Data Analysis on New articles

Kan, Jie-Sheng

Chiu, Lok Ying

Li, Chia-Lo

HKUST

Research Objectives

This study aims to analyze financial, economic, or reputational risk in Hong Kong. By examining a large dataset of news articles and social media posts, the study aims to identify patterns and sentiment across articles that indicate potential risks. A topic modeling model will be designed to extract keywords and visualize them on a time series graph, enabling the identification of hot topics. Sentiment analysis will also be conducted to understand public or market sentiment over time. Overall, the research aims to enhance understanding of risks in Hong Kong and provide insights for risk mitigation.

Data Description

Data distribution

After removing the duplicates, theThe dataset comprised of ¾ of Chinese news and ¼ of English news. The news in this dataset is posted between 2021-10-01 and 2023-09-30 23:58:00. In each month, there are around 4000 – 5000 news recorded, but March 2023 has extra number of records, which is about 7000. There are four news sources, C(Social), K(Blog), L(Forum), and U(Web), but most of them are from U(Web).

Data Cleaning

We cleaned the data by first, removing the column ‘docid’, then we removed the duplicated rows, added column 'language', indicating the language of the news, and deleted all the spaces in 'pubname'.

Most articles contains useless phrases like «繼續閱讀 » , and some data are captured from youtube, which include more noise. All of this require proper model which being able to identify and clean it in an efficient manner.

Methodology

We apply sentiment analysis, topic modeling, and time series analysis on external financial and economic data to aggregate the results.

Topic Modeling using BERTopic

BERTopic is a topic modeling technique that uses pre-trained BERT models to identify and cluster similar documents based on their semantic meaning. It efficiently handles large datasets and provides visualizations to understand topic distributions and relationships.

Model Structure

The BERTopic model structure involves four main steps: document embedding, Dimensionality Reduction, clustering, and topic representation. Then we use zeroshot modeling to allow customization of the topic. Together form a pipeline to perform topic modeling.

Tokenization (extra step for processing Chinese text)

To address the tokenization challenges with Chinese characters in BERTopic, a customized approach is implemented. It involves importing a Chinese stop word list from TSCP and utilizing the Jieba library for accurate data parsing. Jieba handles Chinese text segmentation, while the stop word list filters out irrelevant words. These modifications enhance BERTopic's ability to process Chinese text and improve topic extraction.

Embeddings

BERTopic starts with transforming our input documents into numerical representations. Although there are many ways this can be achieved, we use sentence-transformers ("paraphrase-multilingual-MiniLM-L12-v2") as it is quite capable of capturing the semantic similarity between documents and suitable on multilingual (Chinese is targeted) document embedding.

Dimensionality Reduction

As embeddings are often high in dimensionality, clustering becomes difficult due to the curse of dimensionality. A solution is to reduce the dimensionality of the embeddings to a workable dimensional space. UMAP is used since it can capture both the local and global high-dimensional space in lower dimensions.

Clustering

After reducing the dimensionality of our input embeddings, we need to cluster them into groups of similar embeddings to extract our topics. We use HDBSCAN as it is quite capable of capturing structures with different densities.

Zero-shot Topic Modeling

Zero-shot Topic Modeling is a technique for identifying predefined topics in large document collections. It enables the discovery of expected topics based on prior knowledge or input from domain experts. Additionally, this method allows for the creation of new topics that may not align with the predefined ones. It offers flexibility and can be applied in various scenarios to explore topic extraction possibilities. We manually curated a set of finance and economy-related topics and keywords. Through several iterations of topic modeling and observing the results, we were able to identify 9 specific topics that are of primary focus for risk management. However, we also allowed the model to generate additional topics beyond these predefined ones. This approach combines manual curation with the flexibility of the model to ensure comprehensive coverage of relevant topics while exploring potential new insights.

```
event_topic_list  =  [
        ['加息',  '降息',  '鮑威爾',  '港元',  '拆息',  '金管局','利率',  '國債',  '匯率',  '外匯',  '央行','聯儲局',  '美元',  '美聯儲',  '基點',  '美聯'],
        ['股市',  '指期',  '指數','股價',  '本港',  '香港'],
        ['預算案',  '減稅',  '財政',  '支持',  '人民',  '政府',  '融資',  '政策',  '財政赤字',  '財政措施','本港',  '香港'],
        ['增長',  '衰退',  '經濟',  '增速',  '消費',  '預測',  '通脹',  '市場',  '購買力',  '信貸',  '經濟指標',  'GDP',  '本港',  '香港',  '中國'],
        ['房地產',  '地產',  '房地產投資信託',  '房地產開發',  '樓市',  '樓價',  '住宅市場',  '房屋貸款',  '本港',  '香港',  '中國'],
        ['貸款',  '銀行',  '貸款產品',  '信貸',  '貸款利率',  '貸款條件',  '按揭',  '本港',  '香港'],
        ['數字貨幣',  '比特幣',  '區塊鏈技術',  '加密貨幣交易所',  'ICO',  '加密貨幣監管',  '虛擬貨幣',  '加密貨幣市場'],
        ['黃金',  '黃金價格波動',  '黃金儲備',  '避險資產',  '金價',  '黃金市場'],
        ['原材料',  '油價','原油',  '石油',  '每桶'  ,'減產',  '庫存',  '需求'  '大宗商品',  '農產品',  '金屬',  '能源',  '期貨交易',  '天然氣']
]
```

Dynamic Topic Modeling

Dynamic topic modeling (DTM) refers to a range of techniques that focus on analyzing the changes and evolution of topics over time. These methods provide insights into how topics are represented and manifested across different points in time. By including timestamps and splitting the dataset into subsets based on time periods (e.g., 6-10 months), you can analyze the data's temporal aspect and account for variations in data distribution. This approach helps identify trends and patterns that evolve over time, especially for time-sensitive data.

Sentiment Analysis using SnowNLP

We perform sentiment analysis to get the emotional tone behind a series of words. SnowNLP is a Python library designed for analyzing Chinese text. We use it to determine how possible each piece of news is to be in a positive tone.

Tokenization and Stopword Removal

The model first divides the news into words, then it removes the stopwords, which are considered to be of little value in helping to understand the text.

Bayes Theorem

Using the pretrained Bayes model, the model produces the probability of each piece of text being positive. An output value closer to one indicates that the text is highly possible to be in a positive tone, while output values closer to zero indicate higher probabilities of text being in negative tone.

Results and Findings

Topic Modeling using BERTopic

Topic Modeling

- The three figures should be read in sequence, with each graph complementing the others. By focusing on topic 0, which pertains to keywords such as ['加息', '降息', '鮑威爾'...']. We can discern a clear pattern. The frequency of articles related to this topic aligns with the timing of interest rate announcements, occurring roughly every two months. Notably, around the dates of March 22, May 3, and July 26, 2022, when the Federal Reserve made interest rate announcements, we observe a significant surge in the frequency of articles covering this topic. This indicates a strong correlation between the Fed's interest rate decisions and the attention given to this subject.
- Another simple reference that we can draw from Figure 3 is that as the frequency of [加息, 降息] in the year 2023 decreases, there's a clear uptrend in S&P 2023, resulting in a 17% performance from March 2023 to Sep 2023, just before a slight uptick in Oct 2023.

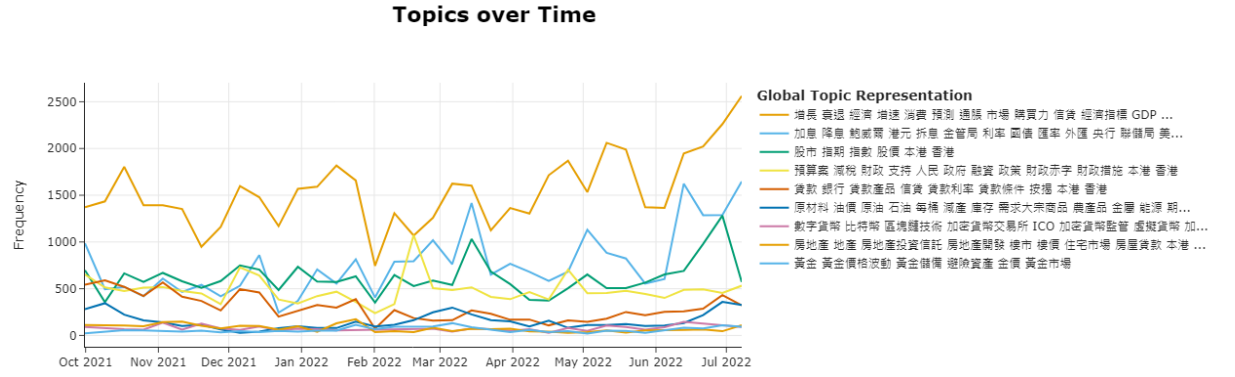


Figure 1_First_Batch

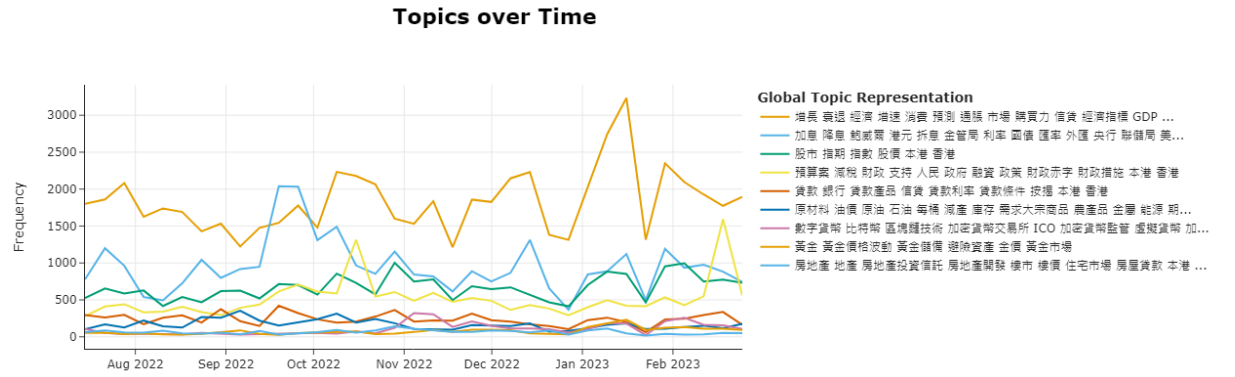


Figure 2_Second_Batch

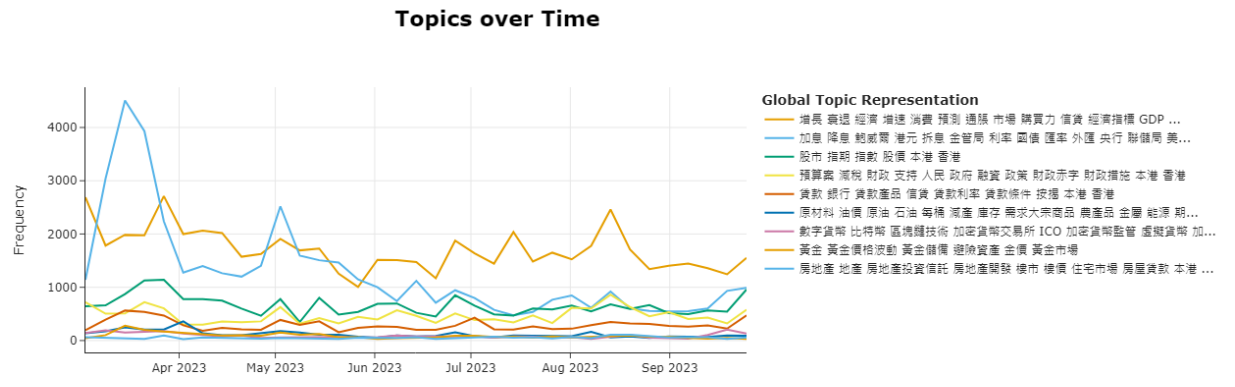
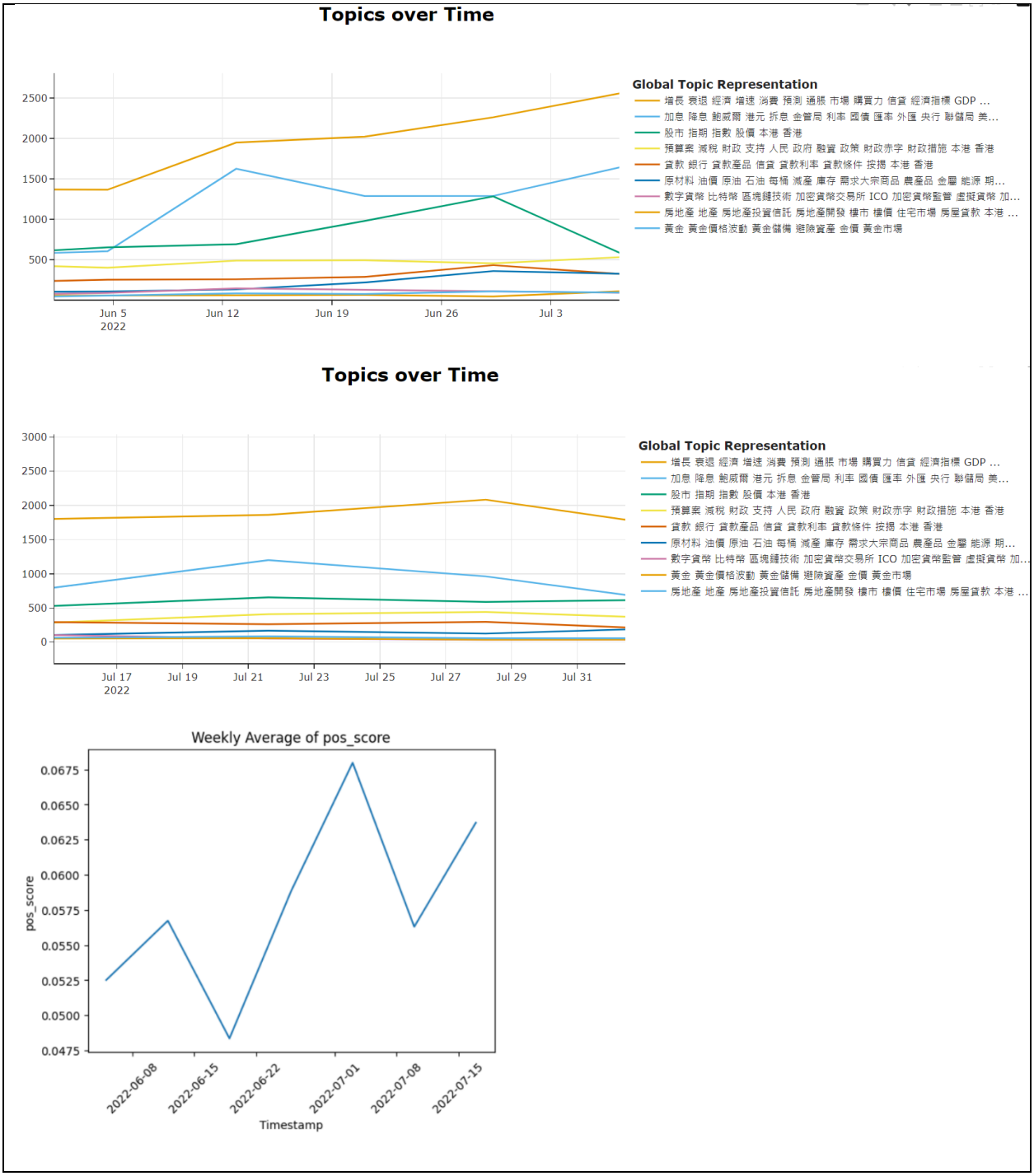
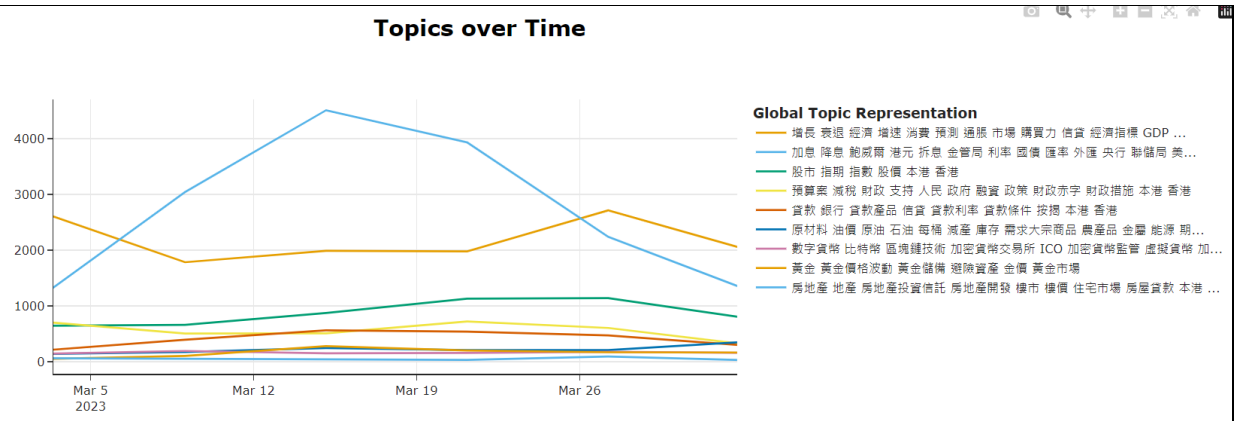
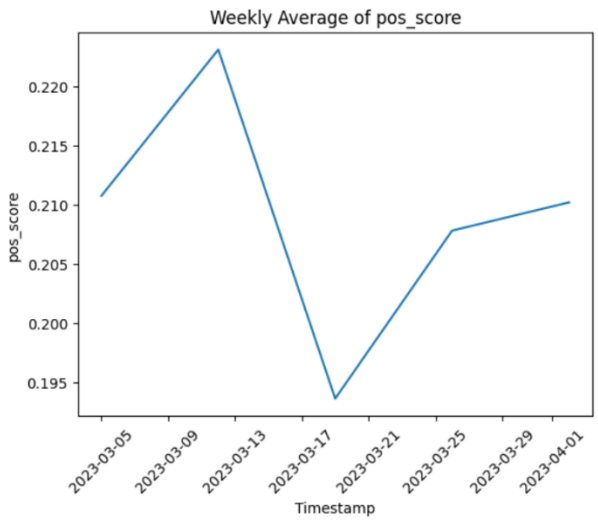


Figure 3_Third_Batch


Topics over Time
2500
2000
1500
1000
500
Jun 5
2022
Jun 12
Jun 19
Jun 26
Jul 3
Global Topic Representation
增長 衰退 經濟 增速 消費 預測 通脹 市場 購買力 信貸 經濟指標 GDP ...
加息 降息 鮑威爾 港元 折息 金管局 利率 國債 匯率 外匯 央行 聯儲局 美...
股市 指期 指數 股價 本港 香港
預算案 減稅 財政 支持 人民 政府 融資 政策 財政赤字 財政措施 本港 香港
貸款 銀行 貸款產品 信貸 貸款利率 貸款條件 按揭 本港 香港
原材料 油價 原油 石油 每桶 減產 庫存 需求大宗商品 農產品 金屬 能源 期...
數字貨幣 比特幣 區塊鏈技術 加密貨幣交易所 ICO 加密貨幣監管 虛擬貨幣 加...
房地產 地產 房地產投資信託 房地產開發 樓市 樓價 住宅市場 房屋貸款 本港 ...
黃金 黃金價格波動 黃金儲備 避險資產 金價 黃金市場
Topics over Time
3000
2500
2000
1500
1000
500
0
Jul 17
2022
Jul 19
Jul 21
Jul 23
Jul 25
Jul 27
Jul 29
Jul 31
Global Topic Representation
增長 衰退 經濟 增速 消費 預測 通脹 市場 購買力 信貸 經濟指標 GDP ...
加息 降息 鮑威爾 港元 折息 金管局 利率 國債 匯率 外匯 央行 聯儲局 美...
股市 指期 指數 股價 本港 香港
預算案 減稅 財政 支持 人民 政府 融資 政策 財政赤字 財政措施 本港 香港
貸款 銀行 貸款產品 信貸 貸款利率 貸款條件 按揭 本港 香港
原材料 油價 原油 石油 每桶 減產 庫存 需求大宗商品 農產品 金屬 能源 期...
數字貨幣 比特幣 區塊鏈技術 加密貨幣交易所 ICO 加密貨幣監管 虛擬貨幣 加...
黃金 黃金價格波動 黃金儲備 避險資產 金價 黃金市場
房地產 地產 房地產投資信託 房地產開發 樓市 樓價 住宅市場 房屋貸款 本港 ...
Weekly Average of pos_score
pos_score
0.0675
0.0650
0.0625
0.0600
0.0575
0.0550
0.0525
0.0500
0.0475
2022-06-08
2022-06-15
2022-06-22
2022-07-01
2022-07-08
2022-07-15
Timestamp





Sentiment Analysis on Top of Topic Modeling

Now that we see the trending topics among the news throughout each week, we want to focus on the topics that were popular and see if there is certain patterns in the sentiments.

From graph 1, 2, we can see that topics related to keywords such as 經濟 (Economy), 消費 (Consumption), 市場 (Market), GDP (Gross Domestic Product) is very popular between the start of June 2022 and the end of July 2022.

In March 2023, news about 加息(Raise interest rates), 降息(Lower interest rates), 港元(Hong Kong dollar), 利率(Interest rate), 匯率(Exchange rate) were trending. The positivity of news under this topic fluctuates over the weeks, with a significant dip around mid-March and a recovery towards the end of the month.

The high frequency periods that we singled out for Topic 0 and Topic 1 both happen before FOMC meetings, a meeting where the Federal Reserve Board determines the future interest rate. In 2022, we saw a clear uptick in frequency for topic 0, representing an increase in worry in recession as shown in the negative sentiment. For Topic 1 interest rates, it appears that the frequency always increases steeply before the FOMC meetings and drops steeply as the news is released. In this scenario, the public appears to be more concerned with “increasing the interest rate” as shown in the negative sentiment. We find that explaining each topic might require further context in the economic and financial sense, and we shall conduct more thorough deep dives onwards.

Limitations and Areas of Improvements

For Bertopic, due to its modular characteristics, we have a lot of space for improvement, for example, finding a better clustering model, fine-tung topic representation using LLM, or generally looking for better and more advanced model for each module and allocate more computation power for training and testing. But most importantly, we should do many iterations of trial and error to identify the topic of interest to better fit the dataset. While identifying more specific topic, we can look for more external index to validate our model predictions and performance, the results above showcase a simple but robust demonstration of such actions.

For sentiment analysis, since we fitted the data using a light weight model, we will apply more advanced and complex model to produce better interpretations. At the same time, we should also analyze the news in other languages, equip the emoji and likes count in the social media articles to assist the analysis.

Conclusion

By using BERTopic for topic modeling,