12-2018

On learning psycholinguistics tools for English-based Creole languages using social media data

Pei-Chi LO

Ee-peng LIM
Singapore Management University, eplim@smu.edu.sg

On Learning Psycholinguistics Tools for English-based Creole Languages using Social Media Data

LO Pei-Chi
School of Information System
Singapore Management University
Singapore
pclo.2017@phdis.smu.edu.sg

LIM Ee-Peng
School of Information System
Singapore Management University
Singapore
eplim@smu.edu.sg

Abstract—**The Linguistic Inquiry and Word Count (LIWC) tool is a psycholinguistics tool that has been widely used in both psychology and sociology research, and the LIWC scores derived from user-generated content are known to be good features for personality prediction [1], [2]. LIWC, however, is language specific as it relies on counting the percentage of pre-defined dictionary words occurring in the content. For content written in English Creoles which are languages based on English, the original English LIWC may not perform optimally due to its lack of words which are only used in the English Creoles. In this paper, we therefore study the learning of LIWC for an English Creole using *word embeddings*, a way to encode contextual meaning of words in a vector representation. We particularly focus on an English Creole known as Singlish (which is a popular English creole in Singapore and it contains words from non-English languages including Malay, Chinese, Chinese dialects, and Indian languages). Instead of a manual effort to construct LIWC for Singlish, we automate the construction of a Singlish-specific LIWC dictionary, called S-LIWC by learning a word embedding model using a large corpus of Singapore tweets, and extracting new words semantically similar to the LIWC dictionary words. We show that the S-LIWC can be used to predict LIWC summary variables. Moreover, we conduct a personality prediction experiment on Singapore university students using their Facebook status updates. Our results show that our personality prediction method using S-LIWC outperforms that using LIWC for most personality traits. We finally show some interesting case examples of explaining the weaknesses and strength of S-LIWC.**

Index Terms—**Language-specific psycholinguistics tool, Personality prediction**

I. INTRODUCTION

A. Motivation

The Linguistic Inquiry and Word Count (LIWC)[1] is a widely-used tool in psychological research using textual information. It has a program which counts the percentage of words reflecting emotions, part-of-speech tag, writing style, and even social status in given text. The latest version of LIWC, LIWC2015, consists of 92 categories for extension and comparison. Other than derived score measures for the summary variables (e.g., analytical thinking) and language

---

[1]http://www.liwc.net

metrics (e.g., word per sentence), LIWC has different word categories each covering a word list. For example, its *money* word category covers a list of words including "atm", "bank", "cash", "debt", and "tax".

Unlike the earlier versions, LIWC2015 also includes four new *summary variables*, namely, *Analytical Thinking*, *Clout*, *Authentic*, and *Emotional Tone*, to capture the degree of analytical thinking, social status, honesty, and emotion expression in the content respectively. The score formulas of these summary variables have not been published, but are said to be derived using an in-built dictionary [2], [3]. Despite multiple attempts to decipher the score formulas, none of them could validate their formulas so far.

LIWC has been utilized in a variety of research tasks including social status prediction, linguistic style differentiation, and even depression detection [4]. LIWC has been shown to be effective for both formal and informal textual content such as social media content [1], [2]. As LIWC relies on counting the pre-defined dictionary words specific to English content, its applicability has been limited to mostly English content. While there are non-English variants of LIWC, it requires extensive efforts to create these LIWC variants for analyzing non-English content. Most of the word lists for these non-English LIWC's are compiled by multiple psychology experts over a long period of time [5]–[7].

Even when a language is English-like, the peculiarities in speaking and writing styles as well as in the choice of words may still require a different LIWC to be constructed. The construction of such LIWC variants for English-based languages remains to be challenging.

For example, Singapore is a city-state comprising immigrants from China, India, Malaysia and other countries. Table I shows the composition of languages spoken in Singapore. Over the years, with English used as the common language among Singapore users from different origins, English has morphed into a creole language known as *Singlish*. Singlish is based on English but incorporates words and lexical rules from non-English languages including Chinese, Chinese dialects, Bahasa Melayu (i.e., Malay), and even Indian languages. Naturally, it

Language	Proportion
Madarin	36.3
English	29.8
Malay	11.9
Hokkien (Chinese Dialect)	8.1
Cantonese (Chinese Dialect)	4.1
Tamil	3.2
Teochew (Chinese Dialect)	3.2
Others	3.4

is a language which can only be understood by Singapore users [8]. In the written form, non-English alphabetical words are invented. For example, in the Singlish sentence *"The weather is quite jialat today"*, the word *"jialat"* is a southern Chinese dialect word that describes a bad or disastrous situation. There are also grammatically incorrect sentences in Singlish. For example, the tweets *"Why is the train so crowded? It's driving me crazy"* and *"Wah MRT so crowded now siao liao loh"* are written by Singapore users. While they share the same meaning, the former uses pure English and the latter uses both Chinese-dialect and English with a loose grammar. Due to these non-English words and sentence structures, analyzing Singlish content using LIWC may lead to compromised accuracies in different content analysis tasks.

B. Research Objectives and Approach

In this work, we therefore aim to develop an big data approach to learn a LIWC variant for an English-based creole language. This automated approach hopefully will reduce the amount of efforts of LIWC construction without compromising the accuracy of content analysis. The data to be used for learning the relevant words in the Creole language come from social media making this proposed approach generalizable to other similar languages. In this project, we specifically apply this approach to learn S-LIWC, a LIWC variant for Singlish.

Challenges. There are several challenges in this research. Firstly, the automated construction process requires an expansion of word dictionary for different LIWC word categories. Particularly, we need to determine words that are semantically similar to existing words in the original LIWC word dictionary. We want to exploit the existence of both English words and their corresponding non-English words in the content. Learning a good semantic similarity measure for English and Singlish words based on the context of their occurrences is necessary. Ideally, this process does not require human efforts or labeled training data. For example, the LIWC's dictionary words "lazy" and "boring" hopefully can be determined to be similar to their equivalences "malas" and "sian" in Singlish respectively.

The second major challenge involves the validation of the new LIWC variant. In the past, much human efforts have been spent on validating LIWCs. In our research, we propose

to validate S-LIWC both manually and by task. The former involves human annotators to determine the correctly inferred relevant LIWC dictionary Singlish words returned by our proposed S-LIWC construction method. The latter applies S-LIWC to personality prediction task using the social media content generated by Singapore users. For S-LIWC to be of good quality, we expect relevant Singlish words to be inferred for S-LIWC, and the personality prediction accuracy using S-LIWC on Singlish social media content to be better than that using the original LIWC.

Key Ideas. Unlike construction of LIWC word dictionary for pure non-English languages, both English and non-English words co-exist in an English-based creole language such as Singlish. We therefore expect the English LIWC words and their similar non-English words to be found in a creole-language content corpus. We therefore exploit the similar context of similar words in the corpus using a word embedding model. Word embedding model is an emerging technique that learns a distributed representation of words that capture the semantic of the words' context. For example, if "jialat" and "bad" are semantically similar words, they will be mapped to nearby locations in the embedding space. One could therefore search the neighborhood of every LIWC word to find the other similar Singlish words so as to construct a new LIWC variant.

The second idea is to derive the summary variables of English-based creole language, even though the formulas to compute these variables are not published. So far, most of the non-English LIWC variants including Dutch-LIWC and Chinese-LIWC are not able to return summary variable scores. In S-LIWC, we want to be able to predict summary variable values using the new word dictionary.

The third idea in this work is to predict Big-Five Personality of Singapore social media users using S-LIWC, and to evaluate the results against that using the original LIWC. Personality prediction is useful in many user profiling and recommendation applications, and past works have shown that LIWC can be used to predict personality with good accuracy [1].

Contributions. In the following, we summarize our novel contributions as follows:

- We propose an automated approach to create LIWC for English-based creole languages based on word embeddings learned from more than 160 millions tweets generated by about 150,000 Singapore users. By learning the words' context, we are able to find similar words to match words found in the original LIWC word dictionary.
- We apply the proposed approach on creating S-LIWC, a LIWC variant for Singlish. We present the results of using S-LIWC to derive the summary variables. We demonstrate that most of the summary variables can be predicted with higher accuracy compared with using LIWC.
- We also evaluate LIWC and S-LIWC in the personality prediction task using a Singlish Facebook dataset. This experiment shows that using S-LIWC yields better prediction accuracy than LIWC.

II. RELATED WORK

In this section, we review the related works on cross-lingual LIWC construction and user profiling using LIWC. We compare them with our work on English-based creole LIWC construction and personality trait prediction using LIWC respectively.

A. Construction of The Linguistic Inquiry and Word Count.

LIWC was first proposed in 1990s [10] to determine a person's psychological state and trait by his or her written English text. LIWC can be applied to the prediction of attentional focus, emotionality, social status, and even deception [4]. It has therefore been widely used among psychology research.

The development of LIWC 2011 is based on linguistic psychology studies, synonym dictionaries, and word list such as PANAS [11]. In the construction process, a word list was first generated for each word category. Every word in the list is then judged separately by 3-6 judges to decide whether it is to be kept in the final word list. Multiple rounds of judgement are then conducted by psychology experts to determine the modification of word list until they achieve at least 93% agreement across all word categories. This process is known to be labor-intensive and time consuming. To our knowledge, there is no work automating the construction of LIWC or LIWC variants.

Beyond English-written content, there are also works that translate LIWC into different languages. For example, Huang et al. translated the original LIWC to traditional and simplified Chinese, following the manual process similar to that for the original LIWC [5]. Wolf et al. found German equivalences of LIWC word list and demonstrated the robustness of German-LIWC by applying it to determine the text quality of E-Mails [6]. The LIWC dictionary is subsequently and manually translated to Dutch [7], Spanish, French [12], Russian, Italian [13], and Brazilian [14]. Other than manual compilation, Van and Boot proposed a framework to build Dutch LIWC using Google translation [15]. To improve the machine translated result, they design a pipeline that filters out wrongly translated words, adds function words and removes mis-categorized words by referring to dictionary and online word lists. Google translation is however not available for most English-based Creole languages.

B. LIWC Applications in Personality Prediction.

The prediction of personality traits using human-generated content has been widely studied in both psychology and computer science domains. Golbeck et al. combined LIWC with structural features, activity and preference, as well as personal information reported by users on Facebook to predict their personality [1]. They further showed that good prediction accuracy can be achieved using LIWC features using the Twitter content generated by users [16].

Instead of simply LIWC features and other structural features, Wei et al. proposed a framework to predict Big-Five personality using heterogeneous information obtained from Weibo which include text, avators, emoticons and response patterns [17]. They suggested that the combination of different kinds of data can yield more than 30% accuracy improvement in personality trait prediction over that using LIWC features only [18].

III. DATASETS

There are several datasets used in this research. They are grouped by the way we use them as described below.

A. Dataset for learning S-LIWC.

SG Twitter Dataset: We first construct a Twitter dataset that contains Singlish content so as to construct the word dictionary for S-LIWC. This dataset is constructed by first identifying a set of well known Singapore Twitter user accounts as seeds. From these seed accounts, we crawl the followers and followees selecting those based in Singapore (according to the profile location of these accounts) to be added to our Singapore user set. We then repeat the crawling and user selection steps on these newly selected users until no more users can be added. For the final set of about 150,000 users, we crawl all their tweets posted from January 2017 to July 2017. This amounts to about 161 million tweets. Subsequently, we will use this SG Twitter dataset for word embedding model learning.

B. Datasets with Ground Truth Personality Scores.

We use two datasets with user contributed social media content and ground truth personality scores assigned to users. These personality scores are obtained by users completing the personality questionnaire survey.

GW Dataset: This dataset consists of all Facebook posts of 93 university students in Singapore collected in an earlier work [19]. Many of these posts contain Singlish words. Each student in the dataset has at least one Facebook post, and 59 of them have completed a 50-item-IPIP-FFM personality survey [20]. The GW Dataset includes personal attribute information including full name, gender, education background, and Facebook pages liked by the students, but these attributes are not used in this work.

MyPersonality Facebook Dataset (myP): MyPersonality dataset is a publicly available dataset for the shared task in Workshop on Computational Personality Recognition, 2013 [21]. 9917 Facebook posts from 250 Facebook user are included, as well as the users' Big-Five Personality scores. Note that the content of this dataset is mainly in English.

C. Dataset with Social Media Content in Singlish:

GenFB: We construct this dataset by collecting 127,339 Facebook posts from 115 Singapore politician's official Facebook fan-pages. These are politicians active in the Singapore's Facebook scene. Among the many Facebook users posting on these fan-pages, We select 1500 users who have public profiles, and collect all their posts. In the end, we have 708,243 posts which include some Singlish content. Note that we do not have ground truth personality scores for users in this dataset. Hence, this dataset will be used for Summary Variable prediction only.

IV. SINGAPORE-LIWC (S-LIWC)

In this section, we describe the approach to construct Singapore-LIWC word dictionary. The key idea is to expand the existing LIWC word dictionary with new non-English words using a word embedding model, e.g., word2vec [22]. In word embedding, every word is assigned a vector in a n-dimensional vector space such that it is close to other words sharing the similar semantics. We elaborate the steps of this proposed approach as follows:

Step 1: Text Corpus Construction. Word embedding is an unsupervised learning model that is performed on a large text corpus. In our research, we use the SG Twitter Database as the text corpus as it contains the Singlish content covering both English and non-English words. The SG Twitter Database also covers words from LIWC word dictionary. Hence, we can use word's context to find the new words to be added to any of the 92 LIWC word categories.

Step 2: Text Preprocessing. In this step, we remove URLs, user tags (e.g., @userid) and punctuations from the tweet content as we do not expect them to be added to S-LIWC. We also convert all hashtags (e.g., #sg50) to normal words by dropping the # symbol as many hashtags are known to carry meaningful information. Finally, all words are converted to lowercase.

Step 3: Language-Specific Word Tokenization. This step aims to tokenize sentences in tweets into sequence of words. Each language requires a different NLP library to perform word tokenization. We therefore first detect the language(s) used in each tweet at the sentence-level.

For example, the tweet "The weather is quite nice today. 出去玩咯" contains an English sentence followed by another Chinese sentence. We segment the tweet into sentences using NLTK sentence tokenizer[2]. The Python package langid[3] is then used to detect the language of each sentence. Finally, we tokenize every detected Chinese sentence into words using Jieba[4], and NLTK word tokenizer for other language sentences. We also find sentences written in multiple languages. In that case, we use Jieba for the tokenization since Jieba can handle English word segmentation while NLTK could not handle Chinese words.

Step 4: Word Embedding Modeling. We concatenate all words in a tweet to create a document, and train a 500-dimensional CBOW word embedding model using gensim[5]. This word embedding model helps us find the words most similar to a given target word. Notation wise, we use w to represent a word, and $E(w)$ to denote the corresponding 500-dimensional embedding vector representation. Since the model is trained using Singapore tweets, we expect some Malay, Chinese Dialect words, Indian and netspeak words to be included.

Step 5: Candidate Word Selection. The original LIWC's word dictionary consists of 16,343 seed words. To construct the S-LIWC's word dictionary, we find the top K most similar words for each seed word in LIWC's word dictionary using cosine similarity. Formally, the similarity between a seed word w_s in LIWC's word dictionary and a word w in SG Twitter Database is defined by the cosine similarity of $E(w_s)$ and $E(w)$. These K similar words are then added to every LIWC word category the seed word belongs to. In our experiment, we empirically set $K = 10$ to ensure that we can always find sufficient new words similar to the seed words in LIWC. Note that, the candidate similar words of different seed words can be overlapping. At the end of this step, we obtain 23,999 distinct candidate words.

Step 6: Word polarity disambiguation. In LIWC's sentiment-carrying categories (e.g., Positive Emotion (Posemo) category which covers 542 words, and Negative Emotion (Negemo) category which covers 609 words), words with opposite semantics may be wrongly assigned very similar word embedding vectors due to their similar context in the training text corpus. Moreover, in Step 5, we also found words that are irrelevant to the seed word added as candidate words. Such noises might be hashtags or names of company and its product (e.g., *Pandora* and *charm*). The word polarity discrimination step is thus introduced to distinguish noises from the real synonyms as described in [23]. This step essentially exploits the directional displacement between words in embedding space that captures some semantic relationship. Figure 1 shows an example of two words "good" and "bad" related by opposite semantic meanings and represented in the embedding space. As described in [23], any other vector pairs w_s and w_s' with such directional displacement might have similar semantic relationship.

We therefore follow this intuition and train a classifier to distinguish noises from actual synonyms which we want to add to S-LIWC. We collect synonyms and antonyms of LIWC seed words from Oxford Dictionary API[6] as the ground truth. As in the dictionary there exists multiple senses for each word, we recruit an annotator to manually select the senses that fit the LIWC word category. After filtering, 2,000 antonym pairs (negative samples) and 2,000 synonym pairs (positive samples) are collected to train our classifier. After all the data points are collected, we extract the word vectors from our pre-trained word embedding model, and use the hadamard product of the two word vectors of a word pair as input to the Logistic Regression classifier. This classifier achieves precision and recall of 83.63% and 63% using 10-fold cross validation. This suggest that with the classifier, we are able to filter out more than half of the noises from the candidate word lists obtained in step 5.

In order to evaluate the performance on polarity disambiguation, we recruit a native Singaporean volunteer to manually evaluate the quality of our S-LIWC dictionary before and after disambiguation. 2,000 seed-candidate pairs across all

---

[2]http://www.nltk.org
[3]https://github.com/saffsd/langid.py
[4]https://github.com/fxsjy/jieba
[5]https://radimrehurek.com/gensim/index.html

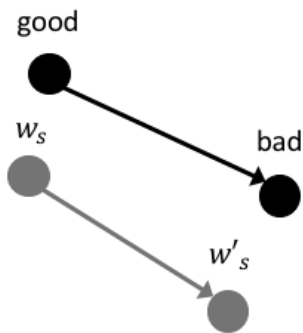[6]https://developer.oxforddictionaries.com/

Fig. 1. Extraction of Opposite Sentiment Polarity in Word Embedding Model

categories are randomly selected and assigned to the volunteer. For each pair, the volunteer can choose one of the following three options:

1) The seed and candidate word are simialr words and thus should be put under the same LIWC category,
2) the seed and candidate word are not similar, but the candidate word is compatible with the LIWC category the seed word belongs to, and
3) the seed and candidate word are not similar, and the candidate word is incompatible with the LIWC category the seed word belongs to.

We compute *strict precision* (i.e., only regard the first option as a successful extension) and *relaxed precision* (i.e., both the first and second options are consider successful) for each word category. The unfiltered S-LIWC gets 5% for strict precision, and 16.5% for relaxed precision. After disambiguation, the average strict precision and relaxed precision of all word categories reaches 42% and 65% respectively, and among all word categories, Function Words achieves the best performance with 45.2% strict precision and 70.3% relaxed precision. The result shows that our method can effectively remove irrelevant words from the candidate list, and generate decent word lists.

Finally, we obtain a list of 9,640 words and the LIWC categories they belong to, and the list is later converted to LIWC application readable format. Table II shows five LIWC categories with the most words added in, and some example words that are not able to be found in English thesaurus or dictionaries.

V. SUMMARY VARIABLE PREDICTION

Task Setup. As we have constructed the S-LIWC data dictionary, we now conduct a summary variable prediction study using S-LIWC as part of our evaluation. Since LIWC2015, LIWC provides the scores of four summary variables measuring *authenticity*, *clout*, *analytical thinking* and *emotional tone* of any given piece of content. Summary variables are important in user profiling as they directly characterize the abilities and social status of the user.

Each summary variable has its value between 0 and 100, and is said to be derived from a pre-defined set of *word*

category features using a proprietary formula. Every word category feature is computed from a standard set of words, and the feature score is converted to percentile that reflects the content author's position in a normal distribution curve. Other than analytical thinking summary variable [2], the definitions of other summary variables are not revealed in the research literature [3], [24], [25]. Some of the prior research suggested formulas for authentic and clout variables but failed to provide the correct coefficients of the formulas.

To conduct evaluation of S-LIWC in the summary variable prediction task, we need to derive the the correct formula that maps the LIWC word category features to each summary variable. We are left with three summary variables without published formulas.

We learn the formulas using a English-based social media dataset, i.e., myP dataset, as training data. With the learned formulas, we can then perform summary variable prediction for the GW Dataset and GenFB Dataset using either using LIWC or S-LIWC. This is possible because both LIWC and S-LIWC share the same set of word category features even when their category words are different. As part of the learning step, we measure the error of a summary variable prediction by the difference between the ground truth summary variable score and the predicted summary variable score returned by the corresponding learned formula using LIWC word category features. If the prediction error is small, the learned model is considered accurate and can be appropriately used together with word category features of S-LIWC.

Feature selection. We learn a linear regression model for each summary variable based on three kinds of features:

- *Formula features*: These are scores returned by important features and coefficients found in the existing known formulas (only for analytical thinking, clout and authenticity) [3], [24], [25].
- *Source features*: These are Word category features included in the proposed formulas [3], [24], [25], and
- *Correlated features*: We conduct Spearman Correlation Test to determine the correlated LIWC word category scores and Summary Variable scores using the training dataset. Word categories having p-value<0.1 are considered significantly correlated and hence their scores are used as features as shown in Table III. As emotional tone summary variable does not have a previously proposed formula, we use only the highly correlated word category scores as features.

As the linear regression model is designed to recover the unpublished formula of each summary variable, we want the model to predict the summary variable values returned by the LIWC2015 package as accurate as possible.

Evaluation on English-only Content. This experiment aims to determine if our trained formulas or models can predict Summary Variables accurately for English-only content. Table IV shows the Mean Average Error (MAE) result of Summary Variables predicted by Linear Regression. All experiments were conducted on the MyP dataset using 10-fold

TABLE II
TOP 5 WORD CATEGORIES AND ADDED CANDIDATE EXAMPLES

LIWC category	Distinct Words Added	LIWC seed	Candidate	Source Language	Meaning
Affect	2,296	**fantasi**	kenikmatan	Malay	Fun
		unlucky	suay	Chinese dialect	Unlucky
		disturb	disiao	Chinese dialect	To irritate
		adorable	kyoot	Netspeak	Cute
		bestie	bestf	Netspeak	Best friend
Relativ	1,848	**finishing**	chionging	Chinese dialect	Rushing to finish something
		stop	stahp	Netspeak	Stop That and Halt Please
		age	岁	Chinese	Age
Bio	1,656	**noodle**	mee	Chinese dialect	Noodles
		tea	teho	Chinese dialect	Black tea
		dinner	iftar	English	Meal at sunset for Muslim
Verb	1,325	**couldve**	coulda	Netspeak	Abbreviation for "could have"
		excite	eggcited	Netspeak	Excite
		hang	lepak	Malay	Relaxing/Chilling
		shook	sh00k	Netspeak	Shook
		sigh	haiz	Malay	The sound of Sigh
Negemo	1,237	**hate**	h8	Netspeak	Hate
		gossip	mengumpat	Malay	Gossiping

cross validation. The results show that the prediction of authenticity and clout summary variables using highly correlated word category features outperforms both formula and source features. For analytical thinking, only the formula features are used (i.e., formula from [2] and the linear regression yields very small MAE results of 7.36.

Evaluation on English-based Creole Content. We now apply the learned prediction models (trained using MyP dataset) on both GW and GenFB datasets using S-LIWC word category features. As GW dataset is small, we use Leave-One-Out cross validation. For GenFB dataset, we apply 10-fold cross validation. Recall that GW and GenFB have both English and non-English content. We evaluate the error between (i) the summary variables returned by LIWC2015 for the English content, and (ii) the summary variables predicted by our learned model for the non-English content. We would like this error to be small.

As shown in Table IV, the prediction models yields highly accurate results. For Analytical Thinking, the MAE results for both GW and GenFB datasets are superior than that for MyP dataset. For other summary variables, the MAE results are comparable across the three datasets. GW dataset particularly sees significantly better MAE using correlated features. These results suggest that our learned models using correlated S-LIWC word category features can already predict summary variables with good accuracy. For Authenticity and Clout, the models using the correlated features perform better than other features.

Finally, we examine how our predicted summary variables using S-LIWC are correlated with the ground truth summary variables returned by the LIWC2015 tool. We use the predic-

tion model trained with S-LIWC scores using myP dataset, and test on GW and GenFB datasets using S-LIWC scores. The result is shown in Table V. For both GW and GenFB datasets, our predicted scores for all the LIWC summary variables are significantly correlated to the original summary variable scores (p-value smaller that 0.01). The result suggests that we can accurately determine the summary variable scores using our trained linear regression model and S-LIWC features.

VI. PERSONALITY TRAIT PREDICTION

In this task, we focus on evaluating both LIWC and S-LIWC in the task of predicting personality traits of Singapore users. We want to determine if S-LIWC is more effective than LIWC when used for this task. In this task, both GW and myP datasets are used in model training and testing as their users have personality scores. Recall that the content posts in GW dataset carry Singlish content, while the content posts in myP dataset are generated by native English speakers. The mean and standard deviation of personality scores in each of the Big-Five personality traits of users from myP and GW datasets are shown in Table VI. We also include the Singapore average personality scores mentioned in a past survey [26] in the table for reference.

A. Selection of Correlated LIWC/S-LIWC Features

We conduct Spearsman Correlation Test between LIWC word category features and each Big-Five Personality Trait on both the GW and MyP datasets to determine the word category features that are important to predicting each personality trait. For the GW dataset, Extraversion is shown to be correlated to word category features "you", "family", "social" and "we". Agreeableness is highly correlated with emotion-related LIWC

Emotional Tone		Authenticity		Clout	
LIWC	ρ	**LIWC**	ρ	**LIWC**	ρ
posemo	0.75336	relativ	0.72374	social	0.70567
negemo	-0.53696	time	0.58482	i	-0.53035
affect	0.35032	i	0.52651	affiliation	0.42113
anger	-0.33946	motion	0.48782	you	0.38076
swear	-0.29497	space	0.4763	we	0.36052
risk	-0.28594	prep	0.40572	shehe	0.32711
sad	-0.28496	adverb	0.34494	drives	0.3176
affiliation	0.27497	shehe	-0.32281	negate	-0.28078
sexual	-0.27493	function	0.31203	differ	-0.27166
relig	0.27299	female	-0.2751	female	0.25802
adj	0.2526	achieve	0.2726	they	0.24769
netspeak	0.22384	Sixltr	-0.26149	adverb	-0.24551
assent	0.21558	discrep	-0.24926	relig	0.24344
anx	-0.21244	ppron	0.24526	swear	-0.24111
leisure	0.19806	pronoun	0.22995	informal	-0.22797
reward	0.19703	focusfuture	0.22515	negemo	-0.21692
drives	0.18761	compare	0.20676	Tone	0.19561
		adj	0.18596	male	0.19366
				time	-0.18919
				power	0.18722
				Apostro	-0.18591

ρ: Spearman correlations values

Summary Variable	Feature Set	MyP (LIWC)	GW (S-LIWC)	GenFB (S-LIWC)
Analytical Thinking	Formula	**7.36**	**6.55**	**9.36**
Authenticity	Formula	23.68	19.89	33.83
	Source	16.97	18.89	23.6
	Correlated	**10.03**	**9.96**	**15.92**
Emotional Tone	Correlated	14.59	17.24	22.72
Clout	Source	9.66	5.62	12.3
	Correlated	**4.34**	**6.355**	**8.69**

categories while Neuroticism is mostly correlated with negative emotion features. Percentage of words consisting more than six characters in users' posts as well as work-related words give a positive effect to Conscientiousness. Finally, Openness to Experience is negatively correlated with netspeak. The observations can also be applied to the myP dataset except that Openness to Experience is positively correlated with affiliation and work categories with significance.

To determine whether the LIWC or S-LIWC features can be effectively used for personality prediction, we study the LIWC and S-LIWC features that are significantly correlated with the ground truth personality scores. We find that the two sets of features with strong correlation are almost identical. Therefore, for each personality trait, we select the word categories with p-value <0.1 in the correlation test to be features for the prediction model. Table VII shows these features which are used in the LIWC and S-LIWC based personality prediction task.

B. Personality Prediction

We finally examine if S-LIWC can help to predict personality traits of Singapore users better than LIWC. Only GW dataset is utilized in this experiment as this is the only dataset with both personality trait ground truth and Singlish content. Each user in the dataset has a ground truth score between 0 and 5 for each of the five personality traits. This experiment involves a prediction model training step and a model evaluation step. In the training step, we build a personality trait prediction model using S-LIWC word category features. As GW Dataset is small, we use leave-one-out strategy to obtain the training and test user sets.

LIWC 2015 includes a desktop application to return LIWC word category feature scores. To obtain the S-LIWC word category feature scores, we store S-LIWC word lists in a dictionary file and import the latter to the LIWC desktop application. This way, the LIWC desktop application will return word category feature scores based on S-LIWC. We

GW (S-LIWC)				
	Analytical Thinking	**Clout**	**Authenticity**	**Emotional Tone**
ρ	0.963	0.929	0.943	0.887
p-value	2.99E-33	2.248E-26	6.07E-29	7.064E-21
GenFB (S-LIWC)				
	Analytical Thinking	**Clout**	**Authenticity**	**Emotional Tone**
ρ	0.919	0.6223	0.847	0.7013
p-value	0.00E+00	4.097E-37	1.58E-74	3.32E-25

TABLE VI
PERSONALITY SCORE DISTRIBUTION OF EACH SCORES BETWEEN 0 AND 5
(EX:EXTRAVERSION, NE:NEUROTICISM, AG:AGREEABLENESS,
CO:CONSCIENTIOUSNESS, OE:OPENNESS-TO-EXPERIENCE)

Data Sources		EX	NE	AG	CO	OE
myP	Mean	3.29	2.62	3.6	3.52	4.07
	StdDev	0.86	0.78	0.67	0.74	0.57
GW	Mean	2.98	2.96	3.74	3.41	3.1
	StdDev	0.76	0.73	0.59	0.54	0.61
Singapore	Mean	3.56	**na**	2.76	3.46	**na**
Average	StdDev	0.59	**na**	0.64	0.6	**na**

then train a personality trait prediction model using LIWC features and another prediction model using S-LIWC features. The trained models are applied to the test dataset.

We compare the accuracy of our trained models with a baseline model which always returns the average personality trait score computed using the training set. We utilize Linear Regression to train our prediction model, and measure the accuracy of method using LIWC and S-LIWC by Mean Absolute Error (MAE) using Leave-One-Out cross validation. As shown in Table VIII (with the better results shown in boldface), we observe that S-LIWC outperforms the original LIWC in the prediction results for all personality traits. This shows the effectiveness of introducing additional Singlish words to S-LIWC. Moreover, both S-LIWC and LIWC-based prediction models beat the baseline.

VII. USER EVALUATION OF S-LIWC

In this user evaluation task, we aim to examine the precision of S-LIWC words extracted using our proposed algorithm. Unlike our early evaluation in Section IV, which aims to evaluate the performance of the disambiguation classifier across all word categories, we focus on a single S-LIWC word category here. The annotators are to examine all words under this specific category, and report the by-category precision score.

Three PhD students are recruited for this annotation task. Each candidate word is labeled in the criteria that whether it is semantically similar to the seed word, and at the same time carries the same sentiment polarity.

As there are more than a hundred categories in LIWC, it is impractical to exam them all. Therefore, we only select **Posemo** category for the labeling task. Posemo in LIWC contains 542 seed words and 1,125 candidate words. Some examples of words added are listed in Table IX. Two or three of the annotator agree that 778 candidates out of 1,125 are correctly selected. One or more annotators agree that 849 candidates are correctly selected. We therefore obtain a precision of 69.1% and 75.4% for the two criteria. The precision score is computed as:

$$\frac{\#Correctly\ extracted\ candidates}{\#All\ candidates\ under\ Posemo} \quad (1)$$

This result is reasonable considering that the Posemo word category has been used to predict Summary Variables and personality trait with good accuracies. A more extensive user evaluation of S-LIWC will be included for our future work.

VIII. CASE EXAMPLES

In this section, we examine some examples that could show some limitations of our constructed S-LIWC. This may suggest possible future improvements to our proposed LIWC construction approach. We also examine some case examples that show the improvement of personality trait prediction using S-LIWC. These examples clearly show the need for language-specific LIWC even if the content uses an English-based creole language.

A. Case Examples: Limitations of S-LIWC

Here, we focus on ambiguous words that have been added to word categories of S-LIWC. Such words are detected by human annotators. For example, the seed word "cool" in the "Posemo" category of LIWC is similar to "cold" and "freezing" according to word embeddings due to multiple meanings of "cool". Therefore, the latter two words have been wrongly added to "Posemo" category of S-LIWC. This error could potentially be corrected if we could disambiguate the multiple meanings of "cool" and find the right one for identifying candidate words.

Our second example is the seed word "award" in the "Posemo" category which is matched with another word "oscar" according to word embeddings. Unfortunately, the two words have a parent-child relationship instead of a synonym

TABLE VII
WORD CATEGORIES USED IN PERSONALITY PREDICTION TASK[a]

Big Five Personality Trait	LIWC Word Categories
Extraversion	2nd person (you), Affiliation, Female referents (family), Social Words (social), Semicolons (semiC), 1st person plural (we), Clout, Religion (relig), Nonfluencies (nonflu)
Agreeableness	Affect Words (affect), Netspeak, Positive emotion (posemo), Informal Speech (informal), Exclamation marks (Exclam), Articles (article), Relativity (relativ), Certainty (certain), Function Words (function), Money, Assent, Impersonal pronouns (ipron), Perpetual Processes (percept) , Prepositions (prep), Space, Apostrophes (apostro), Colon, Religion (relig), Adverb, Family, Authentic
Conscientiousness	Words>6 letters (sixltr), Work, Comma
Neuroticism	Nonfluencies (nonflu), Apostrophes (apostro), 3rd pers plural (they), Family, Body, Hearing (hear)
Openness to Experience	Netspeak, Hearing (hear), Affect Words (affect), Positive emotion (posemo), Comparatives (compare), Sexuality (sexual), Friend, Reward focus (reward), Prepositions (prep), Function Words (function),Informal Speech (informal), Space, Insight, Religion (relig), Comma, Emotional Tone, Apostrophes (apostro), Article

[a]We show the abbreviations in parentheses.

TABLE VIII
PREDICTIVE RESULT OF S-LIWC (MAE)

	EX	AG	CO	NE	OE
LIWC	0.5618	0.3308	0.3344	0.6483	0.4735
S-LIWC	**0.5602**	**0.2980**	**0.3013**	**0.5711**	**0.4220**
Baseline	0.583	0.561	0.489	0.575	0.514

one. In this case, more research would be needed to develop a way to pick up such nuance in meaning.

B. Case Examples: Personality Trait Prediction

We focus on identifying case examples that show improvement on personality trait prediction. We note that S-LIWC contributes largest improvement to the prediction of Neuroticism and Openness-to-Experience. We found a Facebook post with content *"Selfie w Ahma <3"* which LIWC fails to account for word category features "netspeak", "prep" (preposition), and "family" because "Selfie" (self-portrait photograph), "w" (with) and "Ahma" (granny in Chinese dialect) do not exist in the LIWC word dictionary. As S-LIWC covers all these three words (i.e.,"selfie", "w", and "Ahma"), we are now able to predict Neuroticism and Openness-to-experience traits more accurately using S-LIWC.

IX. CONCLUSION

The main contribution of this paper is to automate the construction of LIWC dictionaries for English-based Creole languages such as Singlish. This proposed approach significantly reduces the human efforts thereby allowing us to quickly learn more LIWC variants for different English-based Creole languages. Our proposed approach is based upon the use of word embedding model trained using a large collection of social media content, which is in turn used to find similar words for seed words in the original LIWC dictionary. We also

develop an algorithm to disambiguate the polarity sense of candidate words leveraging on the vector arithmetic property of word embeddings. Our user evaluation experiment on a large Twitter text content generated by Singapore users shows that the new words suggested by our approach are highly relevant to the existing word categories in the LIWC dictionary. 67.7% of these words are determined to be relevant by human annotators.

We also develop prediction models for deriving the scores for LIWC Summary Variables (Analytical Thinking, Clout, Authentic, Emotional Tone) for Singlish. These models are shown to perform accurately using both LIWC and S-LIWC word categories. The same strategy can be adopted for developing prediction models for summary variables for other English-based creole languages.

Finally, we evaluate the usefulness of S-LIWC by conducting a personality prediction experiment. Our experiment shows that we can use S-LIWC to predict personality more accurately than using LIWC for most personality traits for Singapore users.

Looking ahead, more research can be performed to refine and evaluate our proposed approach further. We recognise that new netspeak words have been identified to be relevant to LIWC word categories. This suggests that LIWC should be revised from time to time, perhaps using our proposed approach.

For example, one could automatically extract popular netspeak words and their meanings from social media or forums and add them to the LIWC. Finally, we only focus on personality prediction and user evaluation on Posemo category in this work. More investigation can be done to examine the usefulness of S-LIWC in other applications, such as depression detection, deception prediction, and to conduct scalable user evaluation on other word categories using crowdsourcing.

TABLE IX
FEATURE ADDED FOR POSEMO AFTER HUMAN EXAMINATION

LIWC	Distinct Words Added	LIWC Seed	Candiate	Source Language	Meaning
Posemo	1,125	**relax**	relek	English	Relax in Malay style spelling
		relax	nua	Chinese dialect	Relax
		tolerance	manusiawi	Malay	Humane

REFERENCES

[1] J. Golbeck, "Predicting Personality with Social Media," *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems CHI EA 11*, pp. 253–262, 2011.

[2] J. W. Pennebaker, C. K. Chung, J. Frazee, G. M. Lavergne, and D. I. Beaver, "When small words foretell academic success: The case of college admissions essays," *PloS one*, vol. 9, no. 12, p. e115844, 2014.

[3] E. Kacewicz, J. W. Pennebaker, M. Davis, M. Jeon, and A. C. Graesser, "Pronoun use reflects standings in social hierarchies," *Journal of Language and Social Psychology*, vol. 33, no. 2, pp. 125–143, 2014.

[4] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[5] C.-L. Huang, C. K. Chung, N. Hui, Y.-C. Lin, Y.-T. Seih, B. C. Lam, W.-C. Chen, M. H. Bond, and J. W. Pennebaker, "The development of the chinese linguistic inquiry and word count dictionary." *Chinese Journal of Psychology*, 2012.

[6] M. Wolf, A. B. Horn, M. R. Mehl, S. Haug, J. W. Pennebaker, and H. Kordy, "Computergestützte quantitative textanalyse: Äquivalenz und robustheit der deutschen version des linguistic inquiry and word count," *Diagnostica*, vol. 54, no. 2, pp. 85–98, 2008.

[7] H. Zijlstra, T. Van Meerveld, H. Van Middendorp, J. W. Pennebaker, and R. Geenen, "De nederlandse versie van de 'linguistic inquiry and word count'(liwc)," *Gedrag Gezond*, vol. 32, pp. 271–281, 2004.

[8] J. T. Platt, "The Singapore English Speech Continuum and Its Basilect 'Singlish' as a 'Creoloid'," *Anthropological linguistics*, pp. 363–374, 1975.

[9] CIA, *The World Factbook - Singapore*, 2018. [Online]. Available: https://www.cia.gov/library/publications/the-world-factbook/geos/sn.html

[10] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: LIWC 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, 2001.

[11] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales." *Journal of personality and social psychology*, vol. 54, no. 6, p. 1063, 1988.

[12] A. Piolat, R. J. Booth, C. K. Chung, M. Davids, and J. W. Pennebaker, "La version française du dictionnaire pour le liwc: modalités de construction et exemples d'utilisation," *Psychologie française*, vol. 56, no. 3, pp. 145–159, 2011.

[13] F. Alparone, S. Caso, A. Agosti, and A. Rellini, "The italian liwc2001 dictionary," *LIWC. net, Austin*, 2004.

[14] P. P. Balage Filho, T. A. S. Pardo, and S. M. Aluísio, "An Evaluation of the Brazilian Portuguese LIWC Dictionary for Sentiment Analysis," in *Brazilian Symposium in Information and Human Language Technology*, 2013.

[15] L. van Wissen and P. Boot, "An Electronic Translation of the LIWC Dictionary into Dutch," 2017.

[16] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from twitter," in *PASSAT/SocialCom*. IEEE, 2011, pp. 149–156.

[17] H. Wei, F. Zhang, N. J. Yuan, C. Cao, H. Fu, X. Xie, Y. Rui, and W.-Y. Ma, "Beyond the Words," *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining - WSDM '17*, pp. 305–314, 2017. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3018661.3018717

[18] D. Sewwandi, K. Perera, S. Sandaruwan, O. Lakchani, A. Nugaliyadde, and S. Thelijjagoda, "Linguistic features based personality recognition using social media data," in *National Conference on Technology and Management (NCTM)*. IEEE, 2017, pp. 63–68.

[19] W. Gong, E.-P. Lim, F. Zhu, and P. H. Cher, "On unravelling opinions of issue specific-silent users in social media," *ICWSM*, pp. 141–150, 2016.

[20] L. R. Goldberg *et al.*, "A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models," *Personality psychology in Europe*, vol. 7, no. 1, pp. 7–28, 1999.

[21] F. Celli, F. Pianesi, D. Stillwell, and M. Kosinski, "Workshop on computational personality recognition (shared task)," in *Proceedings of the Workshop on Computational Personality Recognition*, 2013.

[22] Y. Goldberg and O. Levy, "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method," *CoRR*.

[23] T. Mikolov, W.-T. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations." in *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, vol. 13, 2013, pp. 746–751.

[24] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: Predicting deception from linguistic styles," *Personality and social psychology bulletin*, vol. 29, no. 5, pp. 665–675, 2003.

[25] M. A. Cohn, M. R. Mehl, and J. W. Pennebaker, "Linguistic markers of psychological change surrounding september 11, 2001," *Psychological science*, vol. 15, no. 10, pp. 687–693, 2004.

[26] A. G. Thalmayer and G. Saucier, "The questionnaire big six in 26 nations: Developing cross-culturally applicable big six, big five and big two inventories," *European Journal of Personality*, vol. 28, no. 5, pp. 482–496, 2014.