# Semi-supervised Analysis of Media Attitudes toward Geopolitical Entities at the End of 2022

**Hantao Hong**
University of Michigan
Ann Arbor, Michigan
hhantao@umich.edu

**Jacky He**
University of Michigan
Ann Arbor, Michigan
plianghe@umich.edu

## 1 Problem Description

### 1.1 Introduction

The goal of this project is to use Natural Language Processing to gain insight into 12 mainstream media's attitudes towards 5 geopolitical entities (USA, China, Taiwan, Russia, and Ukraine) at the end of 2022. We analyzed a large corpus of tweets from official accounts of the media on Twitter to determine the overall sentiment towards these entities. A fine-tuned Transformer model(BERT) was developed and evaluated for sentiment analysis. By comparing BERT to the baseline Naive Bayes Model, we found a better model to do the project on how the media perceives geopolitical entities(GPEs). The results of this project have the potential to provide valuable insights for researchers, journalists, and policymakers alike.

### 1.2 Input & Output

To analyze the attitudes of media toward GPEs, we first trained BERT and Naive Bayes model based on a sampled and cleaned annotated dataset whose topic is related to politics. After training, validating and testing these two models, we evaluated and chose the better model to do the media's sentiment analysis.
**Input**: English tweets related to 5 GPEs from the media Twitter accounts during December 2022
**Output**: The sentiment score and respective graphs of 12 media's attitudes toward the 5 GPEs, ranging from 0 to 1.

### 1.3 Challenges

- Initially, we aim to construct a semi-supervised model. The rigorous definition of semi-supervised is supervised learning on a relatively small labeled dataset, and unsupervised learning on a relatively large unlabeled dataset. However, the main streamed media are partially likely to concentrate on these 5 GPEs during a small period of time. As a result, the data we could collect is limited. This indicates that the dataset for supervised learning will be larger than that for unsupervised learning. So a challenge is that we could not collect enough data for unsupervised learning.
- Based on the characteristic of sentiment itself, it is hard to set up a standard quantitative scale for measuring the extremeness of a sentiment from a sentence. This is even a difficult task for human beings, which is another challenge for this project.

### 1.4 Contributions

Jacky He mainly focused on dataset selection, web crawling, and the respective data cleaning. Hantao Hong mainly focused on constructing the models and selecting the best one then training and testing. From the workload perspective, the contributions from these two authors are even.

## 2 Related work

Sentiment analysis or polarity detection between entity-to-entity (Park et al., 2021) aims at deciding the positive or negative attitude between several entities within a sentence or quantifying the degree of sentiments embedded in a text (Gilbert and Hutto, 2014). Transformers models, like BERT, can contribute positively to the performance (Devlin et al., 2019). However, there is no such research studying the different media's attitudes toward different GPEs, especially in December 2022. Here is the new thing in our project.

## 3 Methodology

### 3.1 Annotated Datasets

We found a dataset called **Sentiment 140** intended for the training, validation, and test sets of two models with the objective of pinpointing the better model. This dataset is also designed for training

purposes in preparation for the final sentiment analysis, which will be performed on the crawled data. There are 1.6 million annotated tweet data with sentiments in Sentiment 140. To make the training set more related to our topic, which is about media's attitudes towards GPEs, we adopt Linguistic Inquiry and Word Count(LIWC) to filter the data when its 'Politics' score is over 0, which means this tweet is related to Politics. After filtering, we get **5246** for the negative attitude data, and **6725** positive attitude data. One example of a negative tweet is "Trump launches US-China trade war." One example of a positive tweet is "Ukraine is a good friend of the USA." It is suitable for our task because the topics are related to Politics, the type of the text is tweet and it is a well-annotated dataset.

In order to get a balanced dataset, we randomly selected **5000** from **5246** for the negative attitude data, and **5000** from **6725** positive attitude data with the **pandas.DataFrame.sample** method. After all the work is done, we will get an annotated dataset with size **10,000** and **5000** annotated as positive and **5000** annotated as negative.

### 3.2 Web Crawling Datasets

We employed web crawling techniques to collect data from Twitter using the Twitter API. Our date range of interest spanned from December 1, 2022, 00:00:00 EST to January 1, 2023, 00:00:00 EST. Subsequently, we compiled a list of 12 media outlets, as we aimed to investigate their respective attitudes toward 5 geopolitical entities, including the United States, China, Taiwan, Russia, and Ukraine.

To retrieve relevant tweets, we constructed queries such as 'US OR "United States" OR USA from:CNN since:' + start date + ' until:' + end date, which allowed us to gather tweets containing the terms "US," "United States," or "USA" during December 2022. On average, we obtained 5.23 related tweets per day from a single media outlet's Twitter account. Then, we implemented the same data-cleaning methodology used for the annotated datasets on the newly crawled data.

### 3.3 Data Cleaning

For this part of the work, we mainly aim at the data we web crawled from Twitter. Thanks to the contribution of the web crawling API for Twitter, this part of work is relatively straightforward. We remove the "RE" which stands for "retweet" and remove the hyperlinks in the tweets that the media set for their reports on their own websites. For the Naive Bayes model, we also remove the stopwords.

### 3.4 BERT

According to the satisfactory performance of BERT on Natural Language Processing, we decide to select it as a candidate for our project.

- Neural model: Transformer
- Number of layers: 12
- Loss function: CrossEntropyLoss
- Optimizer: AdamW
- Parameters: 110 Million

## 4 Experiments

### 4.1 Data Preparation

To split the dataset into training, validation, and test sets, we apply **sklearn.model_selection.train_test_split** method. First, for the **5000** annotated as positive, we split it into **4500** and **500**. And the same operation will be on the dataset annotated as negative. Then, we concatenate the **4500** positive data and **4500** negative data to get the **9000** training-validation set. We concatenate the **500** positive data and **500** negative data to get the **1000** test set. Furthermore, we adopted 5-fold validation and split the training-validation set into a training set with **7200** data and a validation set with **1800** data. (Figure 1)


Figure 1: Data Preparation for Training/Validation/Test

### 4.2 Model Selection

In order to find the best model to perform the task, we will select between two models. They are the Naive Bayes model and BERT.(Figure 2)

### 4.3 Classify on the Selected Model

We will train the selected model with the annotated data after cleaning. We will make predictions for the sentiments of the tweets we web crawled. And we can classify the sentiment as "POSITIVE" and "NEGATIVE" from the tweet. For each media

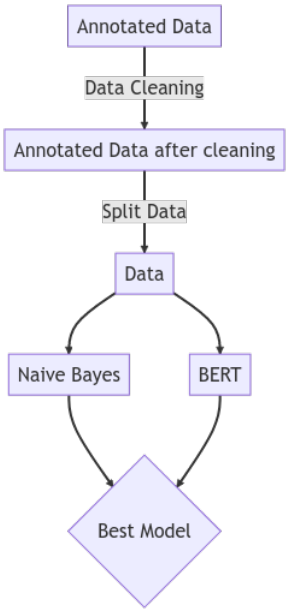


Figure 2: Model Selection

outlet and geopolitical entity pair, we will calculate a score that signifies the extent of positive and negative sentiment expressed in the data (Figure 3).

### 4.4 Baseline – Naive Bayes & TF-IDF

We regard Naive Bayes Classifier as our baseline. Since there is only one hyperparameter for Naive Bayes classifier, after hyperparameter search, we will adopt the hyperparameter for baseline approach as **alpha = 1.6**(Figure 4). we achieve an accuracy rate of **71.72%** and AUROC(Area Under ROC curve) of **0.7948** on the validation set(Figure 5). This value is the baseline performance and will be used to evaluate the performance of our fine-tuned BERT model.

### 4.5 Fine-tune BERT

To fine-tune our Bert Classifier, we need to create an optimizer. We optimized the model based on hyperparameters:

- Learning Rate(AdamW): [1e-5,1e-4,1e-3]
- Batch Size: [16,32,64]

### 4.6 Hyperparameter Search

During training, we adopted cross-validation on the 5-fold validation set. After hyperparameters search within 5 epochs, we find the best hyperparameters as below

- Learning Rate(AdamW): 1e-5

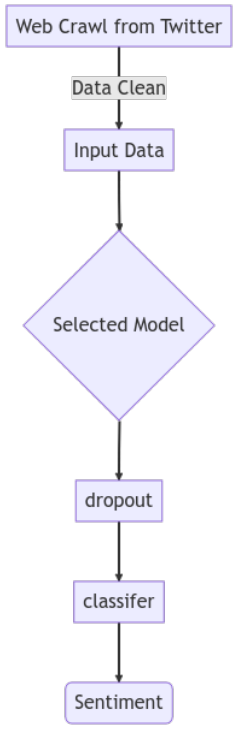


Figure 3: Classify on the Selected Model

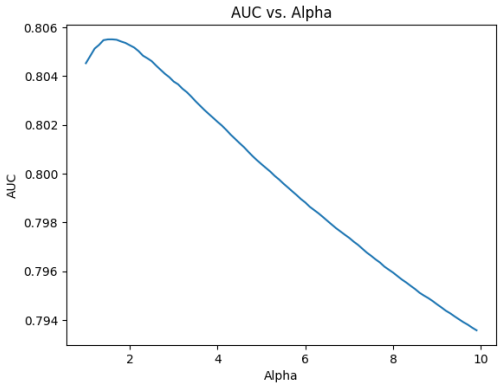


Figure 4: AUROC vs. Alpha

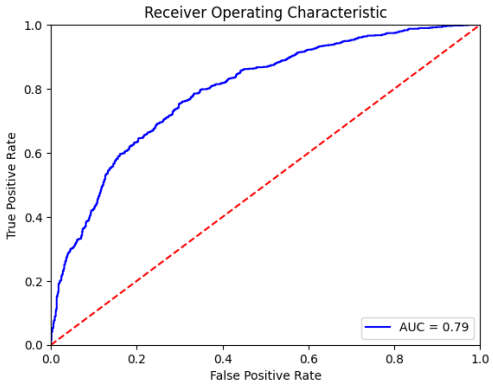


Figure 5: AUROC plot for Naive Bayes & TF-IDF on Validation Set

- Batch Size: 32

We calculated the loss and the accuracy of the validation set. We will still adopt Accuracy and AUROC to evaluate the performance of our model on test set.

### 4.7 Training

We will train the model on GPU provided by Google CoLab. The following is the process of the training:

- Zero out gradients calculated in the previous pass
- Perform a forward pass to compute loss
- Perform a backward pass to compute gradients
- Update the model's parameters

### 4.8 Evaluation Metrics on Validation set

Our current evaluation Metrics is listed below:

- Accuracy: 80.61%
- AUC: 0.8798

(Figure 6) There are pro and con for the evaluation metrics. Pro: AUROC presents the probability that two randomly-selected samples are correctly ranked, which shows it is a measure of correct ordering of the classification. Con: The imbalance dataset will have effect on the performance of AUROC. However, the dataset we choose here is balanced. So we can ignore this con.

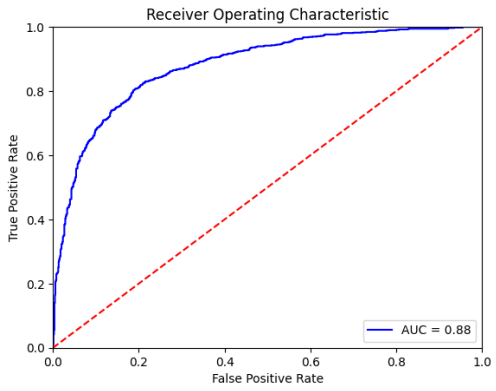


Figure 6: AUROC plot for BERT on Validation Set

### 4.9 Performance on Test Set

After training, we test the model on the test. As we mentioned above, the test set includes **500** data labeled as positive. So we adopt **Softmax** on the output from BERT to get the probability. Then we compare the probabilities from Softmax with the threshold. If the probability is greater than the threshold, the model will classify the tweet as positive. It should be noted that the sentiment classification task here is even difficult for human, so we set the threshold as **0.6** instead of the default as 0.5, giving us safe predictions on the test set.

- Classified as Non-Negative out of 500: 466
- Accuracy: 0.932

### 4.10 Predictions on Unlabeled Dataset

We get the following matrix(Figure 7) when we predict on the dataset from web crawling after cleaning. It should be noted that we still set the threshold as **0.6** for the same reason above.

| | USA | China | Taiwan | Russia | Ukraine |
|---|---|---|---|---|---|
| CNN | 0.248276 | 0.142857 | 0.000000 | 0.302326 | 0.300000 |
| nytimes | 0.741935 | 0.150000 | 0.000000 | 0.276596 | 0.281690 |
| BBCWorld | 0.205128 | 0.162791 | 0.000000 | 0.122449 | 0.244186 |
| TheEconomist | 0.409091 | 0.306011 | 0.272727 | 0.204301 | 0.346734 |
| Reuters | 0.370968 | 0.176282 | 0.240000 | 0.218543 | 0.196203 |
| WSJ | 0.900000 | 0.148148 | 0.000000 | 0.211268 | 0.102273 |
| CGTNOfficial | 0.543210 | 0.680636 | 0.500000 | 0.215190 | 0.240000 |
| XHNews | 0.527778 | 0.741379 | 0.666667 | 0.433333 | 0.187500 |
| globaltimesnews | 0.274678 | 0.528465 | 0.295082 | 0.576923 | 0.272727 |
| SCMPNews | 0.226829 | 0.264596 | 0.156250 | 0.250000 | 0.223301 |
| ChinaDaily | 0.323171 | 0.609524 | 0.520000 | 0.571429 | 0.500000 |
| PDChina | 0.220000 | 0.795359 | 0.500000 | 0.588235 | 0.200000 |

Figure 7: Results

### 4.11 Results Visualization

The following plots are the visualization of the data we get.

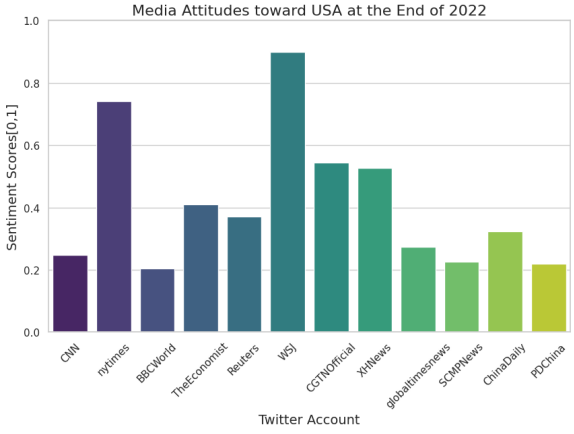


Figure 8: Media Attitudes Toward USA at the End of 2022

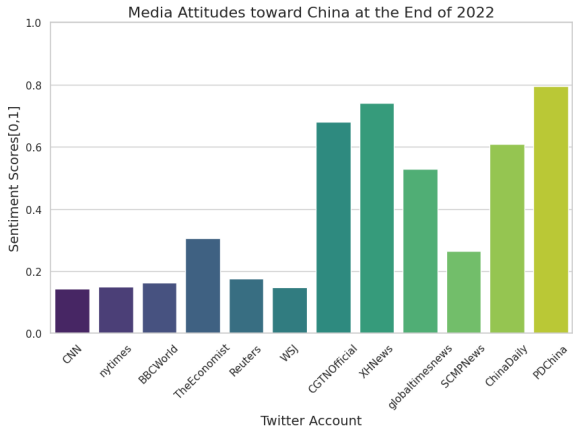


Figure 9: Media Attitudes Toward China at the End of 2022

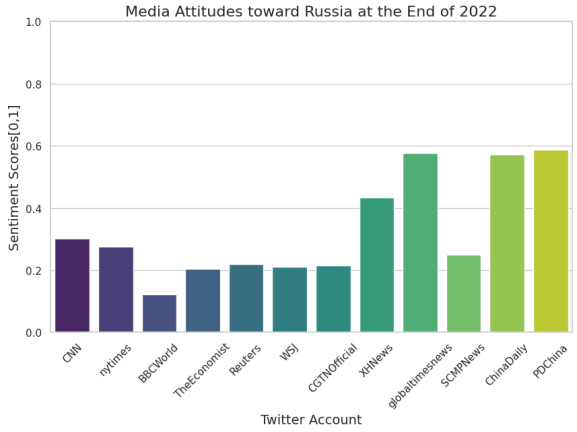


Figure 10: Media Attitudes Toward Russia at the End of 2022

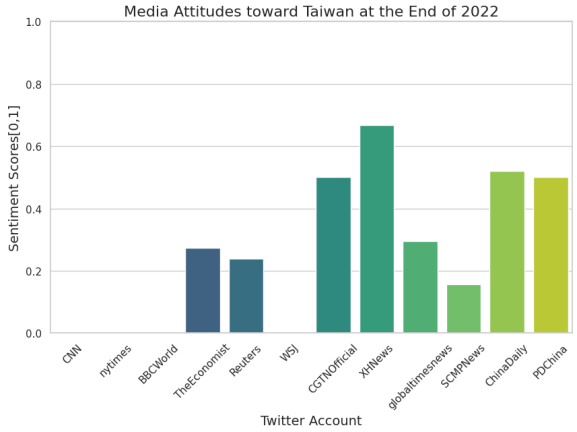


Figure 11: Media Attitudes Toward Taiwan at the End of 2022

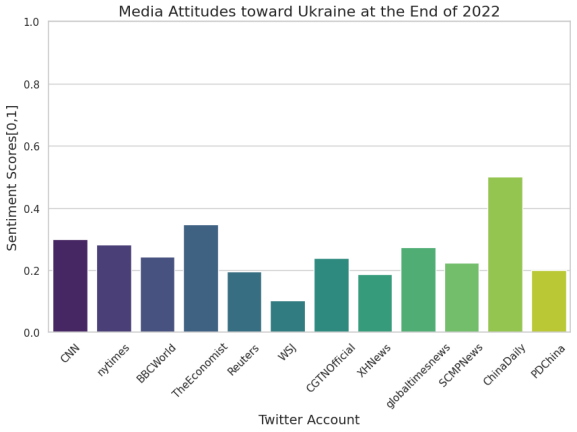


Figure 12: Media Attitudes Toward Ukraine at the End of 2022

### 4.12 Discussion

There are several points that should be noted.

- To verify the accuracy of the sentiment scores generated from the web-crawled data. We manually annotated 100 web-crawled tweets for each person. For the result, 90% of the sentiment labels we classified manually is the same as the sentiment analysis generated by the BERT model.
- We can see that there are some bars in the histogram missing, e.g. the plot for Taiwan. The reason behind this is that the media such as CNN, New York Times and BBCWorld didn't post any tweets about Taiwan during that period of time.
- Just as we mentioned before, according to the rigorous definition of the "Semi-supervised" model, the size of the labeled dataset should be much smaller than that of the unlabeled dataset. But due to the feature of the tweets we choose, the mainstream media is hardly possible only focus on these 5 GPEs during that specific period of time.

As discussed above, we should notice that the target of our project itself is to analyze the media's attitudes towards 5 specific Geopolitical Entities during a very specific period of time(December 2022). And the target itself, analyzing the attitudes will be a kind of analyzing the sentiments, it will be sometimes ambiguous. In future work, if more data could be collected and with more time, we would try BERTweet, which is more specific to Twitter. In addition, we would try some other different models

such as T5 and Recurrent Neural Networks. Then, we would do a comparison of different models and find the best one to do the sentiment tasks.

## References


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.

Eric Gilbert and Clayton J Hutto. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, pages 216–225.

Kunwoo Park, Zhufeng Pan, and Jungseock Joo. 2021. Who blames or endorses whom? entity-to-entity directed sentiment extraction in news text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4091–4102, Online. Association for Computational Linguistics.