Cross-Cultural Analysis in Social Media Comments

Risheng Lian (rl3162)

Directed Research Report

Advisor: Prof. John R. Kender

Columbia University

December 28, 2021

Abstract

In the context of globalization, people of different cultural backgrounds are often faced with the same big events such as the Covid-19 pandemic, opportunities and challenges brought by artificial intelligence, climate change issues and so on. Thanks to the very popular social media applications, we can easily get some comments and responses from different affinity groups upon a specific international event [1]. We think it is very meaningful to collect and analyze these distinct cultural viewpoints through natural language processing techniques. In particular, we conducted word embedding methods to learn geometrical encodings at first, then we focused on sentiment analysis of comments about Covid-19 pandemic events written in English and Chinese. Our results show that there are significant divergences in valence scale between responses from American and Chinese cultures.

Keywords: Word embedding, middle-frequency words, GloVe model, Sentiment Analysis, VADER

1 Introduction

The goal of our research project is to study the differences and similarities in the social media comments towards the same events across several affinity groups. At the beginning, we try the word embedding techniques in order to get representations useful for our final tasks. After that, we would like to apply some sentiment analysis model with interpretability, which means it is not a black-box model and has excellent performance as well. With our model capable of labeling the reviews and comments as positive, neutral or negative, we would like to quantify the emotional labels further based on the valence scale of social media comments, i.e. how positive or how negative these comments can be. Hence, we could use the statistical results to compare the culture differences between different affinity groups.

In this directed research report, we will mainly focus on the dataset constructed by Luvena Huo [2] in her previous work. In view of this dataset, we apply the Global Vectors (GloVe) [3] for word embedding. Our expectation is to obtain some good representations for semantic classification or regression. Next, we explore a simple but powerful rule-based model for sentiment analysis: VADER [4]. Combined this model with other essential toolkits such as Google-translate API, NLTK [11] and jieba [5] packages, we could conduct the sentiment analysis task and obtain the sentiment intensity score (valence-based) for corpus in our dataset.

The main contribution of this research project can be summarized as follows:

- We apply GloVe model to get meaningful word embeddings on Kaggle dataset and explore the feasibility to use this methods on the dataset that we are interested.

- We investigate the dataset collected by Luvena and do sentiment analysis using VADER model for the corpus written in English.
- We explore the VADER model combined with NLTK and jieba toolkits to conduct sentiment analysis for the corpus written in Chinese, and compare the quantified result.

2 Related Works

2.1 Word embedding methods

Word embeddings are a learnable representation for text in which words with similar meanings tend to be close to each other geometrically. It could capture the semantic, syntactic context or a term and helps understand how likely it is to have the similar meaning with other terms in an article, a micro blog, and so on. We expect to exploit this property to see if we could obtain useful representations for sentiment analysis. Among the word embedding methods, Word2Vec and GloVe are very popular these days. Both models could learn spacial encodings of words from how frequently these words appear together in a corpora. We choose to use GloVe model because many work suggested that when we control the training hyperparameters, the embeddings generated from these two methods tend to be very similar in downstream NLP tasks. In addition, compared with Word2Vec, it is easier to parallelize the implementation [3]. Thus, the training process could be more efficiently.

2.2 Mainstream emotion models

Emotion classification and regression are widely discussed topic in psychology. Many Natural Language Processing (NLP) researches pay close attention to Ekman's model, where anger, fear, sadness, joy, disgust and surprise are the six basic emotions. Other works [2, 7] in NLP begin to focus on Russel's circumplex model [8] which suggests the emotion states can be represented as a two-dimensional spaces with two independent axises: valence and arousal. Recently, some NLP studies [9] indicate that these emotional axises could be regarded as dependent and such model design could be beneficial to emotional tasks. In this research project, we focus on one single emotional dimension only, which is valence.

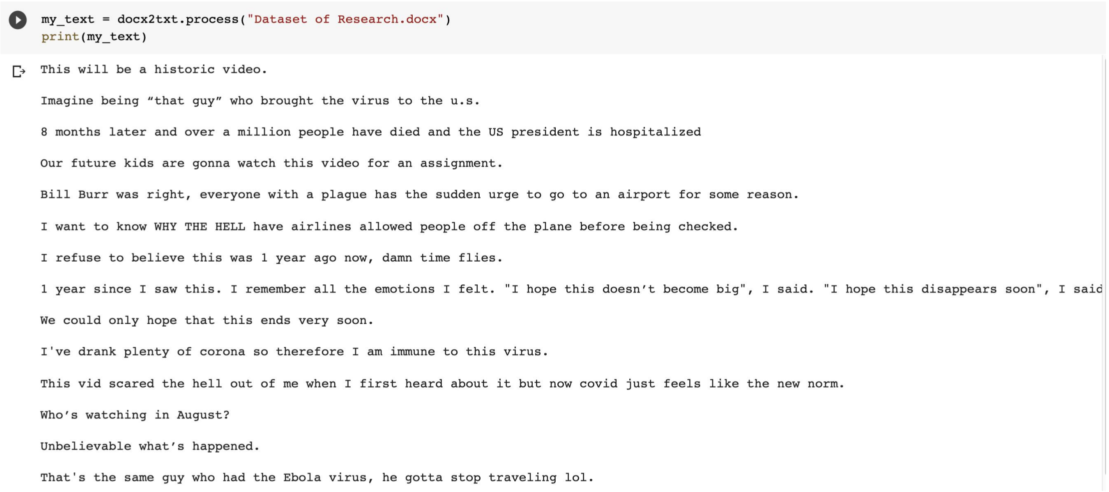
2.3 Sentiment Analysis models

Sentiment analysis is a very active area of study in the field of NLP researches. The goal is to extract and analyze subjective information from corpus such as opinions and attitudes [10]. Much work are conducted based on machine learning approach [1], even if they could achieve high performance, there exists many shortcomings. For example, these methods derives features that are not easily human interpretable, which could harm the ability of models to generalize, modify and extend [4]. Apart from that, many works rely on sentiment lexicons approach, where words are manually categorized in many classes (like positive or negative) or associated with real-valued scores for specific sentiment intensity. This approach is very robust and highly interpretable by humans, we will adapt VADER lexicons as well as some generalizable heuristic rules introduced in [4] to perform our sentiment analysis task.

3 Materials and Methods

3.1 Experiment datasets

We mainly focus on the dataset constructed by Luvena in her previous work. It is consist of comments toward Covid-19 topic in both English and Chinese from YouTube webpage (See **Fig. 1-2**). The specific timeline for these comments is early 2020, during the initial outbreak of the coronavirus. Since both the American people and Chinese people are in the same situations for this global pandemic, we could make a comparison between these two affinity groups. In addition, it is a relatively small corpus since sentences in English and Chinese are less than 100.



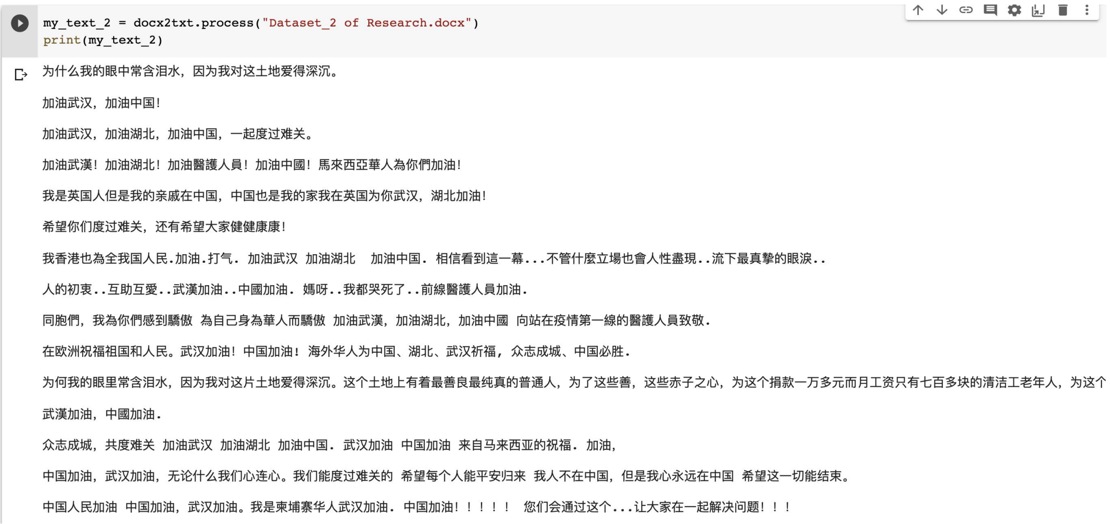
```
my_text = docx2txt.process("Dataset of Research.docx")
print(my_text)
```

```
This will be a historic video.

Imagine being “that guy” who brought the virus to the u.s.

8 months later and over a million people have died and the US president is hospitalized

Our future kids are gonna watch this video for an assignment.

Bill Burr was right, everyone with a plague has the sudden urge to go to an airport for some reason.

I want to know WHY THE HELL have airlines allowed people off the plane before being checked.

I refuse to believe this was 1 year ago now, damn time flies.

1 year since I saw this. I remember all the emotions I felt. "I hope this doesn’t become big", I said. "I hope this disappears soon", I said

We could only hope that this ends very soon.

I've drank plenty of corona so therefore I am immune to this virus.

This vid scared the hell out of me when I first heard about it but now covid just feels like the new norm.

Who’s watching in August?

Unbelievable what’s happened.

That's the same guy who had the Ebola virus, he gotta stop traveling lol.
```

Figure. 1. Examples of English comments in our dataset.

3.2 GloVe for word embeddings

First of all, we would like to try the Global Vectors (GloVe) to obtain the word embeddings. This model is often used to deal with text similarity tasks and named entity recognition. And it could produces a vector space with meaningful geometrical structure. We expected to obtain some useful representations for semantic classification or regression, then we could plug them into models like neural networks to perform these tasks.



```
my_text_2 = docx2txt.process("Dataset_2 of Research.docx")
print(my_text_2)
```

```
为什么我的眼中常含泪水，因为我对这土地爱得深沉。

加油武汉，加油中国！

加油武汉，加油湖北，加油中国，一起度过难关。

加油武漢！加油湖北！加油醫護人員！加油中國！馬來西亞華人為你們加油！

我是英国人但是我的亲戚在中国，中国也是我的家我在英国为你武汉，湖北加油！

希望你们度过难关，还有希望大家健健康康！

我香港也為全我国人民.加油.打气. 加油武汉 加油湖北  加油中国. 相信看到這一幕...不管什麼立場也會人性盡現..流下最真摯的眼淚..

人的初衷..互助互愛..武漢加油..中國加油. 媽呀..我都哭死了..前線醫護人員加油.

同胞們，我為你們感到驕傲 為自己身為華人而驕傲 加油武漢，加油湖北，加油中國 向站在疫情第一線的醫護人員致敬.

在欧洲祝福祖国和人民。武汉加油！中国加油！ 海外华人为中国、湖北、武汉祈福， 众志成城、中国必胜.

为何我的眼里常含泪水，因为我对这片土地爱得深沉。这个土地上有着最善良最纯真的普通人，为了这些善，这些赤子之心，为这个捐款一万多元而月工资只有七百多块的清洁工老年人，为这个
武漢加油，中國加油.

众志成城，共度难关 加油武汉 加油湖北 加油中国. 武汉加油 中国加油 来自马来西亚的祝福. 加油，

中国加油，武汉加油，无论什么我们心连心。我们能度过难关的 希望每个人能平安归来 我人不在中国，但是我心永远在中国 希望这一切能结束。

中国人民加油 中国加油，武汉加油。我是柬埔寨华人武汉加油. 中国加油！！！！！ 您们会通过这个...让大家在一起解决问题！！！
```

Figure. 2. Examples of Chinese comments in our dataset.

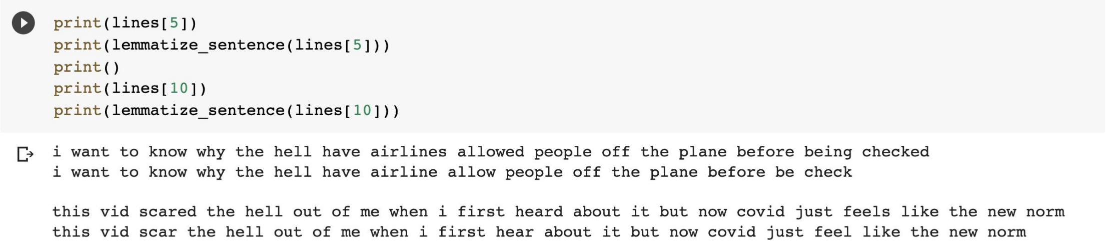
3.2.1 Compute the co-occurrence matrix

A co-occurrence matrix have specific entities in rows (ER) and columns (EC). For instance, if we just treat the adjacent words in a sentence as pair of words that co-occurs, then we could get such matrix shown in **Table.1**.

The aim of this matrix is to present the time counts each ER appears in the same context as each EC, as a result, more similar words are more likely to co-occur. In general, Glove is a co-occurrence based model. It starts by going through the entire corpus and constructing a co-occurrence matrix. In our project, if two specific words are no more than five positions apart, we treat them as a valid count. Before computing this matrix for our dataset, we need to use NLTK toolkit [11] to perform word tokenization, stemming and lemmatization for English comments (**Fig. 3**.). These data pre-processing techniques indeed helps considering our dataset are not large enough.

	Cats	are	cute	Moonlight	is	beautiful
Cats	1	1	1	0	0	0
are	1	1	1	0	0	0
cute	1	1	1	0	0	0
Moonlight	0	0	0	1	1	1
is	0	0	0	1	1	1
beautiful	0	0	0	1	1	1

Table. 1. Examples of co-occurrence matrix for the following text: “Cats are cute. Moonlight is beautiful. ”



```
print(lines[5])
print(lemmatize_sentence(lines[5]))
print()
print(lines[10])
print(lemmatize_sentence(lines[10]))
```

```
i want to know why the hell have airlines allowed people off the plane before being checked
i want to know why the hell have airline allow people off the plane before be check

this vid scared the hell out of me when i first heard about it but now covid just feels like the new norm
this vid scar the hell out of me when i first hear about it but now covid just feel like the new norm
```

Fig. 3. Examples of word lemmatization for English corpus. Specifically, the upper sentences are the original comments in our dataset, and the lower ones are the result of lemmatization.

3.2.2 Training process

Denote our co-occurrence matrix as X and the vocabulary size of our dataset as V, then X is a V × V matrix. The primary goal of GloVe method is to fit a rank-K approximation of X,

$$X \approx R\tilde{R}^T$$

where R and $\tilde{R}^T$ are V × K matrix. K is our hyperparameter for feature dimension, V > K. In this case, each row $\boldsymbol{r}_i$ of R is the K-dimensional representation of a word. And each entry is approximated as

$$x_{ij} \approx r_i^T \tilde{r}_j$$

Thus, we can minimize the squared error, it is actually the same as Principle Component Analysis (PCA).

$$||X - R\tilde{R}^T||^2 = \sum_{i,j}(x_{ij} - r_i^T\tilde{r_j})^2$$

Finally, the objective function J is defined as follows:

$$J(\theta) = \sum_{i,j} f(x_{ij})(r_i^T\tilde{r_j} - log(x_{ij}))$$

$$f(x) = \begin{cases} (x/x_{\max})^{\alpha} & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases} .$$

where f is our weighted function. Usually, word counts are a heavy-tailed distribution, so the most common words will dominate the objective function, this is the reason why we use the logarithm of x in J.

In particular, we will split the dataset into training, validation and testing parts, and use the gradient decent methods to minimize the value of objective function at training process. We will also use validation dataset to figure out the best settings of hyperparameters.

3.3 Sentiment Analysis on English and Chinese comments

Next, we explore a simple but powerful rule-based model for sentiment analysis: VADER. This model is constructed and empirically validated through a gold-standard list of lexical features, which are attuned to sentiment in microblog-like contexts, so it is interpretable and very suitable for the dataset. We apply this model and combined with other essential toolkits such as Google-trans API, NLTK and jieba packages to conduct the sentiment analysis task and obtain the sentiment intensity score (valence-based) for corpus in our dataset.

3.3.1 Lexicons and standard rules

Unlike the traditional machine learning approach, the VADER model relies on its sentiment lexicon to make the decision. Not only could it produce the polarity(positive/negative) label for word, sentence or article, but also score the intensity on a scale from -4 to 4. It mostly inherits the elements of existing well-established lexicons such as LIWC, GI and ANEW [12]. Besides, it adds additional lexical features such as emoticons and acronyms, which are commonly used in social media to express feelings. Based on the wisdom of the crowd approach, it has established polarity and intensity scores of sentiment valence for over 9,000 lexical features. These items comprises the VADER sentiment lexicon. For example, the word "okay" has a positive valence of 0.9, "good" is 1.9, and "great" is 3.1, whereas "horrible" is –2.5, the frowning emoticon ":(" is –2.2, and "sucks" and "sux" are both –1.5 [4].

Furthermore, the model add several generalizable heuristics based on grammatical and syntactical cues to convey changes to sentiment intensity. For example, using the exclamation point or capitalization words could increase the absolute score of intensity, without changing the polarity labels; the contrastive conjunction "but" signals a shift in sentiment polarity, with the sentiment of the text following the conjunction being dominant [4]. In short, associated with these heuristics, VADER goes beyond the typical bag-of-word models.

3.3.2 Natural language processing for Chinese part

In this project, we apply VADER model to get the quantified valence sentiment results of English comments directly. As for Chinese comments, since VADER's lexicon does not contain any Chinese lexical features, we will adopt Google trans API to convert the Chinese comments into English at first. However, due to the property of Chinese texts, idioms, proverbs and even the space character could potentially change the results of Google trans API completely. In fact, the segmentation task for Chinese corpus could be very challenging. As a modification, we choose to use the "Jieba" [5], which is known as one of the best Chinese text segmentation tools, to split the characters in a more reasonable way. And then apply Google trans to convert these word segment into English. Although this could make the translated sentences not so coherent, it could still retain the contextual and logical relationship somehow. Since the VADER is sentiment lexicon based model, this modification has no conflict with our model. Therefore, we expect to obtain a more accurate sentiment analysis result for Chinese corpus.

4 Results

4.1 GloVe embeddings

In this section, we will show some graphs of learned representations of GloVe, these are high dimensional vectors, so we choose to apply the t-Distributed Stochastic Neighbor Embedding (t-SNE) [13] as our data exploration and visualization method. Figure 4 provides an evidence that using GloVe could indeed generate some meaningful structures of word representations where words having the similar semantic meaning or commonly used together tend to form a cluster.

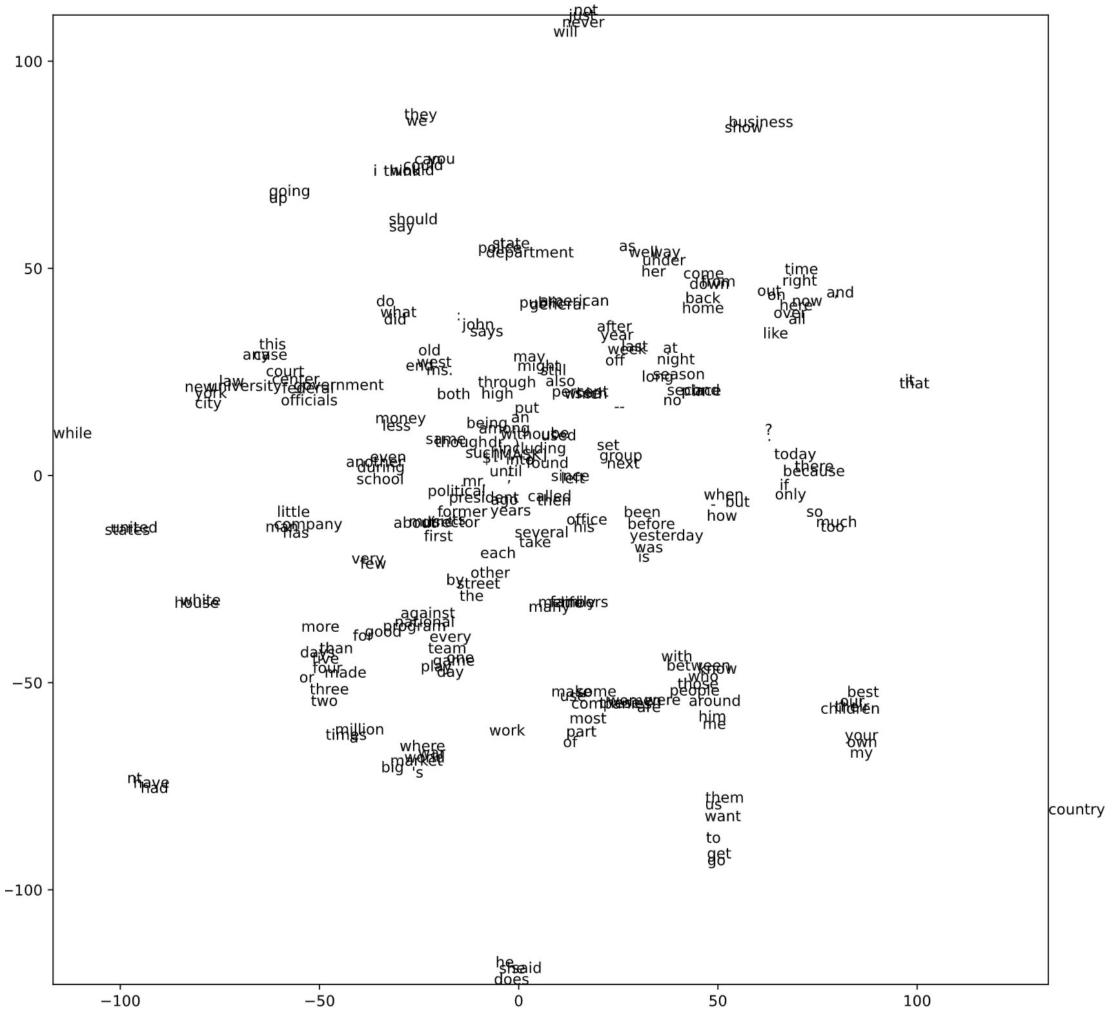


Fig. 4. Visualization of word representations learned by Glove. Specifically, we use the toy examples from [14] to train and validate the model, and choose the best feature dimensions according to the validation loss. In this case, we choose perplexity = 3, iterations = 1000 and early_exaggeration = 10 for t-SNE parameter settings.

We could also use middle-frequency extraction technique to remove the less important words form the vocabulary. For example, high-frequency stop-words and words rarely appears in our corpus. To do this, we need to obtain the word count plot (Figure 5-6) in advance.

However, such model does not work very well on our interested dataset. Figure 7 indicates that the middle-frequency words of our corpus about Covid-19 pandemic [2] do not cluster based on their semantic similarities. One of the possible reasons would be that this corpus is not large enough to build a co-occurrence matrix with sufficient information for the task. We need to collect more documents related with such topic to perform the GloVe method.

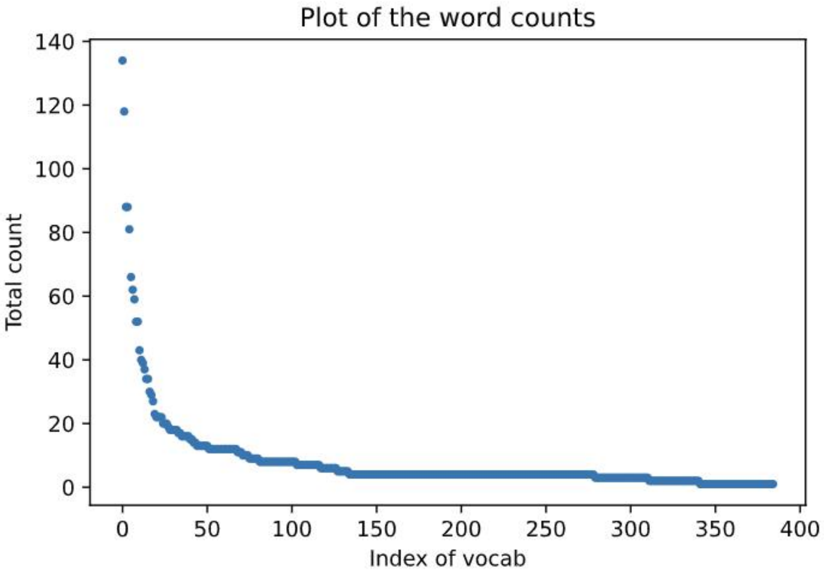


Fig. 5. Plot of the word counts from our dataset [2], the vocabulary size is 385.

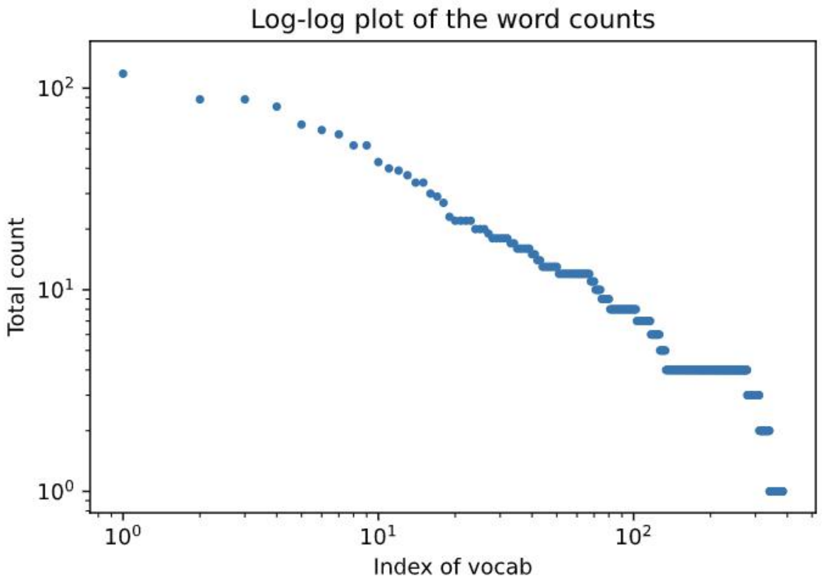


Fig. 6. Log-log plot of the word counts, we can use this graph to define the middle-frequency words for our dataset (the word index ranges from 24 to 80).

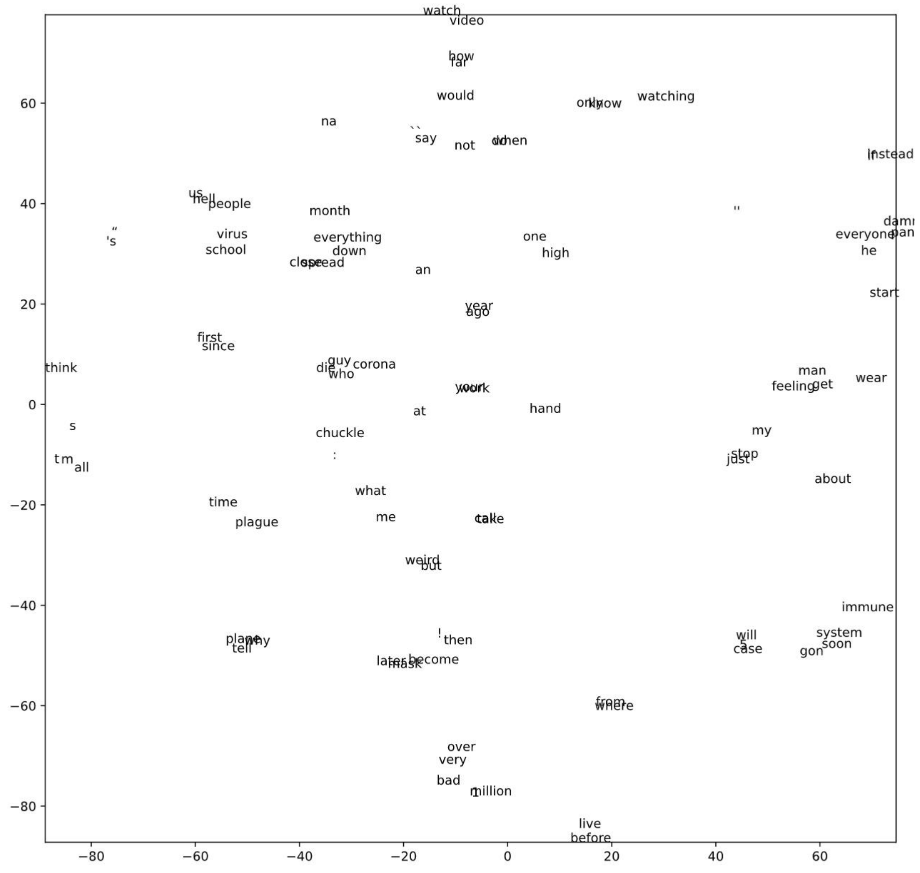


Fig. 7. Visualization of word representations learned by Glove using our dataset [2],

filtered by middle-frequency words. We choose perplexity = 3, iterations = 1000 and early_exaggeration = 10 for t-SNE parameter settings.

4.2 Comparison of VADER sentiment analysis

4.2.1 Result of English and Chinese comments

In this section, we will compare the sentiment analysis results, i.e. the valence scores of our dataset using VADER. In particular, we normalize these scores from -1 to +1, where the minus sign means negative attitude, zero means neutral, and the plus sign means positive attitude. From those histograms (Figure 8), we can see that the valence of English comments is like a normal distribution from negative to positive attitudes. Most of them are neutral, as the strength of emotion increases (i.e. the absolute value of valence scores), the corresponding density tend to decrease.

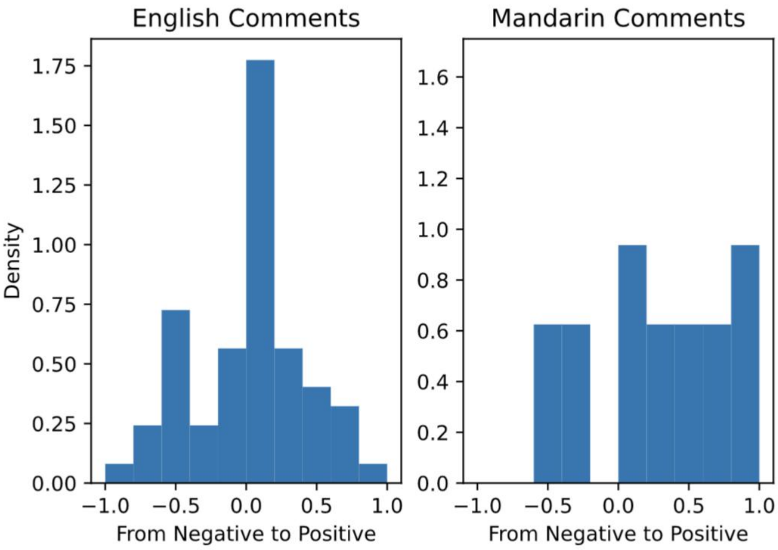


Fig. 8. Histograms of valence scores from negative to positive, before using Jieba

toolkit. Mandarin comments stands for comments written in simplified Chinese.

As for Chinese comments, valence scores are tend to be distributed in areas that represent positive emotions. The density for each intensity is very close, and it looks like an approximately uniform distribution from -0.5 to 1.0, which suggests that the Chinese comments of our dataset are more positive than English comments.

4.2.2 Changes by using Jieba toolkit

When we adopt Jieba toolkit to perform Chinese word segmentation, the result of Chinese comments becomes slightly different from the previous one (Figure 9-10), but this does not change the fact that valence scores of reviews in Chinese tend to be more positive, and our conclusion still holds. The modified methods generates a histogram which is more informative to us, since the previous one are close to uniform distribution. We believe the new sentiment analysis result of Chinese comments are more accurate mainly because Jieba does a better job than Google trans API in separating the words, idioms and proverbs from Chinese corpus. Bad segmentation and several space characters could potentially interfere with translator to convert Chinese texts into English correctly.

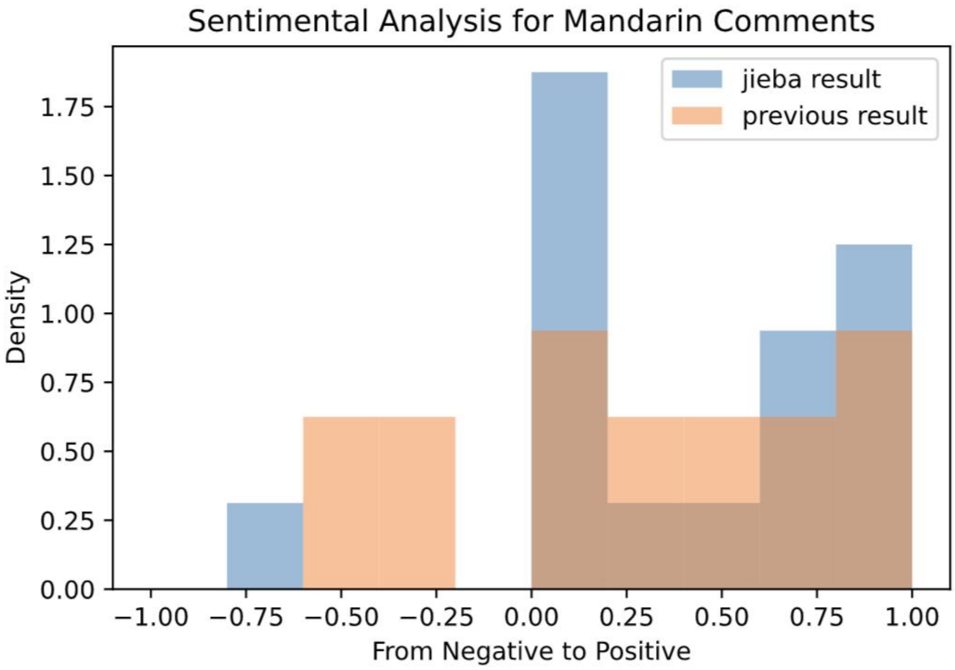


Fig. 9. Histograms of valence scores from negative to positive, the blue one represents result after using Jieba, and the orange one represent the previous result.

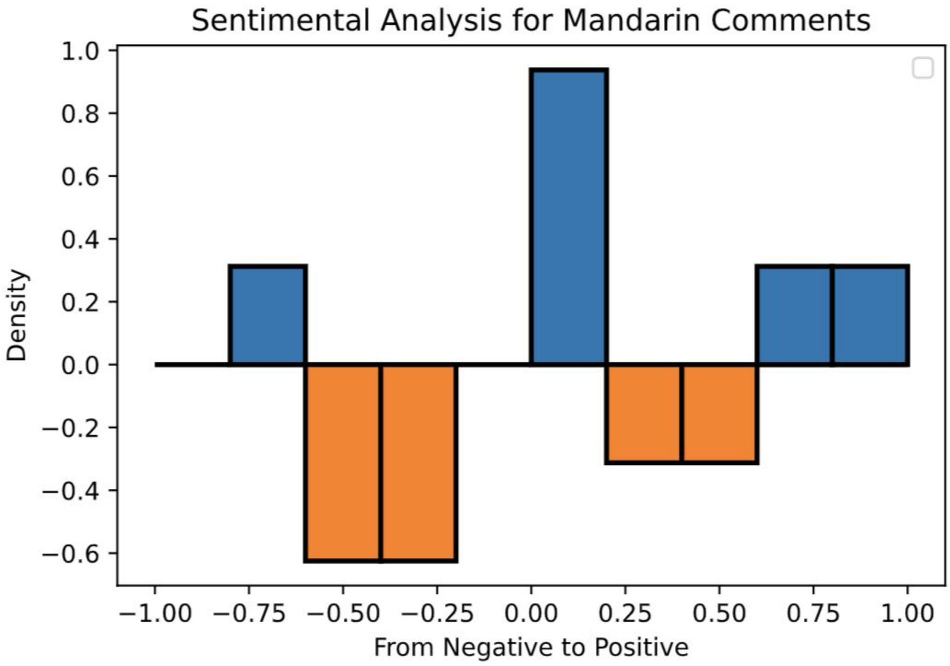


Fig. 10. Histograms of valence scores from negative to positive, the blue one represents that the distribution after using Jieba has more density in the specific region, while the orange one represents the previous result has more density. We can

see after using Jieba, there are more neutral scores and more extreme valence scores than before.

Discussion

In this research project, we adopt Glove embedding methods to learn word representations at first. As a consequence, the final result on our corpus [2] does not show a well-established structures compared with toy examples [14]. The reason might be that our dataset is relatively small, so it is difficult for the co-occurrence matrix to reflect the real semantic relationships of word pairs. One possible solution would be collecting more documents related to the Covid-19 topic, then we can get rich information to build the co-occurrence matrix. Once we can achieve this, we could then use the learned word embeddings to perform some sentiment analysis tasks [1].

We also compare the difference between the valence score of English comments and Chinese comments by using VADER models. As many NLP works suggest [6,7,8,9], we could try to extend our current model to a new one which can produce multi-dimensional scores for emotion given the comments especially from social medias. These multi-dimensional metrics could contain valence, arousal and dominance, regardless of whether these emotional dimensions are dependent or not. One possible research directions would be using the deep learning techniques. Since we will collect the large dataset for Covid-19 topic, it is possible to train a deep learning model with supervised learning. However, this approach is not very human-interpretable. Another possible research directions for the future work could be modify and extend the sentiment lexicons, the lexical features could not only have polarity labels and intensity values in valence scale, we can use the similar methods like inheriting the elements of existing well-established lexicons which contains information for multi-dimensional emotions. Or we could use the wisdom of the crowd approach to get the reasonable scores and labels by human subjects.

Acknowledgments

The author would like to thank Luvena Huo, Mehul Goel and Angela Zhang for valuable discussions and comments on the research project.

Reference

[1] Zhang, Lei, Shuai Wang, and Bing Liu. "Deep learning for sentiment analysis: A survey." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8.4 (2018): e1253.

[2] Huo, Luvena. "Cross-Cultural Differences in Responses to News Videos." (2021)

[3] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

[4] Hutto, Clayton, and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 8. No. 1. 2014.

[5] Sun, Junyi. "Jieba" (Chinese for "to stutter") Chinese text segmentation: built to be the best Python Chinese word segmentation

module. (2012) https://github.com/fxsjy/jieba

[6] Ekman, Paul. "Are there basic emotions?." (1992): 550. https://psycnet.apa.org/doiLanding?doi=10.1037%2F003-295X.99.3.550

[7] Preotiuc-Pietro, Daniel, et al. "Modelling valence and arousal in facebook posts." Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis. 2016.

[8] Russell, James A. "A circumplex model of affect." Journal of personality and social psychology 39.6 (1980): 1161.

[9] Xie, Housheng, et al. "A multi-dimensional relation model for dimensional sentiment analysis." Information Sciences 579 (2021): 832-844.

[10] Pang, Bo, and Lillian Lee. "Using very simple statistics for review search: An exploration." Coling 2008: Companion volume: Posters. 2008.

[11] Loper, Edward, and Steven Bird. "Nltk: The natural language toolkit." arXiv preprint cs/0205028 (2002).

[12] Pennebaker, James W., Martha E. Francis, and Roger J. Booth. "Linguistic inquiry and word count: LIWC 2001." Mahway: Lawrence Erlbaum Associates 71.2001 (2001): 2001.

[13] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.11 (2008).

[14] Ba, Jimmy. http://www.cs.toronto.edu/~jba/a1_data.tar.gz