



**BT4222**  
**Group 15**  
**Word2Vec Analysis and Comment Generation using**  
**r/Singapore Data**  
**(Singaporean English)**  
**Project Report**

Daniel Lee Wei Sheng | A0167810E  
James Chua | A0167073A

<b>1 Project Objective</b>	<b>3</b>
<b>2 Business Problem</b>	<b>3</b>
<b>3 Word2Vec Analysis</b>	<b>3</b>
3.1 Datasets Used	3
3.2 Data Preprocessing	4
3.3 Word2Vec Training	4
3.4 Word Similarity	4
3.5 Historical Semantics	4
3.6 New Words Added into Singaporean English Vocabulary	6
3.7 How Singaporean's Opinions on Issues have Changed	6
3.8 Analysing Gender Bias amongst Singaporeans	7
3.9 Insensitivity to other Countries or Groups of People	7
3.10 Differences in Results Between a Model with a Context Window Size of 2 vs 10	8
3.11 Differences in Results Between a Model with Dimension Size of 300 vs 60	9
<b>4 GPT-2</b>	<b>9</b>
4.1 Data preprocessing	10
4.2 Inference parameters	10
4.3 General article generation	10
4.4 Singlish text generation	11
4.4.1 Walao eh	11
4.4.2 Knn ccb	11
4.5 Singaporean context text generation	12
4.5.1 SINGAPORE	12
4.5.2 PAP	13
4.5.2 Singaporean meme generation	13
<b>5 Conclusion</b>	<b>14</b>
<b>6 Bibliography</b>	<b>15</b>

# 1 Project Objective

While there has been an abundance of semantic analysis based on embeddings produced by Word2Vec, there has been a lack of more thorough analysis, especially for less common languages. We observe that the English spoken in Singapore as one that exists on a linguistic spectrum. On one hand, the colloquial form is highly derived from Southern Chinese languages, as well as Malay. We define this creole language as Singlish. On the other hand, within more formal contexts, and among those with a higher socio-economic status, one tends to hear Singaporean Standard English, which is closely mutually intelligible with other forms of English<sup>1</sup>.

In this paper, we aim to analyse the language written by Singaporeans online, which includes both Singlish and Singaporean Standard English. In order to obtain a sufficient corpus size, we utilised the Pushshift API which allows the retrieval of comments on the subreddit r/Singapore. This amounted to 3.5 million comments from 2012 to 2019 November.

We performed two different broad analysis of Singaporean English:

1. Historical and bias semantic analysis based on Word2Vec (Daniel)
2. Singaporean text generation with GPT-2 (James)

## 2 Business Problem

Deep language models that are capable of generalizing to Singaporean forms of English is an important step in providing services to customers. For example, in chat bots, a language model is trained for intent classification. Thus, a chat bot used in Singapore has to be able to understand the English written in Singapore, which could differ with English from other countries.

Other uses include sentiment analysis, where a language model fine tuned on Singaporean English would perform better on Singaporean comments as compared to one fine tuned on general internet text. One could use sentiment analysis to gauge a product / service reception by the Singaporean public, and make adjustments if necessary.

## 3 Word2Vec Analysis

### 3.1 Datasets Used

In order to study how Word Vectors have changed over time, we decided to split up the large combined data set into a 2012-2016 dataset and 2019 dataset. This would give us 1.2

---

<sup>1</sup> Modelling Variation in Singaporean English, Leimgruber 2009  
<https://jakobleimgruber.ch/dbFile/48/u-c661/u-c661/dissertation.pdf>

million comments from 2012-2016 and 0.8 million comments from 2019, which would be sufficient to train separate Word2Vec models on. In addition, we still utilised the combined data set from 2012-2019 for other forms of analysis such as assessing gender bias or insensitivity to people of other backgrounds amongst Singaporeans.

## 3.2 Data Preprocessing

To normalise the text data we removed all non English alphabets and lower-cased the words. We then utilised Regular Expressions to normalise contractions to their longer forms, such as “won’t” to “will not” and “I’ll” to “I will”. Due to there being a Singaporean internet slang we had to standardise certain spellings of terms commonly used by Singaporeans. For example, “ft” was converted to “foreign talent” and “gahmen” was converted to “government”. After removing stopwords from the data set we proceeded to lemmatise the text data. In order to lemmatise the words accurately we had to extract the Part of Speech tag for each word. For example the word “running” would only lemmatise to “run” if the verb Part of Speech tag is extracted and inputted into the lemmatization function. If not, the default procedure would be that the lemmatization function will just assume every word is a noun and “running” will not be lemmatised at all. Finally, we formed bigrams to add into the dictionary along with unigrams. We extracted bigrams that had a minimum count of 30 within the corpus.

## 3.3 Word2Vec Training

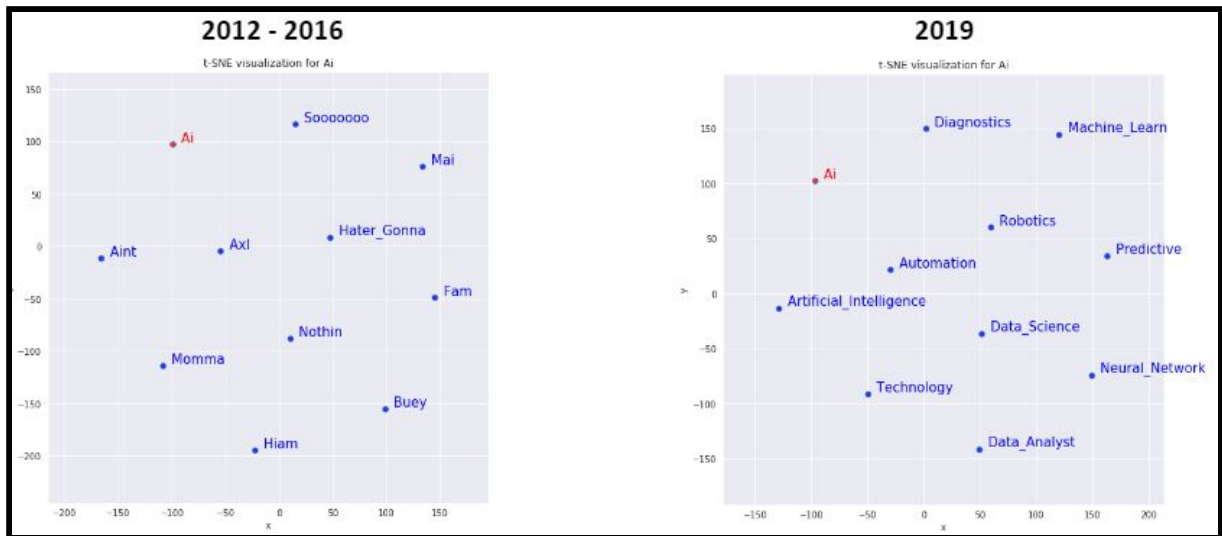
We decided to utilise Word2Vec’s Continuous Bag of Words Model to understand the semantic relations of words used in Reddit Singapore based on the context they are used in. For each dataset we used, we trained a separate model. We also adjusted the parameters, by trying a context window size of 2 and 10. We used a dimension size of 300 and 60 as well for the word vectors.

## 3.4 Word Similarity

Utilising the word vectors created we are able to measure the similarity between words based on the Cosine Similarity, with a high Cosine Similarity Score indicating a closer semantic relation. We are able to display this by visualising the t-SNE plots of the specified word and its top 10 most similar words.

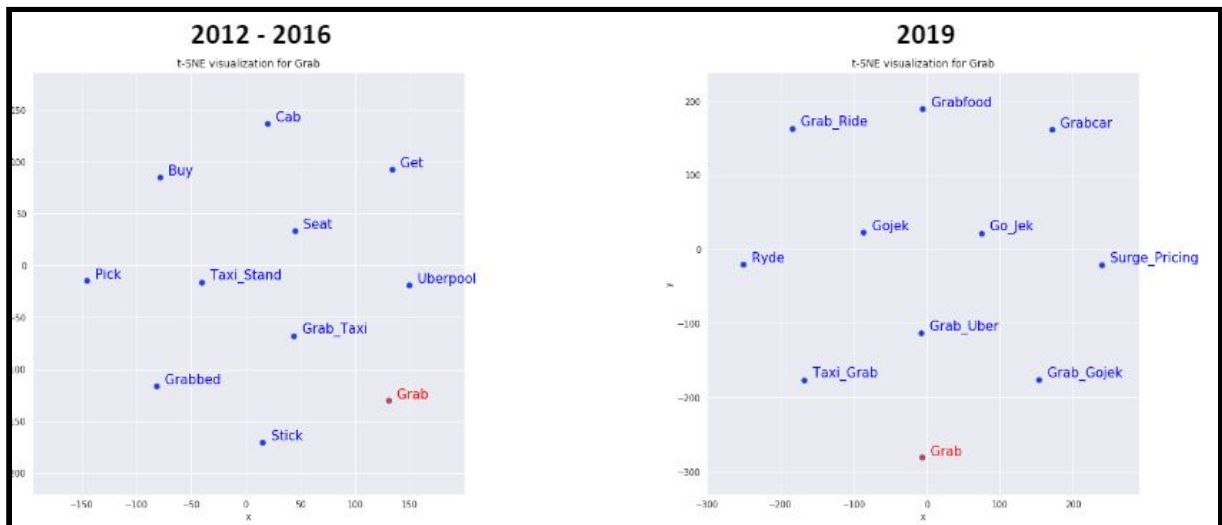
## 3.5 Historical Semantics

We compared the Word Vectors from the 2012-2016 and 2019 dataset. Looking at the most similar words derived, we can analyse how the usage of words commonly used by Singaporeans’ have changed over time. The similar words indicate what the specified word meant to Singaporeans or shows what it is related to.



**Figure 3.1: t-SNE Visualisation of the word “AI” and its Top 10 Most Similar Words**

With the rise of Data Science and Technology in the past few years, we decided to look at how the usage of the word “AI” changed. In the past it was frequently used to form Singlish / Hokkien phrases such as “Ai Mai”, “Ai Buey”, “Ai Hiam”. However, there has been a stark shift in 2019 with the word “AI” mostly being used to refer to Machine Learning and Technology.

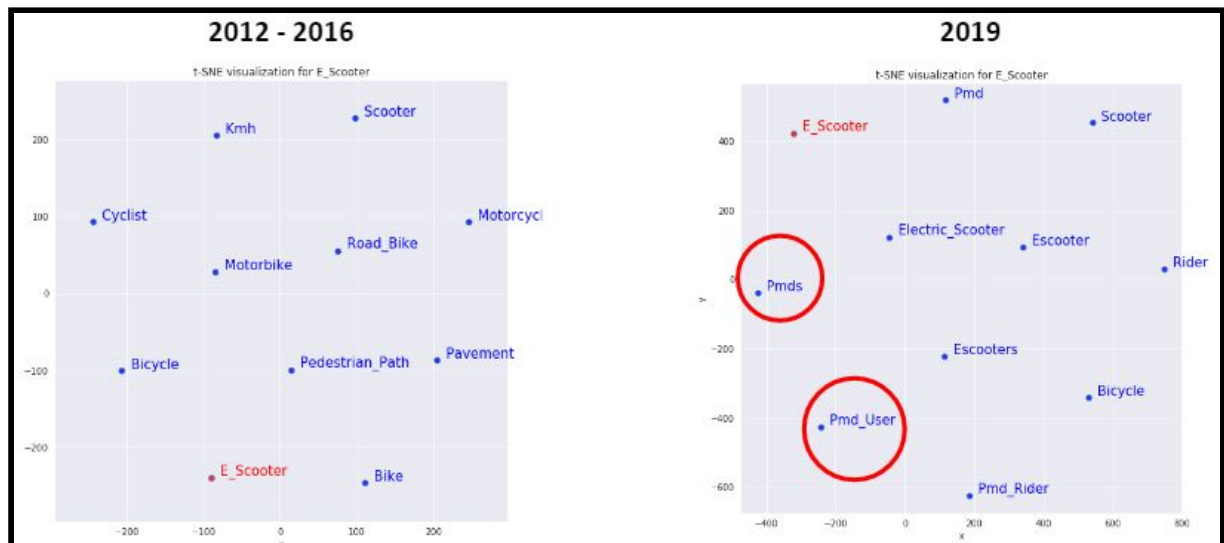


**Figure 3.2: t-SNE Visualisation of the word “Grab” and its Top 10 Most Similar Words**

With Grab expanding its business from just a ride hailing Mobile app to food delivery, and with Grab being a monopoly for a time with Uber leaving Singapore, we can see that the word “Grab” has changed from being used to refer to the actual verb of taking something, to just being used to refer to the Grab company, its businesses, or its competitors such as Ryde and Gojek

### 3.6 New Words Added into Singaporean English Vocabulary

Even in a short span of a few years we can see that there have been new words commonly used by Singaporeans, as these new words have popped out in the 2019 data set amongst the highest similarity of a specified word.

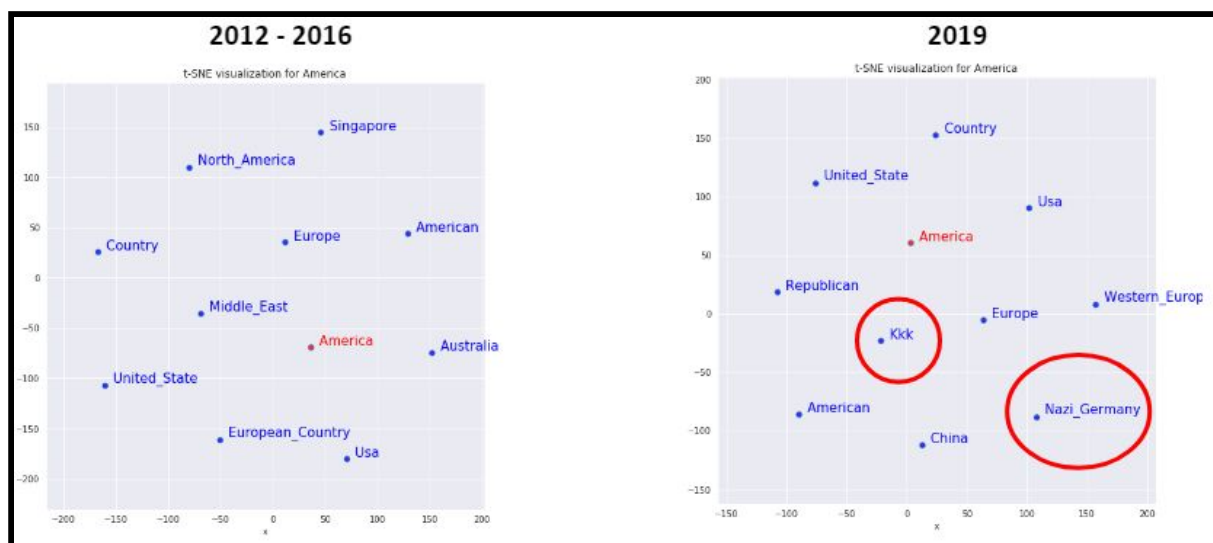


**Figure 3.3: t-SNE Visualisation of the word “E Scooter” and its Top 10 Most Similar Words**

In this example, we can see that by looking at the most similar words to “E Scooter” the word PMD standing for “Personal Mobility Device” only started to appear in the 2019 data set and was not frequently used in 2012-2016. This may be due to the introduction of the term Personal Mobility Device by the Land Transport Authority (LTA) in the Active Mobility Act in 2018. Moreover, recently there has also been a lot of controversy regarding the banning of Personal Mobility Devices on walking pavements in Singapore.

### 3.7 How Singaporean’s Opinions on Issues have Changed

By looking at the related words to particular entities, we can gather a sense of how the publics’ opinion towards these entities have shifted over time.



**Figure 3.4: t-SNE Visualisation of the word “America” and its Top 10 Most Similar Words**

Singaporean redditors’ viewpoint of America has seemed to change, with racist entities such as the Ku Klux Klan (KKK) and Nazi Germany being related to the word America. We can infer this to be due to the new US president Donald Trump who entered the White House in 2017, who has been a highly controversial public figure.

### 3.8 Analysing Gender Bias amongst Singaporeans

To understand Gender Bias, we looked at the analogical relations between the word vectors.

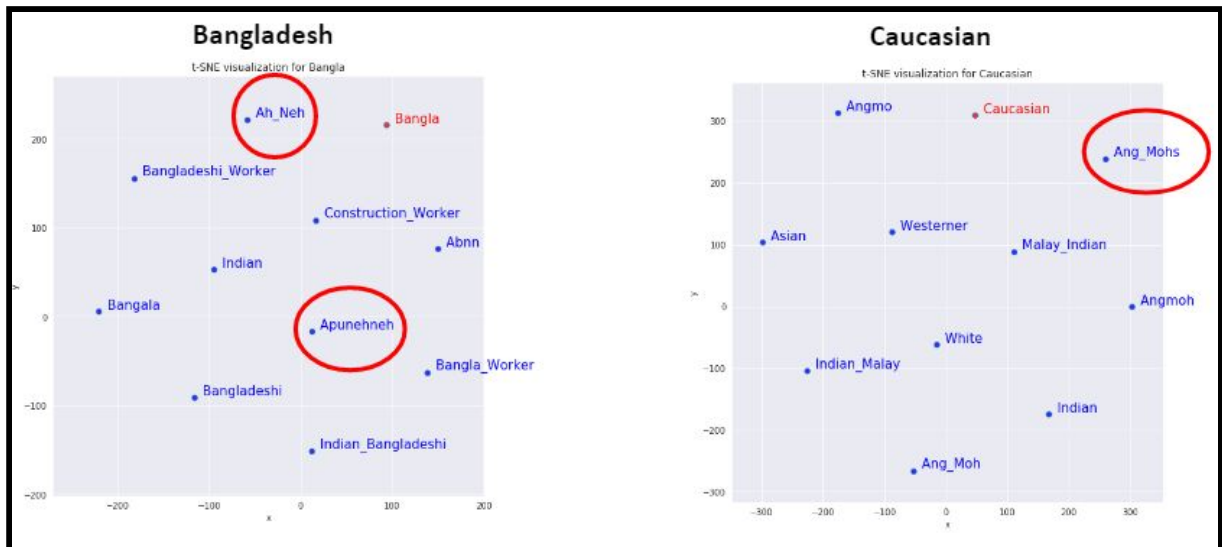
In Word2Vec vectors, **King - Man + Woman = Queen**, therefore we can use this method to look at analogies generated from embeddings trained on r/Singapore comments. We found that the embedding analogies tend to be sexist.

**Strong - Male + Female = Weak**  
**Top - Male + Female = Bottom**  
**Degree - Male + Female = Diploma**  
**Clothe - Male + Female = Change Diaper**

This set of equations clearly display that Males tended to be closer related to more positive and stronger adjectives and verbs. For example, males are related to a Degree which is one step higher in the education path than a Diploma, which females are related to.

### 3.9 Insensitivity to other Countries or Groups of People

We decided to look at specific ethnic groups or nationalities to see what similar words were extracted. We can look out for derogatory terms within the similar words to understand how Singaporeans feel about these groups.



**Figure 3.5: t-SNE Visualisation of the word “Bangla” and “Caucasian” with their Top 10 Most Similar Words**

The words highlighted in red in the figure above may be considered insensitive. These words have shown to be used by Singaporeans in association with the country “Bangladesh” and the race “Caucasian”

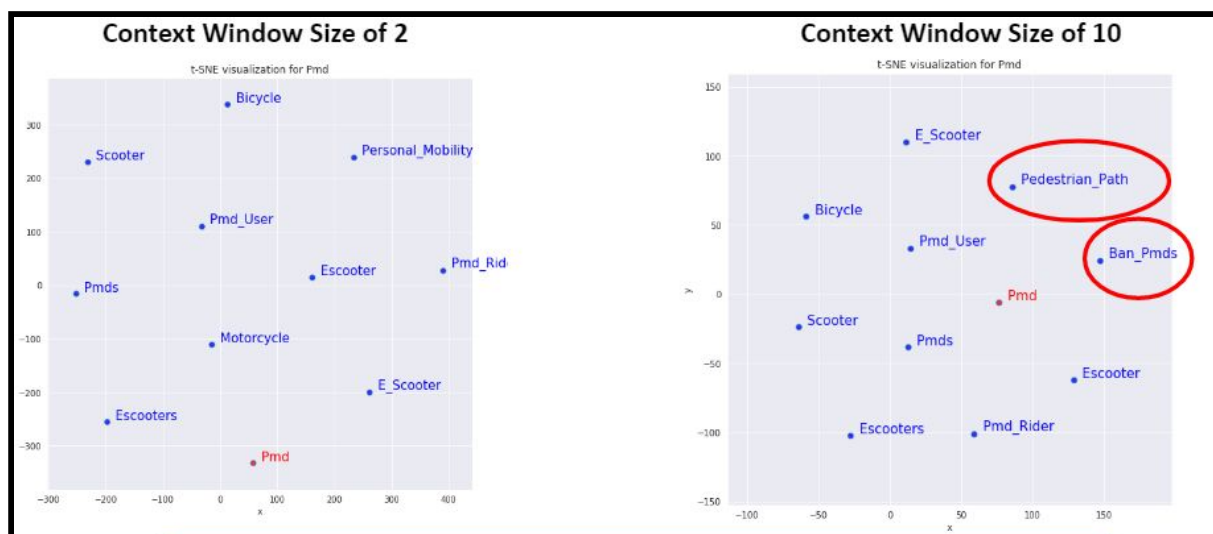
### 3.10 Differences in Results Between a Model with a Context Window Size of 2 vs 10

By comparing the 2 different context window sizes of 2 and 10 respectively, we discovered that the larger window size captured larger amounts of topical and domain information whereas smaller window sizes captured more information about the specified word itself or synonyms. These findings align with a paper by Levy & Goldberg on "Dependency-Based Word Embeddings"<sup>2</sup>

For example when comparing the most similar words from “PMD” in the 2019 dataset, we can see that all the similar words captured in a context window size of 2 were more related to the actual Personal Mobility Device as a vehicle. However with a context window size of 10, we can see that words such as “Pedestrian Path” and “Ban PMDs” start appearing. This is related to the recent issue where Personal Mobility Devices were banned from pedestrian sidewalks.

<sup>2</sup> Levy, & Goldberg. 2014. Dependency-Based Word Embeddings.  
<https://levyomer.files.wordpress.com/2014/04/dependency-based-word-embeddings-acl-2014.pdf>





**Figure 3.6: t-SNE Visualisation of the word “PMD” and its Top 10 Most Similar Words**

### 3.11 Differences in Results Between a Model with Dimension Size of 300 vs 60

To our surprise the results were relatively similar between the model with dimension size of 300 and the model with dimension size of 60. The semantic relations between the words still seemed accurate with a dimension size of 60. After doing further research, we discovered that according to Google Developers<sup>3</sup> that there was a general rule of thumb:

**Number of Dimensions = Vocabulary Size \*\* 0.25**

In the case of our largest model with a vocabulary size of 52552, the number of dimensions should be: **52552 \*\* 0.25 = 15**.

Thus we could look at possibly reducing the number of dimensions to reduce training time as a dimension size of 60 has shown to capture sufficiently accurate information. Furthermore, Google’s rule of thumb suggested that the dimension size should be even smaller at 15.

## 4 GPT-2

GPT-2, was released in 2019 by OpenAI, trained on 40gb on internet text. It is trained by predicting the next word, given all the previous words in the document.

The GPT-2 paper, “Language Models are Unsupervised Multitask Learners”, discusses that training in such an unsupervised method allows for a general language model that is capable of many tasks. To directly train a language model in a supervised learning fashion requires labelled data, which is expensive and time-consuming to create. Therefore, one could first train a language model such as GPT-2 in an unsupervised fashion, and then later finetune

<sup>3</sup> TensorFlow Team. 2017. Introducing TensorFlow Feature Columns.  
<https://developers.googleblog.com/2017/11/introducing-tensorflow-feature-columns.html>

the weights on the specific task. Because of the method of training of GPT-2, it is inherently easy to use it for text generation. Simply give it an initial prompt, and the model will return text that is “chameleon-like” - adapting to the style and context of the prompt.

We use the medium size GPT-2 model which has 355 million parameters. While the 774M and 1.5B parameters version of GPT-2 has been released, it is unfeasible to finetune the model on new text data due to memory issues. We used the gpt-2-simple package which allowed for easy finetuning on new text samples. Using a NVIDIA Telsa P100 GPU, fine tuning the weights against our dataset took 2 hours for 8000 training steps.

## 4.1 Data preprocessing

Unlike many language models, GPT-2 training does not require any preprocessing in the word of tokenization and representation of out-of-vocabulary strings as a special token. Instead, it is able to compute the probability and generate any possible string. GPT-2 handles this with a modified version of Byte Pair Encoding (Sennrich et al., 2015).

Therefore, we did not perform any additional data preprocessing steps, other than taking comments above 100 characters. This means that GPT-2 will predict items such as line breaks, which is important in the syntactic style of articles.

## 4.2 Inference parameters

The main inference parameter to vary when generating text is temperature. Temperature is a value controlling randomness in boltzmann distribution in the final output. Lower temperature results in less random completions, a high value creates more “creative” completions. As the temperature approaches zero, the model will become deterministic and repetitive. We found that a temperature of 0.8 generated text that was interesting but also possible. We also attach text files containing inference results under the folder GPT-2.

## 4.3 General article generation

GPT-2 can generate very believable passages given a sufficient prompt. We highlight the relevant phrases in the text.

For example, we use the prompt “In NUS, a group of students took the module BT4222, Mining Web Data for Business Insights, under Assistant Professor Qiao Dandan. The students had to create a project using Natural Language Processing tools. One of her students, James, did not know what to do for his project and”.

Generated:

*(continued from prompt) so he spent hours creating memes to throw at the other students. A similar phenomenon happened in the other school of the same level. It appears that the students were unable to create a task as simple as pick a random phrase from a common dictionary,*

Comments:

Here we see the success of GPT-2 managing long-range word dependency. Although the word “Language” was mentioned 62 words ago, it managed to infer that “dictionary” was a

probable word to use in this context. Indeed, according to “Attention is all you need” by Vaswani et al. 2017, self-attentive layers are usually much faster than recurrent layers due to reduced computational complexity. Also, by directly providing a path from the inputs, this avoids the vanishing gradient problem faced by RNNs, allowing for long-range dependency.

## 4.4 Singlish text generation

An important goal was to generate Singlish text. While most of r/Singapore data was written in Singaporean Standard English, hopefully there were enough examples of Singlish. We gave the following prompts in order to generate Singlish text.

### 4.4.1 Walao eh

Walao eh is an exclamation literally meaning my father, used to indicate surprise or disappointment.

Generated:

*walao eh shiok, im a meathead*

Comments:

Here we observe the model successfully inferring “shiok” in conjunction with “meathead”. Shiok is another Singlish word indicating something is very enjoyable or delicious. This, together with the word meathead - is a possible Singlish sentence.

Generated:

*walao eh lor. You can't say ill give the poo to you during school*

Comments:

“lor” is a possible singlish particle that indicates an obviousness or sense of resignation.<sup>4</sup> However, “walao eh lor” does not make sense. Furthermore, the sentence as a whole is not coherent.

Generated:

*walao eh, you got scholarship above your means, you like studying?*

Comments:

While the sentence is not very coherent, the sentence is written in a rather Singlish way. Here, “you got scholarship” is generated instead of the Standard English “you have a scholarship”. Singlish employs got-constructions, where the use of the word *got* is similar to the Hokkien *wu* and Mandarin *you* (有)<sup>5</sup>.

### 4.4.2 Knn ccb

Knn ccb are two singlish acronyms for expletives.

Generated:

---

<sup>4</sup> Wee, Lionel. 2004. Reduplication and discourse particles.

<https://scholarbank.nus.edu.sg/handle/10635/130449>

<sup>5</sup> Vaswani, Shazeer, PArmar, Uszkoreit, Jones, Gomez, Kaiser, Polosukhin. 2017. Attention is all you need. <https://arxiv.org/abs/1706.03762>

*ccb knnfk. Didnt see anything wrong with that. I prefer more mala.*

Comments:

Interestingly, “mala”, a popular Sichuanese dish is mentioned. We can also observe “fk” being added, which is another possible expletive.

Generated:

*ccb knn to play liao. But oh well. I won't join an SG group.*

Comments:

The model does predict the word “liao”, a singlish particle similar to the mandarin 了 (le) . However, its use here is not coherent.

Generated:

*ccb knn sian. I mean if you're dead and resting...*

Comments:

“Sian” indicates disappointment or boredom. Here, the model uses it correctly.

Overall, while the model generated singlish words in short-range, most of the comments generated were not very successful. This could be due to the lack of singlish sentences in r/Singapore text which tends to be written in Standard English. It is difficult to acquire sufficient examples for model training.

## 4.5 Singaporean context text generation

Efforts at generating passages and articles in a Singaporean context were far more successful.

### 4.5.1 SINGAPORE

One can use SINGAPORE in caps to generate articles regarding singapore.

Generated:

*SINGAPORE - The Singapore Civil Defence Force (SCDF) and the Singapore Police Force (SPF) will jointly launch an exercise-related exercise in coordination with their relevant counterparts in the region, the Public Transport (Town Council) Board said on Monday (Oct 28).*

*It said the exercise will be an enhanced alert exercise where officers will operate at their very best to prevent any risks associated with suspicious activity.*

*The exercise will be held at the Singapore Civil Engineering School (SCES) and will take place over two days, and involve all OCS.*

*On the design of the exercise, the Public Transport Board said the exercise would have a clear theme which will be the Singapore Civil Engineering School.*

*It will also explore how to improve the training and operations for the exercise.*

Comments:

Generally, the article seems to be rather believable, where the Singaporean civil service coordinates with the public transport authority to look out for suspicious terrorist activities. However, it seems like long-range term dependency went awry here, as the “Civil” in Civil Defence was used to generate the “Civil Engineering School”, which is not coherent.

#### 4.5.2 PAP

The PAP refers to the People’s Action Party, the governing party in Singapore.

Generated:

*PAP's dominance in the region is now so strong that Singaporeans are likely to vote for them. That, coupled with the fact that many of the Singaporeans who voted for Trump are white nationalist, cloaked in explained patriotism, makes me wonder if the country would become as politically impotent as it has been for the past 50 years. Singaporeans are generally apolitical and conservative. They have little interest in politics and politics are their only hobby.*

Comments:

The PAP is successfully associated as a political party. However, the model generates a comment that is more likely to stem from an American context, talking about Trump and white nationalism. This could be explained from how GPT-2 was trained - utilising internet articles that had a much higher likelihood to discuss US affairs. Therefore, this problem could be resolved by training the model for more epochs on Singapore related articles.

Generated:

*PAP MP for GRC Teo Chee Hean's Teck Whye constituency apologised to the families of all the families who had lost their loved ones, as well as to the families of the victims. He added that Mr Teck's conduct has been "a matter of public fascination and reflection" since his election in May 2014. The incident, which happened on December 2 last year, prompted Mr Teck to apologise on behalf of Char Boon Khi Wong, who lost her husband, Jeyaratnam Syajee, a businessman, in a Park Geil Air flight in 2013.*

Comments:

Here a very believable passage is generated, using a real minister’s name, Teo Chee Hean, and real constituency names. The passage is very coherent, with a consistent topic being the loss of lives throughout the text.

#### 4.5.2 Singaporean meme generation

In 2010, the Navy Seal copy-pasta meme was spread throughout the internet, taking various forms:

*“What the fuck did you just fucking say about me, you little bitch? I’ll have you know I graduated top of my class in the Navy Seals, and I’ve been involved in numerous secret raids on Al-Quaeda, and I have over 300 confirmed kills. I am trained in gorilla warfare and I’m the top sniper in the entire US armed forces.....”*

*“What the fuck did you just fucking type about me, you little bitch? I’ll have you know I graduated top of my class at MIT, and I’ve been involved in numerous secret raids with Anonymous, and I have over 300 confirmed DDoSes. I am trained in online trolling and I’m the top hacker in the entire world. ...”*

*“What the fuck did you just fucking say about my gear, you little n00b? I’ll have you know I am a lvl 90 Undead Arcane Mage, and I’ve won so many PVP matches, and I have done raids on every 10 man heroic dungeon. I also have a fuckton of macros and I have a GS of 10K. You are nothing to me but just a lvl 12 gnome hunter. ...”*

We wanted to observe if GPT-2 could generate memes in a Singaporean context. Therefore, the following prompt was given “What the fuck did you just fucking say about me, you little bitch? I’ll have you know I graduated top of my class in the SAF”. Hopefully, the model would use SAF - the Singapore Armed Forces as a dependency to generate text related to Singapore and the Singaporean Army.

Generated:

*What the fuck did you just fucking say about me, you little bitch? I’ll have you know I graduated top of my class in the SAF, with a double degree from NUS and NTU. I have connections in the intelligence community, including being the top marksman in my entire unit and I can practically shoot you down with my fucking simple rifle.*

Comments:

The model successfully generates achievements related to Singaporeans - “a double degree from NUS and NTU”. In the Singaporean army, an award of “marksman” is a possible achievement as well. The generated words therefore are related with the SAF. Furthermore, it finishes in the style of the meme - showing high coherence and long-range dependency throughout the text.

## 5 Conclusion

From the Word2Vec analysis, by comparing two datasets from different time periods, we can see how Singaporeans use certain words can rapidly change over a short span of a few years. This is also known as Historical Semantics. On the other hand, even new words have been added into our vocabulary within a short period. Singaporeans’ opinions on other countries and political issues also can quickly shift .

By studying the Analogical Relations of the word vectors we can detect inherent biases and prejudice that we, Singaporeans, may hold but not even be aware of ourselves, only for it to be uncovered by Word Embeddings. Thus Word Embeddings can be used for a progressive good to create a better world and reveal hidden biases.

We also generate that state of the art models such as GPT-2 can generate realistic text in a Singaporean context. In fact, OpenAI was initially reluctant to release its full 1.5 billion parameter model due to its ability to generate fake news. However, according to OpenAI, it has not observed such instances yet. We had limited success in creating Singlish text, due to the lack of training samples. While sentences involving Singlish words could be generated, overall the Singlish grammar was not well abided to. Future efforts could be improved by training on a more extensive Singlish corpus. With a model trained on Singaporean text data, one can finetune the model for business use, such as sentiment analysis to gauge a product's reception.

## 6 Bibliography

1. Leimgruber. 2009. Modelling Variation in Singaporean English.  
<https://jakobleimgruber.ch/dbFile/48/u-c661/u-c661/dissertation.pdf>
2. Levy, & Goldberg. 2014. Dependency-Based Word Embeddings.  
<https://levyomer.files.wordpress.com/2014/04/dependency-based-word-embeddings-acl-2014.pdf>
3. TensorFlow Team. 2017. Introducing TensorFlow Feature Columns.  
<https://developers.googleblog.com/2017/11/introducing-tensorflow-feature-columns.html>
4. Wee, Lionel. 2004. Reduplication and discourse particles.  
<https://scholarbank.nus.edu.sg/handle/10635/130449>
5. Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, Polosukhin. 2017. Attention is all you need. <https://arxiv.org/abs/1706.03762>
6. Radford, Wu, Child, Luan, Amodei, Sutskever. 2019. Language Models are Unsupervised Multitask Learners.  
[https://d4mucfpksyv.cloudfront.net/better-language-models/language\\_models\\_are\\_unsupervised\\_multitask\\_learners.pdf](https://d4mucfpksyv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf)