகிட்டே இருக்கிறான்? என்னக்கு புவே தாஹான் வா! ஹோக்கியேன் மொழியிலிருந்து வந்த சொல். மலாய் மொழியில்: 'தக் பொலெஹ் தாஹான்'. (From [Singaporelang - What the Singlish](#)) [Less](#)**
3. */tah-hahn, 'ta: ha: n/ a. phr. [Mal. Tahan] Be unable to endure or stand a person, situation, etc., any longer. 2000 Leong Liew Geok "Forever Singlish" in Women without Men 130 .. like when the secretary say / You hold on arh, he's on another line; / So you wait for him to finish - wah piang, talk / So long, boey tahan, some more I kena / Scolding from boss for wasting time. 2001 John Chen The Straits Times, 30 October, H2 Buay tahan.. I was shocked out of my wits.. 'Buay tahan' is Hokkien for 'couldn't stand it anymore'. 2003 Suzanne Sng (quoting Faizah Abdullah) The Sunday Times (LifeStyle), 18 January, L10 Buay tahan. The noise is so loud at night. 2004 Lim Han Ming (quoting James Wong) Streets, 7 May, 9 I already 'buay tahan' (Hokkien for 'cannot cope') with just two kids. (From [Jacklee - A Dictionary of Singlish and Singapore English](#)) [Less](#)*

Fig. 4. Singlish Checker UI: Sentence Results

in the sentence by checking against *MergedSinglishDict*. In this example, the sentence score of 0.9994 returned by our classification model suggests that the sentence is Singlish. For each Singlish word or phrase, Singlish Checker also returns its confidence score, its corresponding definition(s), or meaning(s). The confidence score is assigned by a Singlish word phrase recognition module. Due to space constraint, we shall not elaborate this module in this paper. In the figure, the word phrase “buay tahan” has been detected with a confidence score of 0.7948. There are three definitions found for this word phrase and they come from *Wiki_Singlish*, *Singaporelang* and *SinglishDict* respectively. Note that these definitions have not yet been integrated and shall be a topic for future research.

6 Conclusion

This paper introduces Singlish as a English creole language widely used in Singapore. To help analysing this low-resource language, our research integrates several data sources of Singlish words and their definitions. This combined set of Singlish words to our knowledge serves as the largest Singlish dictionary. To detect Singlish sentences or to select them for downstream research or analysis, we propose a BERT-based classification model utilizing *SinglishBERT* a BERT variant pre-trained or fine-tuned with a large Singlish corpus. Through our experiments, we show that our proposed classification model can achieve very accurate

results. We finally deploy the classification model and Singlish word extraction based on the combined Singlish dictionary in our web demo application Singlish Checker.

As part of our future work, we first plan to automate the recognition of Singlish words/phrases. This will be very important for the understanding of Singlish trend over time. As we recognize new Singlish words/phrases, it is also an important research topic to determine their meanings and senses to help users fully understand the Singlish content.

References

1. Zanelim/singbert. hugging face. <https://huggingface.co/zanelim/singbert>,. Accessed 31 Dec 2010
2. Botha, W.: A social network approach to particles in Singapore English. *World Englishes* **37**(2), 261–281 (2018)
3. Chow, S.Y., Bond, F.: Singlish where got rules one? constructing a computational grammar for Singlish. In: LREC (2022)
4. Chua, H.: Stylistic approaches to predicting Reddit popularity in diglossia. In: ACL (2021). <https://doi.org/10.18653/v1/2021.acl-srw.10>
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
6. Gupta, A.F.: Singlish on the web. In: *Varieties of English in South East Asia and Beyond*, pp. 19–37. University of Malaya Press (2006)
7. Ho, D., Hamzah, D., Poria, S., Cambria, E.: Singlish SenticNet: a concept-based sentiment resource for Singapore English. In: 2018 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1285–1291 (2018)
8. Leow, Y.S., Lo, S.L.: Singlish polarity study using deep learning. In: *First International Workshop on Social Media Analytics for Smart Cities (SMASC)* (2017)
9. Lo, S.L., Cambria, E., Chiong, R., Cornforth, D.: A multilingual semi-supervised approach in deriving Singlish sentic patterns for polarity detection. *Knowl.-Based Syst.* **105**, 236–247 (2016). <https://doi.org/10.1016/j.knosys.2016.04.024>
10. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. In: *NeurIPS EMC2 Workshop* (2019)
11. Silva, A., Lo, P.C., Lim, E.P.: On predicting personal values of social media users using community-specific language features and personal value correlation. In: *ICWSM*, pp. 680–690 (2021)
12. Wang, H., Yang, J., Zhang, Y.: From genesis to creole language: transfer learning for Singlish universal dependencies parsing and POS tagging. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **19**(1), 1–29 (2019)
13. Wang, H., Zhang, Y., Chan, G.L., Yang, J., Chieu, H.L.: Universal Dependencies parsing for colloquial Singaporean English. In: *ACL* (2017). <https://doi.org/10.18653/v1/P17-1159>
14. Wong, J.: “Why you so Singlish one?” a semantic and cultural interpretation of the Singapore English particle one. *Lang. Soc.* **34**(2), 239–275 (2005)