Twitter Economy Sentiment Analysis

Chen Wei, Chenxi Wu, Minfeixue Zong, Zhenhao Sun

INTRODUCTION

Sentiment analysis prediction is one of the most common NLP task which involves analyzing a piece of text to determine the underlying sentiment or emotional tone. This can be done using machine learning and other computational techniques to classify text as positive, negative, or neutral.

For this project, we are trying to study the relationship between the bank collapse and people's attitude towards the collapsed/being bought bank (SVB, Credit Sussie) or World economy.

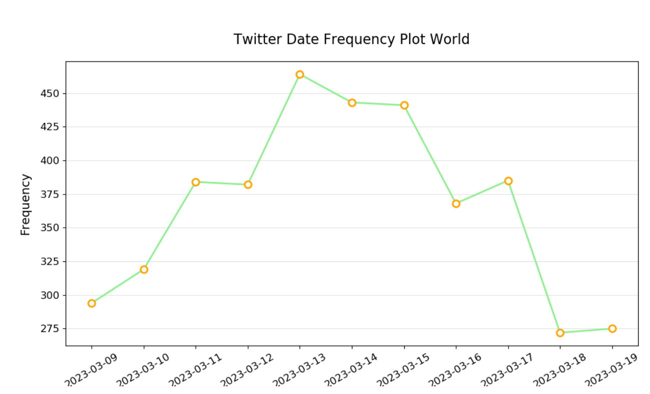
DATASET

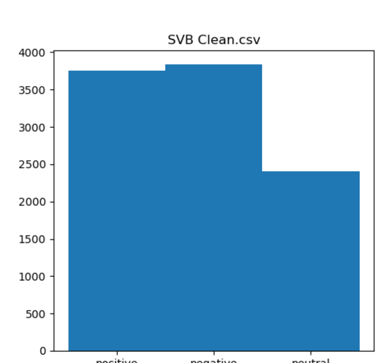
We scraped data from twitter which contains Tweet_id, User, Content, Like_Count, Quote_Count, Reply_Count, Url, Date. After carefully handling sensitive information like User, we want to perform the sentiment analysis task and hypothesis testing on the Content of the twitter. We have 3 datasets, which correspond separately to the twitter world_economy data, twitter credit_sussie data, and twitter SVB data.

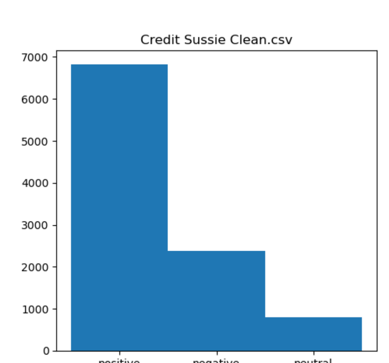
Due to the scraping technical constraint, the world_economy data ranges from the date 2023-03-09 to 2023-03-19 and the credit suisse (2023.3.18 - 2023.3.19) and SVB (2023.3.9 - 3.10) data is just the twitter at the time of the collapse. The visualization is the same as the hypothesis testing part.

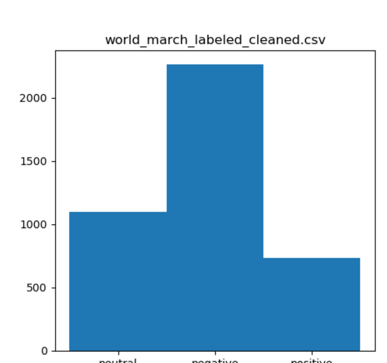
We manually labeled the world_economy sentiment (and only used this dataset) for our model training. We ask 3 of us to label the data and one to organize our votes, use the mode as the overall sentiment and manually deal with the outliers (i.e. some illegal values, the situation where 3 of us all think differently and there is not really a mode). This means if 2 out of 3 think that a twitter is negative, its overall sentiment label is negative. For negative, positive, and neutral sentiments, we use 0, 1, and 2 respectively to represent them in our data.

Data Distribution









Preprocessing

1. Lowercasing
Converting all text to lowercase to ensure consistency.

2. Tokenization
Splitting the text into individual words or tokens

3. Stopwords, punctuation removal
Removing common stopwords (e.g. "the", "and", "is") and punctuation (",", ".") that do not provide meaningful information about the content.

4. Drop NA
Remove the invalid value.

Hypothesis Testing

1. Before collapse, 50% people have negative attitude towards SVB. (z-statistics)

We use a one-sample z-test for proportions. This test will help us determine if the proportion of negative attitudes in our sample. The statistical analysis revealed that less than 50% of sampled tweets expressed a negative attitude towards SVB, as indicated by a p-value of 1.00. However, strong evidence was found (p-value of 0.000) to support the claim that more than 30% of tweets in the sample displayed a negative attitude towards SVB.

2. Before collapse or being bought, people's attitude towards SVB and Sussie Credit are identical.

The Chi-Square test was used to check if there was a significant difference in sentiment labels distribution between two groups. The resulting p-value was 0.96, which is greater than 0.05. We fail to reject the null hypothesis. Therefore, it is concluded that there's no significant difference in the sentiment labels distribution between SVB and Sussie Credit.

3. People have the same attitudes towards the world economy before and after the collapse.

Two-sample t-tests were utilized to compare sentiment scores towards the world economy before and after a collapse. The null hypothesis suggested no average difference in these scores. However, the derived p-value was 0.04, less than the significance level of 0.05. We reject of the null hypothesis. The attitudes towards the world economy were statistically different before and after the collapse.

Machine Learning

1. LDA

Why we used LDA?
- Discover the hidden thematic structure
- Suitable for analyzing and summarizing large text corpora

Evaluation Metric:
- Topic coherence
- Interpretability

Challenges
- Interpretability
- Choosing the hyperparameters
- Model evaluation

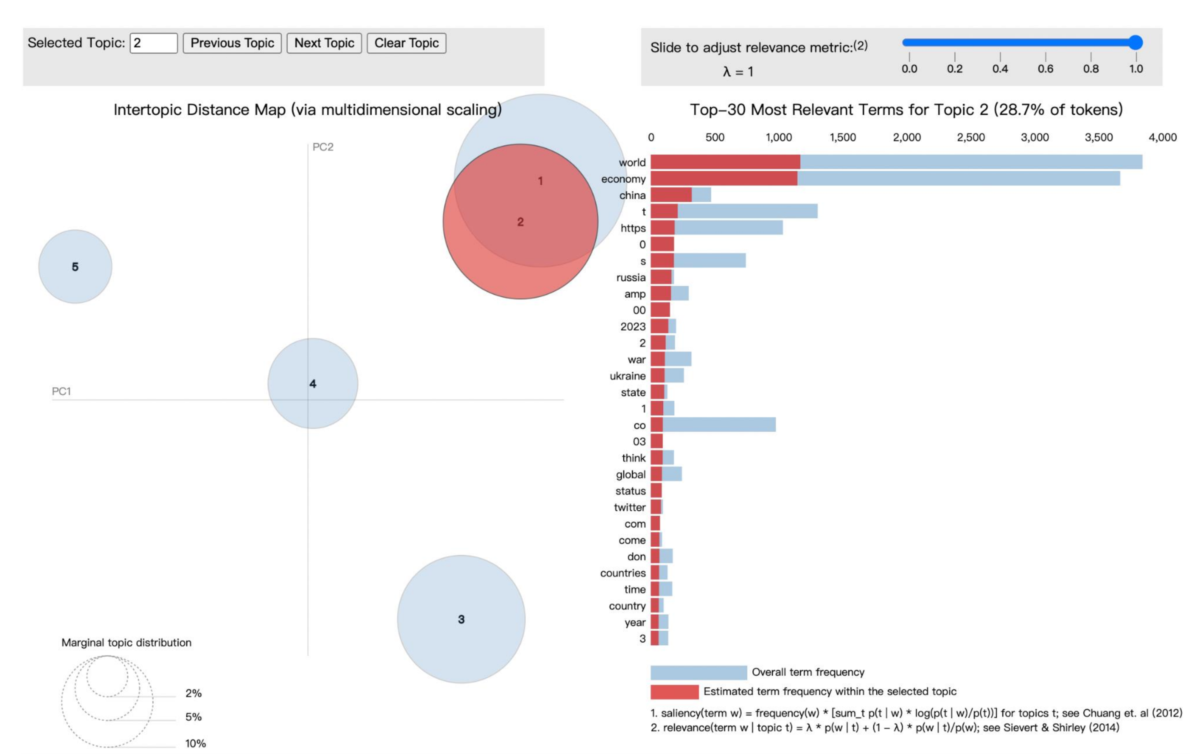
Other methods we considered:

Non-negative Matrix Factorization (NMF)

Latent Semantic Analysis (LSA)

Our preprocessing steps include:
- Lowercasing.
- Tokenization
- Stopword removal
- Lemmatization

Results(Top topics we found):
- how the war affects the world economy
- focuses more on international relationships and the world economy.
- monetary policies and the world economy.



2. Mulnitomial Naïve Bayes
Why Multinomial Naive Bayes?
- Simple and effective probabilistic classifier
- Suitable for classification

Data reconstruction
- Remove special characters
- Balance the dataset using SMOTE
- Split the dataset into training and testing

Classification Report:

	precision	recall	f1-score	support
0	0.80	0.53	0.64	471
1	0.75	0.90	0.82	430
2	0.74	0.85	0.79	455
accuracy			0.76	1356
macro avg	0.76	0.76	0.75	1356
weighted avg	0.77	0.76	0.75	1356

Result:
The model achieved moderate success in predicting the sentiment of the tweets. With an accuracy of around 75%

	precision	recall	f1-score	support
negative	0.73	0.87	0.79	386
neutral	0.75	0.20	0.32	103
positive	0.73	0.73	0.73	320
accuracy			0.73	809
macro avg	0.74	0.60	0.62	809
weighted avg	0.73	0.73	0.71	809

3. SVM
Why SVM?
SVM is effective in finding the optimal hyperplane that separates the different sentiment classes

Data cleaning:
- Lowercase the text
- Remove stop words and punctuation
- Lemmatize the word

Result: The SVM model achieved an accuracy of around 0.72 for sentiment classification.

DISCUSSION

In conclusion, our statistical analyses reveal three key findings:

1. Less than 50% of sampled tweets displayed a negative attitude towards SVB before it collapsed. However, there's strong evidence that more than 30% held a negative sentiment before it collapsed.

2. Before they collapsed, the sentiment towards SVB and Sussie Credit was indistinguishable.

3. A significant shift in public sentiment towards the world economy was found before and after the banks collapse.

Our machine learning model:

1. The LDA model effectively highlighted key themes in global economic discussions, including war effects, international relations, monetary policies, and varying economic viewpoints.

2. The Multinomial Naive Bayes model achieved a decent 76% accuracy in predicting tweet sentiments, performing best with positive sentiments and struggling with negative ones.

3. The Support Vector Machine model showed strength in identifying negative and positive sentiments with 72% accuracy but had notable difficulty recognizing neutral sentiments.

ACKNOWLEDGEMENTS

Throughout the implementation of this project we have received a great deal of support and assistance.

We would like to thank our mentor TA, whose expertise was invaluable in formulating the research questions and methodology. His insightful feedback pushed us to sharpen my thinking and brought our work to a higher level.

We would like to thank each other in our group, for our wonderful collaboration and support.

Last, we would like to thank our professor, who offers such an amazing course and we really learned a lot,