Stylistic Approaches to Predicting Reddit Popularity in Diglossia

Huikai Chua
University of Cambridge
huikai.chua@gmail.com

Abstract

Past work investigating what makes a Reddit post popular has indicated that style is a far better predictor than content, where posts conforming to a subreddit's community style are better received. However, what about diglossia, when there are two community styles? In Singapore, the basilect ('Singlish') co-exists with an acrolect (standard English), each with contrasting advantages of community identity and prestige respectively. In this paper, I apply stylistic approaches to predicting Reddit post scores in diglossia. Using data from the Singaporean and British subreddits, I show that while the acrolect's prestige attracts more upvotes, the most popular posts also draw on Singlish vocabulary to appeal to the community identity.

1 Introduction

Reddit is a popular social media platform which is organized into different sub-forums, called subreddits. Users can submit original content as top-level posts to each subreddit, which other users can then comment on and either up- or down-vote. The most popular posts earn tens of thousands of upvotes.

But what exactly makes a post popular? In this paper, I apply natural language processing (NLP) techniques to predicting the popularity of a Reddit post. As past research has found style to be a strong predictor of community response (Tran and Ostendorf, 2016), I focus on stylistic approaches using punctuation, stopwords and part-of-speech tags, as inspired by Bergsma et al. (2012).

In particular, I investigate how community style endorsement (Tran and Ostendorf, 2016) applies in diglossic Singapore. Linguists have observed that Singapore English is organized along a sociolect continuum from an informal basilect (Singlish), to a formal acrolect, which has minimal features of Singlish and is essentially Standard British English (Gupta, 1991; Zhiming and Huaqing, 2006). Use of the acrolect is generally associated with better education, and therefore higher socioeconomic status. On the other hand, despite top-down efforts from the Singaporean government, the basilect is the dialect used by the average Singaporean in everyday situations, and is closely associated with the Singaporean identity. In fact, Singaporean politicians intentionally include Singlish phrases in election speeches in efforts to appear more down-to-earth and likeable. With competing appeals of identity and prestige between the two, I find that the most popular posts similarly use basilectal lexicon together with the acrolect to achieve the 'best of both worlds'.

2 Related Work

Much research has gone into investigating what makes a social media post popular, including some specifically focused on Reddit. Lakkaraju et al. (2013) controlled for the content of the post by concentrating on image submissions, which are frequently re- or cross-posted to different communities by different authors. They found that the title of a submission played a role in determining its success, where titles specifically engineered towards the community it was posted in (for example, by using community-specific words) performed better.

Tran and Ostendorf (2016) took this a step further and trained separate models for the content (using Latent Dirichlet Allocation (LDA)) and the style of the language used (by replacing topic words with their part-of-speech tags). They computed the Spearman rank correlation between scores and post representations, and found that the style model was much better at predicting of the success of a post than the content model. In other words, they found that these subreddits had their own community style, and posts which are stylisti-

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 93–100

cally more similar to it are more likely to be well-received.

Fang et al. (2016) is the paper which is closest to the aim of this paper. They divided posts into eight different bins which are automatically determined by the score distribution of that particular subreddit, and evaluated model performance using a modified macro F1 score (details in Section 5.1). However, while Fang et al. (2016) focused on modelling the conversational context of a post, I instead focus on modelling the community style.

I take cues from Bergsma et al. (2012) to achieve this. They grouped their features into three broad categories: word (bag-of-words), style, and syntax features. For style features, they defined style words to be punctuation, stop-words, or Latin abbreviations, and replaced all non-style words with their part-of-speech (POS) tags. Meta-features such as average word and sentence lengths were also used. For grammatical features, they included a feature for every unique context free grammar and tree substitution grammar rule, as well as Charniak and Johnson re-ranking features (Charniak and Johnson, 2005). These are parse tree features initially used for re-ranking parser output, and include aggregate features for conjunct parallelism and lexicalized features for sub-trees and head-to-head dependencies.

3 Approach

I adopt Bergsma et al. (2012)'s three-pronged approach to stylometry. For content features, I use Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019); for style features, I used term frequency–inverse document frequency (TF–IDF) vectors (Sparck Jones, 1988) with stopwords and punctuation only; for grammatical features, I used dependency relations and part-of-speech (POS) tags.

3.1 BERT

BERT (Devlin et al., 2019) uses a Transformer (Vaswani et al., 2017) encoder to achieve state-of-the-art performance on a wide variety of tasks. Past investigations have suggested that BERT is not just good at capturing the meaning of sequences, but is also sensitive to the grammar of phrases. Goldberg (2019) ran a series of grammatical test cases and found that BERT performed well on all; Jawahar et al. (2019) suggested that BERT layers encode linguistic information hierarchically, with surface information in lower layers, syntax in the middle, and semantic information at the top. Thus, it seems that BERT would be able to capture both the contents of posts as well as their style, making it particularly suitable for this task.

To capture the content of a post, I used the uncased English BERT$_{BASE}$ model provided by Hugging Face (Devlin et al., 2019; Wolf et al., 2020) to produce post embeddings. Since BERT is designed to encode sequence-level representations (Devlin et al., 2019), I first split each Reddit post into sentences using NLTK's sentence tokenizer. Then, each of the sentences were tokenized and encoded with BERT. Finally, the embeddings for each sentence were averaged to produce the overall post-level representation.

3.2 Grammatical features

I used the spaCy parser to extract dependency relations and part-of-speech (POS) tags. First, I hand-compiled the lists of relations and POS tags from the documentation[1]. Then, the dependency and POS labels for each word were replaced by their positions in the respective lists. I also included the POS labels of the heads of each word. Each of the three vectors (dependency tag, POS label, and head POS label) were L2-normalized.

3.3 Style features

I used stopword TF–IDF vectors for the style features. The vocabulary is predefined to be either a stop-word, using NLTK's English stop-word list, or a punctuation character, from Python's inbuilt string module. NLTK's English stop-word list, consists of 179 stop-words including determiners ('the', 'a'), pronouns ('he', 'she'), prepositions ('before', 'after'), quantifiers ('all', 'some'), among others.

3.4 Model

As I wanted to focus on feature rather than the model engineering, I used a tried-and-tested model for imbalanced class distributions: random forest classifiers. I opted to use the RandomForestClassifier from sklearn. I weighted each class proportional to its frequency in the dataset particular. For each Level $i, 0 \leq i \leq 7$, these are:

$$weight_{Level i} = \frac{\#samples_{Level 0}}{\#samples_{Level i}}$$

[1] https://spacy.io/api/annotation

4 Data

4.1 Data collection

As my aim was to investigate the stylistic characteristics of communities, I selected a subreddit with a distinctive linguistic style – the Singaporean (SG) subreddit.[2] Singaporeans speak a distinctive flavour of English dubbed "Singlish", which has drawn much linguistic interest as the *lingua franca* of different cultural communities. It serves as the vernacular in the diglossic Singapore, where the Standard British English serves as the acrolect.

Therefore, for comparison, I also select the United Kingdom (UK) subreddit.[3] Although the population sizes of the two countries are quite different (roughly 5 million Singaporeans versus over 60 million UK citizens), I found that the subreddit sizes were similar, with roughly 300k participants in SG and 400k participants in UK.

Data was scraped from the two subreddits by querying the Pushshift API.[4] 3 years' worth of posts, ranging from 1 January 2017 to 31 December 2019, were collected for each subreddit. To ensure each post had sufficient linguistic content, I excluded any posts containing less than 101 characters.

4.2 Annotations

I followed the annotation procedure described in Fang et al. (2016). First, all posts with a score below 2 were labelled as the lowest class, Level 0. This threshold was selected for the base class as all new posts are initialized with a score of 1 (Fang et al., 2016). For the next level, the median of the remaining posts was computed and all posts with a score lower than the median labelled as 1. This process is repeated for each of the levels 2-6. Finally, the remaining posts are labelled as the highest class, Level 7. For clarity, the annotation function is given as pseudocode in the appendix (Algorithm 1). The distribution for each subreddit along with the respective class thresholds are summarized in Table 1.

Level	r/singapore		r/UK	
	Size	**Cap**	**Size**	**Cap**
0	15,633	2	9246	2
1	4797	14	2466	14
2	2394	36	1246	64
3	1200	74	633	191
4	620	151	318	531
5	310	284	159	1086
6	156	507	79	1762
7	156	-	80	-
Total	25,266		14,227	

Table 1: Distribution of classes for both subreddits.

5 Quantitative evaluation

5.1 Evaluation metric

I also replicate the evaluation procedure described in Fang et al. (2016). First, the F1 score for each of the Levels 1-7 were computed, treating each sample with a score below that level as a negative example. Then, the final score for that model is obtained by averaging over the F1 scores for each level. Fang et al. (2016) had designed this evaluation metric such that the higher levels, which are of greater interest, are weighted more highly. For example, for the SG score distribution, a model which predicts only Level 1s would obtain an F1 of 0.0789, while a model which predicts only Level 8s would obtain an F1 of 0.176. Level 0 is excluded in computing the average, as using the scheme described above the F1 score would always be 1.

5.2 Results

In total, I tried six different combinations of the three different types of features. First, I tried each of the style features, BERT embeddings, and grammatical (POS and dependency labels) features separately. Then, I tried individually adding the other two types to the weakest baseline, which was the grammatical model. Finally, I tried a combination of all features together. I used stratified five-fold cross-validation and report the average modified F1 score across all folds. The results can be found in Table 2.

In all cases, the models clearly out-performed the simplistic baseline of 0.176 for a model which predicts only the highest class. Although the scores for each model are similar, the results are consistent across the two sub-reddits, r/Singapore (SG) and r/UnitedKingdom (UK). In both cases, BERT performs the best out of the three baselines, and indeed was improved only slightly by 0.02 for SG when other features were added, and not at all for UK.

Between SG and UK, all models performed sig-

[2] reddit.com/r/singapore
[3] reddit.com/r/unitedkingdom
[4] https://github.com/pushshift/api

nificantly better on the UK dataset. This is possibly due to there being a more consistent group style for UK, compared to the diglossic situation in Singapore. It could also be due to the tools used (such as BERT and spaCy) being trained mostly on standard American / British English, and hence performing better on the UK subreddit.

The results are not directly comparable to those achieved by Fang et al. (2016), due to differences in the data used. However, comparing the trends in F1 score across levels reveals some interesting differences. In Fang et al. (2016), the model performed better on lower levels, with an average of nearly 0.60 F1 on the lowest 3 levels, and an average of under 0.50 F1 on the highest 3.

However, in this paper, the models used performed better at higher levels, as can be seen from Figure 1. Though the models start with roughly similar performance for Levels 1 and 2, they gradually diverge as the level increases, for a gap of 0.085 F1 points at the highest. As we will see in the next section, a diglossic situation with two competing dialects makes it a bit more difficult to craft an effective style.

	SG	UK
Style	0.748	0.788
BERT	0.749	0.793
Gram.	0.733	0.781
Gram. + style	0.750	0.792
Gram. + BERT	0.751	0.793
All	0.751	0.793

Table 2: F1 scores for each subreddit for each model.

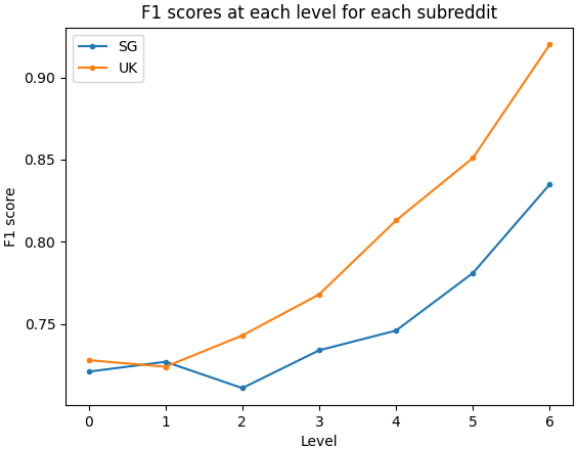

Figure 1: F1 scores for each individual level for the model with all features. The numerical results are given in the appendix (Table 6.)

6 Qualitative evaluation and analysis

In this section, I look at inferences that can be gained by looking at the most important features from each model. While BERT appears to be good at capturing the grammatical relations, it is not as good with more complex relationships. I also find overlaps between the grammatical and stylistic features, with the more specific stylistic features performing better. Finally, I investigate grammatical and lexical similarities between SG posts and the acrolect and basilect respectively. I find that while the most popular posts are most grammatically similar to the acrolect, they also use the most lexicon from the basilect.

6.1 Feature importances

The top 10 non-BERT features, i.e. either a POS, dependency, head_POS tag, or a stopword or punctuation for SG and UK are tabulated in Table 3 and 4 respectively. The POS tags of head words (henceforth referred to as head-POS) are differentiated from the POS tags of words with a 'head_' prefix.

6.1.1 Does BERT capture grammatical features?

Although model performance improved only a little when BERT embeddings were added to grammatical features, the most informative features were completely taken over by BERT features. For SG, the highest ranking non-BERT feature was 'head_ADV' at 15th place, with the next one, 'head_SCONJ', coming in 10 places lower. For UK, the top two were at 5th ('head_X') and 30th ('CCONJ') place respectively. This does suggest BERT is capable of capturing the grammar of a sentence in its embedding, as it seems to have replaced grammatical features when it was added to the model.

Of particular note are the changes in the individual features' rankings. In the grammatical features only model, the top features are occupied by dependency and POS tags; the highest ranking head-POS features for SG and UK are 'head_NOUN' and 'head_VERB' at 12th and 7th place respectively. The relatively higher rankings of head-POS tags after adding BERT suggest that it might not be as good at capturing more complex grammatical relationships.

Rank	Gram. only	Style only	Gram + Style	Gram + BERT (position)
1	punct	.	.	head_ADV (15)
2	PUNCT	the	?	empty dep relation (22)
3	ROOT	?	PUNCT	head_SCONJ (25)
4	DET	to	punct	prt (42)
5	advmod	,	ROOT	head_PUNCT (48)
6	poss	and	the	dative (50)
7	aux	i	advmod	DET (52)
8	NOUN	a	head_VERB	poss (104)
9	det	of	AUX	PUNCT (118)
10	ADJ	in	DET	PART (128)

Table 3: Top 10 non-BERT features for selected models on the SG dataset.

Rank	Gram. only	Style only	Gram + Style	Gram + BERT (position)
1	amod	.	/	head_X (5)
2	ROOT	/	.	CCONJ (30)
3	NOUN	the	ROOT	PART (46)
4	DET	to	head_VERB	amod (48)
5	PUNCT	a	punct	cc (57)
6	punct	i	i	conj (137)
7	head_VERB	and	AUX	advmod (173)
8	PRON	,	head_NOUN	relcl (174)
9	aux	”	PUNCT	SPACE (176)
10	cc	?	amod	NOUN (220)

Table 4: Top 10 non-BERT features for selected models on the UK dataset.

6.1.2 Overlap between grammatical and style features

There is a noticeable overlap between grammatical and style features, where the top-ranked features for grammatical and style mirror each other. For example, punctuation ranks among the most informative style features, particularly for UK where they occupy 5 out of the top 10 spots despite making up only 15% of the roughly 200 style features. Among the 100 grammatical features, the dependency rule ‘punct’ and POS tag ‘PUNCT’ also rank highly. A similar trend can be seen for determiners, which rank highly as both style features (in the form of the stopwords ‘the’ and ‘a’) as well as grammatical features (in the form of the dependency rule ‘DET’). This possibly contributes to the very similar performances of the style and grammatical models.

However, it appears that the more specific style features generally perform better. When grammatical and style features were combined for SG, the specific punctuation characters ‘.’ and ‘?’ appear before ‘PUNCT’ and ‘punct’. Similarly, the determiner ‘the’ appears before the dependency rule ‘DET’. This might explain the difference between the individual style and grammatical models, where the style model performed better on both the SG and UK datasets. Although the top features from both form a common subset, the more specific features found in the style model are better predictors.

6.2 Grammatical closeness

Since the acrolect should be close to Standard British English, I decided to assess this by first computing the Euclidean centre of Level 7 posts from UK. Then, for each of the Levels 0-7, I computed the average Euclidean distances from Singaporean posts to the UK centre. For comparison, I also compute the average distances for UK posts. The distances for each of the three types of features are tabulated in Table 5. Note that, due to different dimensions and normalization, the distances for each feature are not directly comparable to that of other features.

Across all three features, Level 0 SG posts are generally less similar to the UK centre than Level 0 UK posts, possibly due to greater presence of the

basilect. However, at the top level, SG posts are even *more* similar than the original posts the centre was calculated from. This suggests that indeed the Standard British English acrolect holds more prestige and draws greater community endorsement.

Separately, the consistent trend in the Style column where posts from higher levels are more similar to the Level 7 centre than lower level posts supports the hypothesis that there is a community style and posts which are more similar to it receive greater community endorsement.

6.3 Lexical closeness

We can see that stylistically and grammatically, the most popular posts from SG are very similar to British English. However, what about lexically? Singlish has a vocabulary full of borrowed words and phrases from the different cultural groups of Singapore. As mentioned earlier, politicians often try to build rapport by sprinkling speeches with Singlish terms. Would we observe something similar on Reddit? I decide to investigate the prevalence of Singlish terms by level.

Compiling a written Singlish lexicon can be very tricky due to several reasons, including different possible romanizations and lexical change in loanwords (when the word's meaning changes). With this in mind, using my experience growing up in Singapore, I compiled a list of 56 everyday Singlish words and phrases, including alternative spellings where practical. I excluded phrases with specific niches, like the names of foods or military terms (common in Singapore where all males have to enlist for 2 years). The full list of phrases used is included in the appendix.

The average number of such Singlish words or phrases used per 1000 words per post for each of the Levels 0-7 is shown in Figure 2. The results confirm the earlier hypothesis that effective use of Singlish words helps earn more community endorsement. We see a somewhat U-shape in the frequency of Singlish terms; the least popular posts include more Singlish than the middlingly popular posts, likely due to greater influence of the basilect, while posts on the highest levels utilise Singlish vocabulary in tandem with the acrolect to achieve the most popularity.

A reading of the Level 7 texts including Singlish terms confirm that this is indeed the case. For example, one post is written in very eloquent standard English[5], but includes Singlish quotes as well as specific, appropriate Singlish terms (with English explanations in brackets).

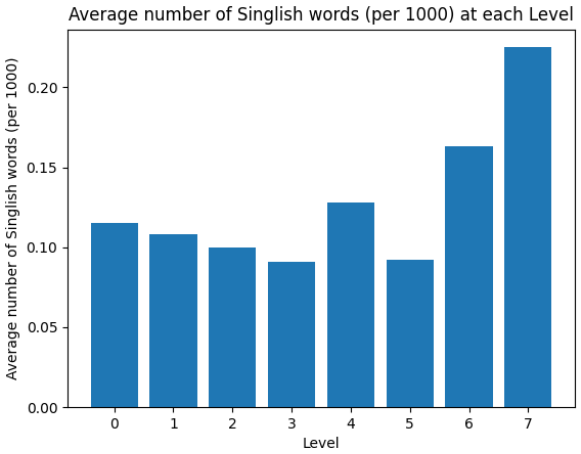

Figure 2: Average number of everyday Singlish terms per 1000 words. The numerical results are given in the appendix (Table 7).

7 Future Work

In the future, I would like to extend this work by including context-free grammar (CFG) features using CoreNLP to compare the use of Singlish grammatical features, in order to further confirm or disprove the theory that the most popular posts use the acrolect, i.e. the least grammatical features from Singlish, despite having the highest prevalence of Singlish terms.

8 Conclusion

In summary, in this paper, I look at the linguistic factors that predict the community response of Reddit posts. I collected data from two Reddit subforums, the Singaporean and UK subreddits. Following Bergsma et al. (2012), I extracted three types of features, broadly grouped as grammatical, stylistic and content features. The models generally show good results, with the stylistic and grammatical models performing comparable to state-of-the-art BERT embeddings.

I investigate also the hypothesis that posts conforming to a group's style receive greater community endorsement (Tran and Ostendorf, 2016). I show that in a diglossic situation, although the acrolect draws greater prestige, the most successful posts draw on features from the basilect in order to connect with the audience.

[5]https://www.reddit.com/r/singapore/comments/8gfewd/the_singaporean_male_version_of_metoo_an_exguards/

Level	BERT		Style		Gram.	
	SG	UK	SG	UK	SG	UK
0	4.45	4.35	0.875	0.847	0.704	0.715
1	4.40	4.31	0.874	0.845	0.688	0.697
2	4.33	4.34	0.877	0.832	0.678	0.716
3	4.39	4.34	0.848	0.830	0.676	0.733
4	4.17	4.35	0.826	0.826	0.632	0.722
5	4.11	4.20	0.808	0.829	0.629	0.711
6	3.79	4.18	0.797	0.814	0.591	0.722
7	3.55	3.91	0.751	0.800	0.513	0.659

Table 5: Average Euclidean distances from the UK Level 7 centre.

Acknowledgments

Dr Andreas Vlachos and Prof Ted Briscoe taught the L101 course for which this project was done. Prof Briscoe also gave feedback on this paper, and suggested submission of this work to the Student Research Workshop (SRW). I would also like to thank my bachelor's and master's thesis supervisors, Dr Andrew Caines and Dr Helen Yannakoudakis, for their guidance on NLP techniques. Dr Caines also reviewed an earlier version of this paper. Finally, I am grateful to the SRW anonymous reviewers for their detailed feedback.

References

Shane Bergsma, Matt Post, and David Yarowsky. 2012. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 327–337, Montréal, Canada. Association for Computational Linguistics.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 173–180, Ann Arbor, Michigan. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hao Fang, Hao Cheng, and Mari Ostendorf. 2016. Learning latent local conversation modes for predicting comment endorsement in online discussions. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 55–64, Austin, TX, USA. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing BERT's Syntactic Abilities. *arXiv e-prints*, page arXiv:1901.05287.

Anthea F Gupta. 1991. Acquisition of diglossia in singapore english. *Child language development in Singapore and Malaysia*, pages 119–160.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. 2013. What's in a name? understanding the interplay between titles, content, and communities in social media.

Karen Sparck Jones. 1988. *A Statistical Interpretation of Term Specificity and Its Application in Retrieval*, page 132–142. Taylor Graham Publishing, GBR.

Trang Tran and Mari Ostendorf. 2016. Characterizing the language of online communities and its relation to community reception. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1030–1035, Austin, Texas. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Bao Zhiming and Hong Huaqing. 2006. Diglossia and register variation in singapore english. *World Englishes*, 25(1):105–114.

A Appendices

A.1 Annotation algorithm

Algorithm 1 Annotator

Input: An array 'score' containing the number of votes for each post
Output: Another array 'classes' containing the annotations

get_indexes(array, condition) ← function which returns the list of indexes x in the array for which array[x] satisfies condition

class_indexes ← new array[8]
class_indexes[0] ← get_indexes(score, x $\leq$ 1)
rest_of_posts ← get_indexes(score, x $>$ 1)
for i, 1$\leq$ i $\leq$ 7 **do**
median ← median score of rest_of_posts
class_indexes[i] ← get_indexes(score, x $<$ median)
rest_of_posts ← get_indexes(score, x $\geq$ median)
end for

class_indexes[7] = rest_of_posts
classes ← new array[len(score)]
for i, 0$\leq$ i $\leq$ 7 **do**
classes[class_indexes[i]] = i
end for

A.2 Extra numerical results

	SG	UK
Level 1	0.721	0.728
Level 2	0.727	0.724
Level 3	0.711	0.743
Level 4	0.734	0.768
Level 5	0.746	0.813
Level 6	0.781	0.851
Level 7	0.835	0.920

Table 6: F1 scores for each individual level for the model with all features.

Level	# terms (per 1000)
0	0.115
1	0.108
2	0.0996
3	0.0906
4	0.128
5	0.0923
6	0.163
7	0.225

Table 7: Average number of everyday Singlish terms per 1000 words.

A.3 List of Singlish words

'abuden', 'act blur', 'agak', 'ai', 'aiya', 'alamak', 'ang mo', 'ang moh', 'atas', 'bao toh', 'barang', 'bo', 'bodoh', 'bojio', 'boliao', 'botak', 'chao', 'chee bai', 'chim', 'cheem', 'chio bu', 'chiong', 'chope', 'gahmen', 'heng', 'huat', 'jialat', 'jio', 'kena', 'kiasu', 'la', 'lah', 'lao', 'leh', 'lepak', 'liao', 'liddat', 'mafan', 'mah', 'meh', 'paiseh', 'ps', 'paktor', 'sabo', 'sia', 'sian', 'siao', 'simi', 'tahan', 'ulu', 'wa', 'walao', 'wayang', 'ya', 'yah'