Social Media Analytics (SMA) Text Mining Pipeline

Bushui Zhang (UID: 10407579), Yazhuo Cao (UID: 10329221),
Yecheng Chu (UID: 10319044), Zhaoyu Zhang (UID:10348535),
Zhengqian Jin (UID:10839527)

1 Background and Introduction

The Beijing 2022 Winter Olympics ended a few weeks ago. Along with the processes of contest is the generations of many popular social topics such as the newly arisen gold medal winner, the changing overall population feeling towards contests and the host location, Beijing and China. Such topics are intuitive to reflect the attention and sentiment from the global population to The Beijing 2022 Winter Olympics.Therefore it is an interesting topic to collect and analysis the global opinions using Social Media Analytics (SMA).

Supported by previous research (Fan & Gordon, 2014), SMA collects feelings and viewpoints from users who posts messages on the social media platforms and analyses the public opinions towards specific tags. In the paper, our group selects Twitter as the source platform for data extraction to answer three specific questions about The Beijing 2022 Winter Olympics by analysing outputs from the SMA pipeline. The three questions we purposed to analysed about the Olympics in this paper are:

1. What were the most talked about topics for the Winter Olympics?

2. What is the dominant sentiment towards the Winter Olympics?

3. Which 2 countries have been mostly talked about during Winter Olympics? Which one do people prefer?

All of them not only reflect where the public attention are towards the Olympics games but also the topics that are appropriate for the SMA pipeline to analyse in terms of Topic Detecting (Stieglitz et al., 2018).

Related tweets about the tag ***#WinterOlympics*** from Twitter were obtained and analysed by SMA pipeline, more detailed approach will be discussed in section 3.1. It should complete some of the following tasks such as Data Pre-processing, Topic Modelling, Sentiment Analysis and Named Entity Recognition for each specific question. All mentioned processes assist us to extract the hidden information about public opinions to answer the three analytic questions.

2 Related Works

Because of the publicly available nature of its text, Twitter has always been a popular source of data for opinion analysis, unlike Facebook, where many posts are hidden from strangers.

As research has shown, social media text has four distinctive features: time sensitivity, short length, unstructured phrases, and abundant information. This brings new challenges and therefore requires new methods for analysis as the classic methods such as Bag-of-Words have limited performance due to the short and unstructured text (Hu & Liu, 2012).

As a result, a variety of methods for Sentiment Analysis in social media have been proposed. Some are lexicon-based, some use statistical-based machine learning methods, others use both.

Both methods have been shown to have their advantages and disadvantages. Lexicon based methods require no training, however the performance is heavily dependent on the quality of the lexicons given. This type of method is flexible and can easily be integrated between different languages. Machine learning methods, on the other hand, require training data and rely on their quality. It performs poorly for misspelt or slang words.

In the past, people have applied recurrent neural networks (RNN) methods (Arras et al., 2017) built using Keras and Tensorflow to predict the emotions of the text by focusing on the relationships between words that have similar performance compared to third party analyser TextBlob (Nemes & Kiss, 2021).

There has also been experimentation with rule-based models for Sentiment Analysis on social media text. The Valence Aware Dictionary for Sentiment Reasoning(VADER) (Hutto & Gilbert, 2014) model was published in 2014 and starts with a gold standard sentiment lexicon validated by humans (Surowiecki, 2005) and makes use of the five rules for sentiment intensity. It was shown to outperform individual human raters and equally matched machine learning algorithms (Hutto & Gilbert, 2014) .

Topic modelling is a way to identify patterns in a dataset, in this case, tweets that include the tag ***#WinterOlympics***. There are existing tools like MALLET and modelling techniques have been proposed using a probabilistic modelling framework based on LDA called the joint sentiment/topic model (Lin & He, 2009).

3 Methodologies

After discussions among among our team and as mentioned above, we decided to use Twitter as the social media platform for our research. In this chapter the SMA research process is described.

3.1 Data Collection

As we aim to analysis the public opinion on the Winter Olympics, we wanted to collect all the tweets on Twitter that tagged *Winter Olympics*.

Since the duration of the Winter Olympics was 19 days from February 4th to 20th, but the Twitter external API can only retrieve tweets within the most recent week, we decided not to use the external API to collect data, but use *snscrape*, a social networking service scraper in Python, to obtain the data needed. The specific commands are as follows:

snscrape twitter-search "#WinterOlympics since:2022-02-04 until:2021-02-20 lang:en" > scraped_tweets.txt

The data obtained by this command is a list of URLs in a txt file. We generate twitter access tokens from a newly registered twitter developer account and use the function in *Tweepy* to convert the URLs into tweets and stored them in a Comma Separated Values (CSV) file.

In the end, we fetched around 50,000 tweets.

3.2 Data Cleaning

In consideration of the diversity of language expressions, all the tweets fetched needs to be standardised to facilitate the subsequent SMA process by either modifying or reducing the symbols and sentences that are meaningless to expressing the meaning of sentences.

The following data cleaning steps are performed on the tweets fetched:

- Remove Hyperlinks
- Remove punctuation
- Remove unnecessary line breaks
- Convert the titles to lowercase
- Remove emojis

The cleaned tweets were stored in a different CSV file, which contains 4 properties of a tweet: *Create Time*, *User*, *Location*, *Tweet Text*.

3.3 Word Cloud

A word cloud is a visualisation of different words in random order and different size. The higher word appearance frequency, the larger the font of the word in the cloud. The cleaned tweets are processed and the overall word frequency are recorded to produce a word cloud as shown in Figure 1

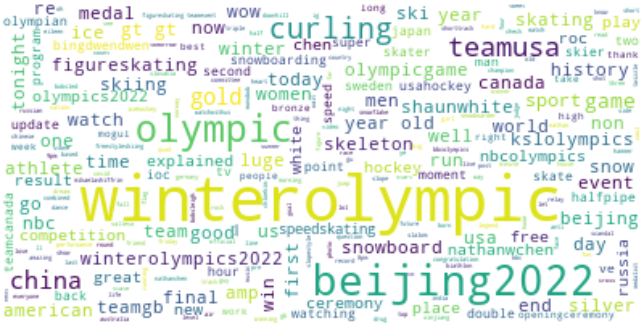

Figure 1: Word Cloud From Tweets

3.4 Topic Modelling

In machine learning and natural language processing, a topic model is a statistical model used to discover abstract "topics" that appear in collections of documents. It is a common text mining tool often used to discover hidden semantic structures in the text body. In SMA, this tool can be used to discover the hidden information in the corpus. Topic modeling has been used to study full-text articles (Lamba & Margam, 2019)(Lei & Chen, 2021). Our topic modeling can be briefly divided into the following steps:(1)Sentence tokenization; (2) Stopword removal; (3) Corpus words to Dictionary; (4) Dictionary to BoW representation; (5) LDA model creation base on custom number of topics: (6) Model visualization using pyLDAvis third-party library; (7) By viewing the generated HTML, you can get the most frequently occurring and most relevant words under each topic, from which you can get the relevant vocabulary modelling for each topic.

For diagrams on topic modeling, please refer to Figure 2,3,9 and 10 below.

3.5 Sentiment Analysis

We are particularly interested in the polarity of tweets about the tag *Winter Olympics*. To achieve this, we used third library NLTK's sentiment analyser to perform polarity researches against every unique tweets. By dividing them into 19 parts according to the create date of each tweet, we collect the *positive* and *negative* sentiment and *compound* scores of the tweets to plot graphs. We represent the compound scores along with the proceeding of

Winter Olympics from opening to ending, and compared their overall percentages in a pie chart. (See Figure 5)

3.6 Named Entity Recognition

We classified the key words in all tweets into different categories such as event, person, place and so on, based on their corresponding labels inside a NER entity.

Through the NER (Named Entity Recognition) package *en_core_web_trf* in *spacy*, we analyzed the country names involved, and stored the number of occurrences of each location name in a dictionary. After sorting the counts of the dictionary, we can get the country with the most occurrence. (See Figure 6)

4 Analysis

4.1 Task 1

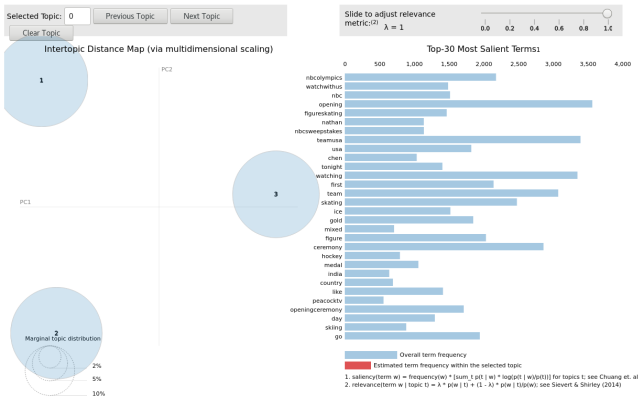

Figure 2: Topic modelling result for the three topics

From the intertopic distance map via multidimensional scaling, the relationships between these topics can be investigated. The map has three circles, each represent a topic. The size of a circle can demonstrate the prevalence of a topic. As shown in Figure 2, Topic 1 makes up the biggest portion of the topics being discussed among the tweets collected, constituting 35.2% of all the tokens. There is only a small difference in size between Topic 1 and the other two, hence, the prevalence of these three topics is similar. Besides, the distances between these three circles differ greatly. It can be deduced that each topic is not quite related to another.

From Figure 3, the most relevant term for Topic 1 is "watching". Some events from the winter Olympics are depicted, such as curling, skating, and skiing. Therefore, the first topic could be summarised as watching events at the winter Olympics. Likewise, the remaining two topics could be described separately as the opening ceremony of the winter Olympics for Topic 2 and the news about Team USA by the NBC Olympic news channel for Topic 3. The topic modeling results for Topic 2 and Topic 3 are specified in the Appendix A.

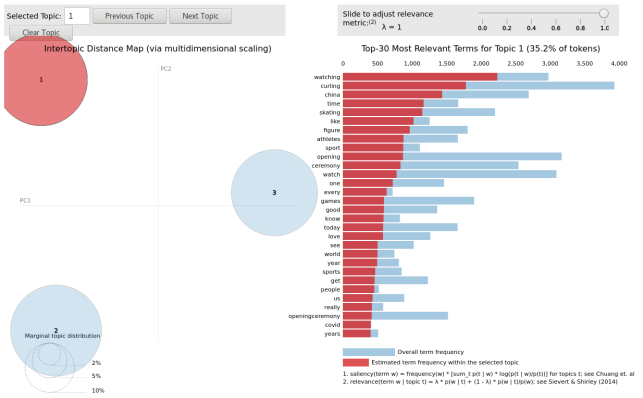

Figure 3: Topic modelling result for topic 1

4.2 Task 2

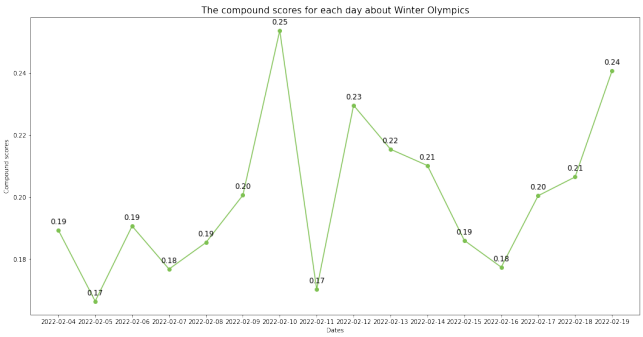

Figure 4: The compound score for each day about Winter Olympics

Figure 4 shows the variation in the daily-averaged compound score, which could reflect people's semantic orientation over the 16-day period from February 4^{th} to February 19^{th}, 2022. That is calculated by dividing the sum of the compound score for each text by the total number of texts for this day. From this graph, it can be seen that the daily-averaged compound score fluctuates all the time without any obvious trend in this period. Specifically,the compound score reaches its maximum of 0.25 on February 10^{th} and drops immediately to 0.17 the next day, which is one of two minimums. Even though there are two days that the compound score falls to 0.17, it is still above 0.05, which can be considered as positive. In summary, the dominant sentiment towards the Winter Olympics is positive from the perspective of the time as the compound score is positive for each day.

This pie chart in Figure 5 represents people's sentiments as positive, negative, or neutral towards the Winter Olympics. Almost half of the Twitter text is positive about this event, with only about 15% of the text being negative. There is approximately three times as much positive information as negative ones. As a result, it can be concluded that the dominant sentiment towards the Winter Olympics is positive based on the quantity of messages.

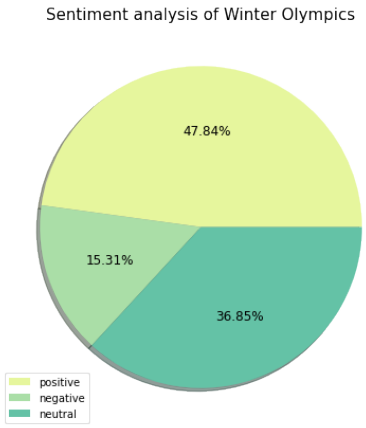

Figure 5: Sentiment analysis of Winter Olympics

All in all, the dominant sentiment towards the Winter Olympics is positive as a result can be obtained from the viewpoints of both time and quantity.

4.3 Task 3

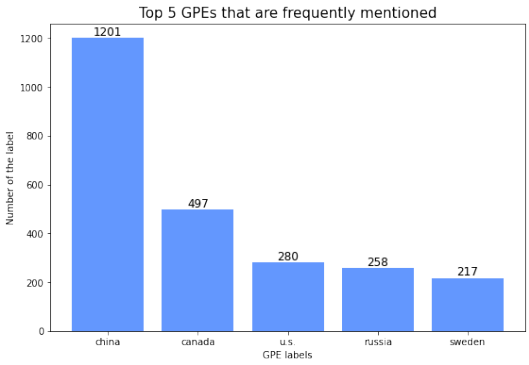

Figure 6: Top 5 GPEs that are frequently mentioned

Figure 6 illustrates the 5 most popular GPE(Geopolitical Entity) labels that are observed for NER. It is clearly shown that China is the most frequently mentioned country, over double of the frequency of the seconded mentioned country, Canada. After finding the 2 countries, Sentiment Analysis was performed on them separately using a similar technique to that in Section 4.2. The resulting daily-averaged compound scores for both of them are plotted on the same graph for comparison, as demonstrated in Figure 7 .

Despite the fluctuation, the compound score for China generally grows throughout the Winter Olympics period. It peaks on February 19^{th}, when the event ends. The only exception occurred on the 13^{th}, when the compound score plunged to 0.05, indicating a change in general attitude from positive to neutral. In contrast to the compound score for China, that for Canada first shows a decline. After that, it edges down from 0.22 to 0.14 before reaching a plateau at about 0.31. After experiencing another shrink, the compound score for Canada gradually rises until reaching a summit at 0.4. Generally speaking, although

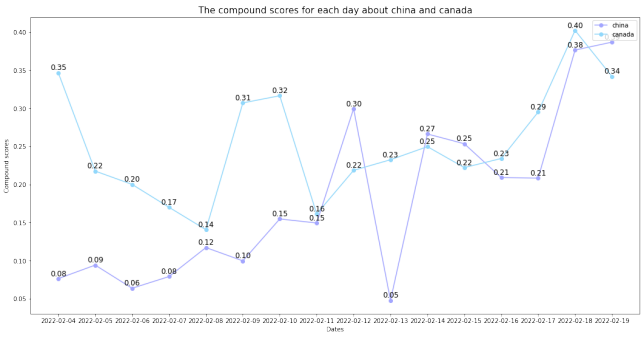

Figure 7: The compound scores for each day about China and Canada

the compound score shows that people have a positive attitude towards both Canada and China, the compound score for Canada surpasses that of China on most of the days, illustrating a more positive attitude towards Canada. The same conclusion can be derived from Figure 8, where a comparison between general sentiment labels is plotted. According to the chart, when talking about China, 43.47% of the comments are positive, while this figure is 51.92 for Canada. In terms of negative comments, Canada merely has 11.82%. In contrast, China has 21.35% negative comments, almost twice that of Canada.

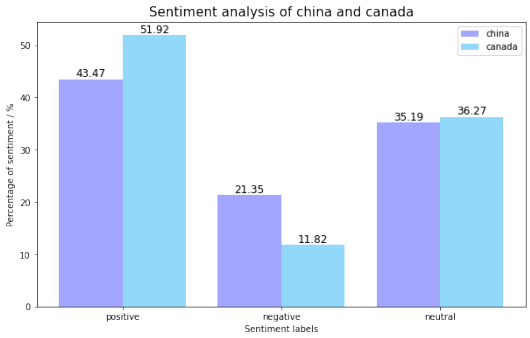

Figure 8: Sentiment analysis of China and Canada

In general, according to our experiment using NER and Sentiment Analysis, the two most talked about countries are found to be China and Canada. Among them, people show a more positive attitude towards Canada.

5 Conclusion

From the results shown above, it can be concluded that users mostly express their opinions about the Winter Olympics events they watched. The opening ceremony is more talked about than any sports event, and for those users who are focusing on news about Team USA at the Olympics, they mostly watched the broadcast from NBC.

The overall sentiment towards the Olympics this year is mostly positive, as shown in Figure 5 and out of the two most talked about countries, China and Canada, Twitter users have a more positive feeling towards Canada.

References

Arras, L., Montavon, G., Müller, K.-R., & Samek, W. (2017). Explaining recurrent neural network predictions in sentiment analysis. *arXiv preprint arXiv:1706.07206*.

Fan, W., & Gordon, M. (2014). The power of social media analytics. *Communications of the ACM*, *57*, 74–81.

Hu, X., & Liu, H. (2012). *Text Analytics in Social Media*, (pp. 385–414). Boston, MA: Springer US.
URL https://doi.org/10.1007/978-1-4614-3223-4_12

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216–225.
URL https://ojs.aaai.org/index.php/ICWSM/article/view/14550

Lamba, M., & Margam, M. (2019). Mapping of topics in desidoc journal of library and information technology, india: a study. *Scientometrics*, *120*, 477–505.

Lei, H., & Chen, Y. (2021). *Exclusive Topic Modeling*.

Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, (pp. 375–384).

Nemes, L., & Kiss, A. (2021). Social media sentiment analysis based on covid-19. *Journal of Information and Telecommunication*, *5*(1), 1–15.
URL https://doi.org/10.1080/24751839.2020.1790793

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*, 156–168.

Surowiecki, J. (2005). *The Wisdom of Crowds*. Knopf Doubleday Publishing Group.
URL https://books.google.co.uk/books?id=hHUsHOHqVzEC

A Appendix:

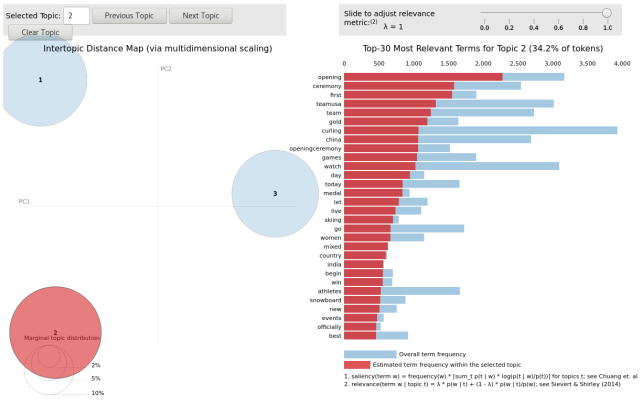

Figure 9: Topic modelling result for topic 2

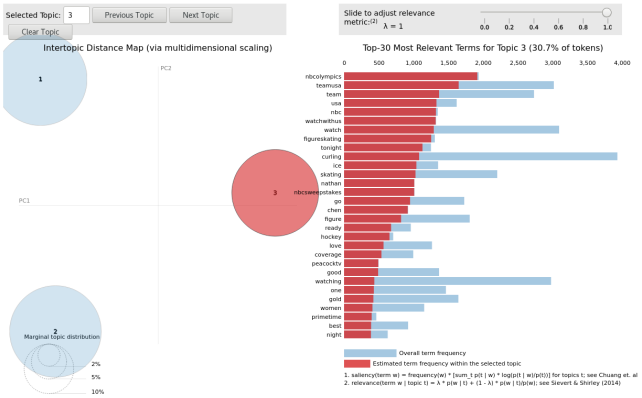

Figure 10: Topic modelling result for topic 3