LIX1

FYP Final Report

Counterparty Risk Alert System for Société Générale Asia

by

POSWAL Abhishek, CHOW Walker Ho Yin, IP Si Hou, NG Ka Fong

LIX1

Advised by

Dr. Cindy Xin LI

Submitted in partial fulfillment of the requirements for COMP 4981

in the

Department of Computer Science and Engineering

The Hong Kong University of Science and Technology

2021-2022

Submission Date: April, 20, 2022

Abstract

Individuals and investment firms lose a great deal of money when one of their investments goes bust. It is even worse when there are all kinds of indicators and news that could have potentially helped minimize the losses but too little time to act upon the situation. Counterparty Risk Alert System is a web application that utilizes machine learning techniques such as natural language processing to predict the sentiment, suggest a summary, and look for user-defined topics in the news of the investors' counterparties.

A model is no good if its results cannot be inferred by the users. Therefore, our user interface is key for putting the analysis of these models together. It provides a custom analysis of each counterparty based on the market data and throws alerts before a major event transpires. The key differentiator between our system and other alert systems is that we leverage state-of-the-art advances in NLP to account for text-based data sources, such as news, to generate alerts, instead of just numeric data many alert systems do. This allows individual and firms to keep track of a large amount of companies without having the intensive labor cost of reading pages upon pages of news.

]Table of Contents

1. Introduction

1.1 Overview

Counterparty Risk Assessment Problem

This final year project is an industrial collaboration with the Data Science team of Société Générale (SG) under their Risk department in Asia. The purpose of this collaboration was to help develop a tool that can help or warn their Credit Risk team. Specifically, our application identifies when one or more of the company's counterparties may have elevated risk factors that are not encompassed in their current set of methodologies and indicators.

To clarify, counterparties refer to companies or entities with which SG has financial transactions or contracts, and counterparty risk refers to the risk of those entities not being able to fulfill their end of a deal. In the worst case, this would mean completely defaulting on a loan from Société Générale.

During our discussions with SG, the specific problem floating up, which was not addressed in their original workflow, was not being able to notice sudden and unexpected events related to one of the counterparties in their portfolios. Some examples of such events included the 2021 collapse of Archegos due to the failure in meeting margin calls, the Japanese brokerage Nomura taking a $2.9 billion loss from that same collapse, or the 2020 bankruptcy of Baoshang. More general or mundane examples could be the resignation of CEOs or fines suffered by one of the counterparties. Most of these examples are related to news events and do not reflect immediately into indicators or metrics they use, such as quarterly earnings. Such sudden and unexpected events are also difficult to notice easily through the manual reading of news or internal newsletters due to the sheer size of their portfolios. To the Credit Risk team, these events are important, because failing to notice them immediately means a delay in a necessary reevaluation of counterparties, increasing the company's exposure time to potentially risky counterparties. With our alert system, Société Générale can have more time to take protective action, like short selling the risky counterparty's stock, to hedge against loss.

Hence, the target for our FYP was to explore different techniques to monitor and even forecast which market events may negatively impact the counterparties in their portfolio, through natural language processing (NLP) on news sources, and then alert them when one of their counterparties has the potential to have increased risk and require human reevaluation. In summary, we aim to create a counterparty risk alert system that does not solely rely on numeric sources but can ingest and make use of non-numeric sources such as text too.

Existing Counterparty Risk Assessment Methodologies

To manage risk within a corporate portfolio, one method is to monitor a set of Key Risk Indicators (KRIs) [1]. KRIs are critical predictors of unfavorable events that can adversely impact organizations. Some of the common indicators that directly relate to risk are liquidity, capital adequacy, profitability, and geographical location. Based on these generic indicators, we have well-defined ratios, such as debt-to-capital and debt-to-equity. Rule-based alert systems constructed with these numerical indicators are already prevalent and widely used. However, these indicators are typically only updated quarterly or half-yearly, depending on the counterparty's balance sheet release frequency. Therefore, they do not provide a full picture of the position of a counterparty at a specific time.

Another approach to monitoring corporate risk is to look at credit ratings. Credit ratings refer to the ability of an entity (an individual, government, or business) to fulfill its business obligations within a certain deadline. They are issued by credit agencies, such as Moody's Investor Services, Standard and Poor's (S&P), and Fitch Group [2]. They are usually more reliable than KRIs in the sense that besides numerical indicators, non-statistical features such as reputations and related news, are considered during the credit rating process. However, they are still very delayed, as it normally takes time for the teams at these agencies to reevaluate and update these ratings, therefore making them insensitive to sudden news changes.

Considering the limitations of numerical alerts, ruled-based systems, and external credit ratings, there is a need to develop an automatic news-based alert system. Traditionally, an automated news-based alert system is difficult to develop, as human languages are ambiguous in such a way that risk judgments cannot be made with simple conditionals. However, with the advancements of machine learning and NLP technologies, it is now possible to encode news texts in numeric representation and hence enable computers to threshold the numbers and trigger alerts.

1.2 Objectives

Our goal in this project was to develop a counterparty risk alert system that not only issues warnings based on price or certain numeric indicators but also based on changes in news and web-based text content. To achieve this goal we mainly focused on these three objectives:

1. Find ways to detect a sudden change in risk based on the news

1.1. Find a way to ingest news into a database. Either through an API or an offline dataset.

1.2. Research and implement a way to sort news by the counterparty.

1.3. Research ways of identifying key events from news articles that may affect the risk factor of a counterparty.

1.4. Research ways to identify the sentiment of financial news and sudden changes in sentiment.

2. Investigate relationships between different times-series data on counterparties for potential correlation or forecasting

2.1. Investigate ways to ingest other sources of time-series data on top of the ones generated from news from methods in the first objective.

2.2. Research how these time-series data relate/correlate and how this can be used to potentially forecast or warn.

2.3. Choose one data source that is indicative of counterparties' credit health and use the other sources as input to do a prediction.

3. Provide a web app that displays information and risks

3.1. Provide a responsive web app that allows users to manage the list of counterparties they want to track in this system, through desktops, tablets, or mobile phones.

3.2. Display alerts and warnings generated (as in the first objective) on the web app prominently and concisely so that the credit team of *Société Générale* can notice any incidents that happen each day.

3.3. Present detailed information of each counterparty and visualize correlations (from the second objective) with UI elements like charts.

3.4. Provide options for users to rate the warnings generated and correct any mistakes of the system.

1.3 Literature Survey

There are massive amounts of financial applications related to market risk. Most of them are trading applications for stocks. In most applications, features like displaying the indicators related to specific stocks are implemented, together with some time-series data. Some of these applications are discussed below.

1.3.1 Stocksnips

StockSnips is a small startup specializing in delivering stock market news sentiment. They have a mobile app of the same name [3]. The app has amassed over ten thousand downloads on the Google Play store since 2016 [4]. Additionally, the company has raised over a hundred thousand USD through seeding rounds from investors [5]. This app is one of the few portfolio apps on the market that includes news sentiment within it.

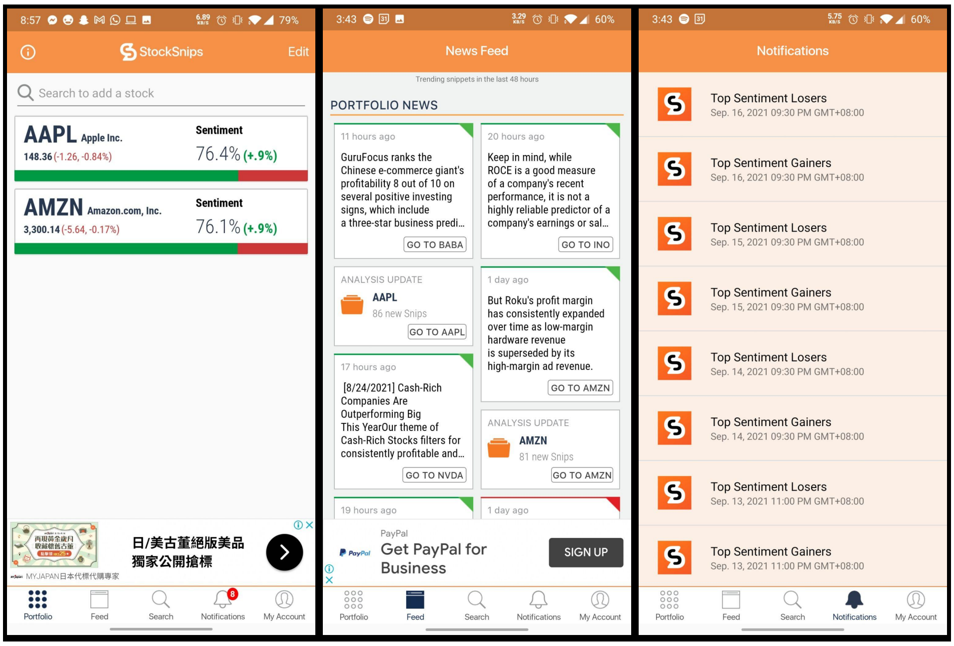


Fig 1. Screenshots of the 3 Main Views of the Stocksnips App

The app itself is rather simple, with 3 main views to it (portfolio, feed, and notifications) as seen from the screenshots in Fig 1.

The news feed view is a simple card UI that displays the most ‘trending’ news from the user’s self-defined portfolio and outside it. Each “news article card” is tagged with the relevant company and is either green or red depending on the sentiment ratio of the sentences. This approach seems to rely on how much in-app attention each article gets and ultimately relies on crowdsourcing to decide which news is the most important, meaning it is not guaranteed for news that affects portfolio entities the most to show up and requires a large number of users for it to be feasible, which means it may not apply to a project until it is deployed and scaled broadly.

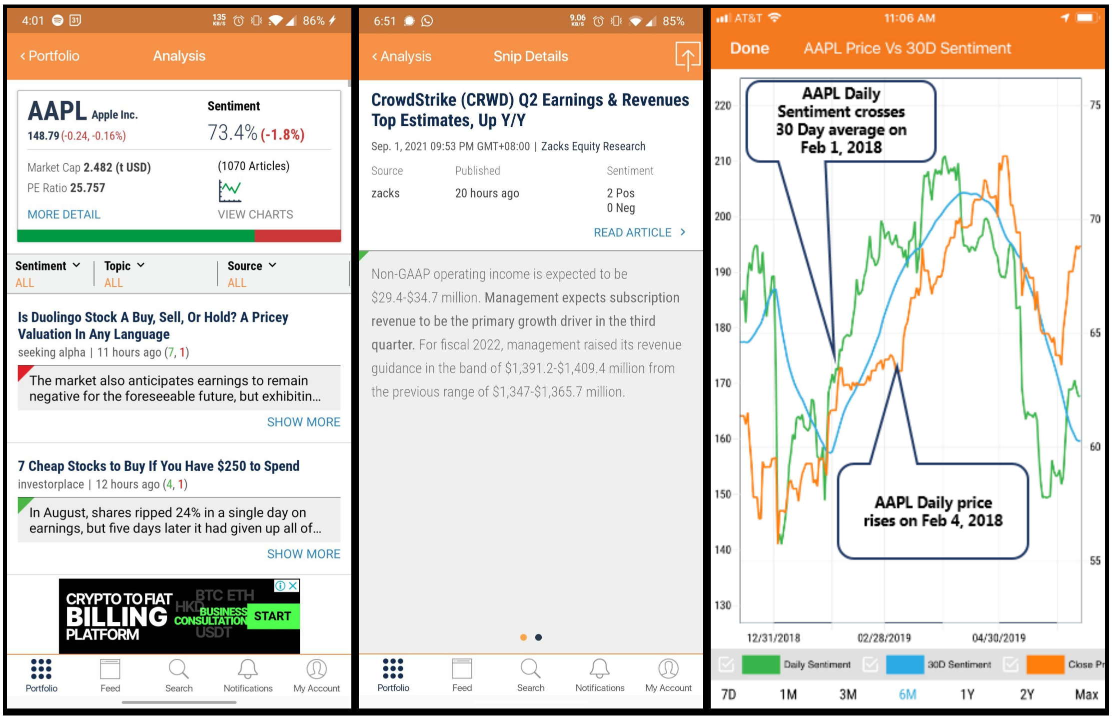


Fig 2. Screenshots of the Views Within the Portfolio View

The portfolio view allows users to see all the stocks in their self-defined portfolio, showing each stock's current price and sentiment percentage along with the daily change. Clicking on each stock shows further financial information as well as a list of the latest articles related to that company. These articles are from a variety of sources (Zacks, WSJ, Barrons, Reuters, etc.) [6] and have a count of the amount of 'good' and 'bad' sentiment sentences within them (see Fig. 2 above). The number of selected sentences with sentiment is low and is displayed to the user. For paid users, they can view the sentiment % plotted against the stock price in a separate view (also seen in Fig. 2).

From a user standpoint, the information presented in the portfolio view is relevant but may not be as actionable or useful as one may initially think. One of the reasons is that sentiment % is an inherently vague descriptor. For example, is 60% bad or is 40% bad? Without providing context through some kind of analytics on how badly a change would affect that company, users may not be able to evaluate these metrics themselves unless they are intimately familiar.

Furthermore, the sentiment % also seems to have some reliability issues. For one, their sentiment % seems overly positive; even across a sample of random stocks and most shorted ones, the sentiment percentage is consistently well over 50% (Appendix 6.1), making it extremely hard to identify companies with a poor sentiment. Another problem contributing to unreliability is that the 'sentiment sentences' it picks out sometimes are not even related to the stock entity in question. For example, in the screenshots above, the article about CrowdStrike was marked under Apple's stock, presumably because the company's name was mentioned in the article. However, in the article, Apple was only mentioned as an alternative stock to buy instead of the article's main subject CrowdStrike [7], and the highlighted sentence that contributes to Apple's positive sentiment was actually about CrowdStrike instead. There is an argument to be made that maybe positive sentiment of companies in the same industry can be good, but given the fact that the article explicitly recommends one stock over the other, the most likely conclusion is that they simply do not do adequate name entity recognition of the sentence they sentimentalize, hence contributing to the unreliability of their metric.

To make these sentiment type metrics useful to end-users managing their portfolios, it may be useful to define some "normal" ranges for these metrics as well as do some correlation analysis with various important values, such as price, to better show users when and what sentiment movement should concern them (some form of prediction could be done too). In terms of reliability of the metric, it may be good to reference a project done for the UK government to track business sentiment [8] where an NLP algorithm was used to identify direct reported speech first, which may provide more accurate sentiment about a company. In general, some sort of preprocessing like entity recognition and postprocessing like sentiment validation should be done to make sure the sentiment is relevant.

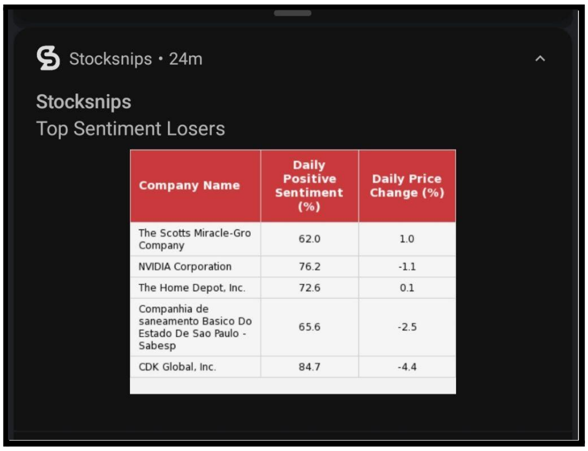


Fig 3. Sample Notification from the StockSnips App

In terms of the alert or notification part of the app, it seems very limited and non-personalized to specific portfolios. There are no options available for users to customize the notifications they can get from the app. The only type of notification users seem to get is at the end of the day, where there seems to be a generic generated image sent to all users of the top sentiment "losers" of that day. These top "losers" are not even specific to the user's portfolio but rather it is done across the whole US stock market. Therefore, there is no real way for a user to find which is the top "sentiment loser" of his portfolio that day which is especially at risk and requires attention. There's also no way to define certain sentiment loss thresholds at which a user wants to be notified when one of his stocks is deemed to be risky or when certain events occur.

In general, the app does not seem to excel at helping to manage large portfolios, as it does not seem to be able to clearly identify and show important risk events to users. It seems that the app is built solely around this one very opaque sentiment indicator and not around being a portfolio alert application that uses the change in global news as an alternative event alert source instead of price.

1.3.2 S&P Global Market Intelligence

To create an early alert system for market risk, the American publicly traded corporation, S&P Global Market Intelligence, suggests a solution to build an early warning system to alert their risk team when there are significant changes in the global market [9]. Having situations that are similar to what Societe Generale is facing, the risk team of S&P Global is required to monitor and assess corporate companies periodically. They thought it would be useful to build a system to keep track of companies in terms of their risk, as well as provide warnings when their creditworthiness decreases significantly. By utilizing the system, S&P is better able to minimize the number of companies that may possibly be delinquent.

S&P Global suggests that a good credit warning system should comprise several features, which will immediately determine the emerging problems of the bank's partners.

1. A bunch of credit assessment tools including comprehensive data, quantitative models, and sophisticated web-based workflow solutions. A range of indicators that evaluate a company should be included in the models, generating a unique credit score that aims for that specific company. Workflow capabilities allow the risk team to set alerts on the system and use color codes such as red, and yellow, to indicate a company's credit risk inside the dashboard of the system.
2. A continually expanding private company database to be cautious of those companies that provide a significant impact on the market.
3. Timely news and research to provide a better and more detailed insight behind the significant changes in a company's financial conditions.
4. Macroeconomic data to determine external factors that can possibly provide an impact on the companies in the bank's portfolio.
5. Ongoing support to describe the capabilities, enhance the effectiveness of the risk team, and leverage the solution to its fullest.

Without a doubt, the system improves the efficiency of the credit team by alerting them automatically. Each company's creditworthiness can be determined by its automatically generated credit score. A well-rounded view of the breaking news is delivered to the credit team immediately, for the purpose of instant evaluation of the negative impact caused. Also, the dashboard and visuals display the information modularly and assist collaboration between colleagues. They are able to understand the flow thoroughly with ongoing support and training.

However, there is a massive amount of news and counterparties displayed to the credit team through the system. The news is displayed randomly and is not sorted in terms of importance. The credit rating team needs to read the news themselves and evaluate its potential risks, which is not efficient enough, as reading a large amount of news takes a significant amount of time.

1.3.3 Loan Default Forecast in Europe

In the Oxford Academic Journal of Financial Econometrics, a dataset of 12 million residential mortgages was used to investigate loan default behavior in various European countries [10]. The default occurrences were modeled as a function of borrower characteristics, loan-specific variables, and local economic conditions. The performance of a set of machine learning algorithms was compared to the logistic regression (benchmark), finding that they perform significantly better in yielding predictions [10]. It was found that the interest rate and the local economic characteristics were the most crucial features explaining the default rate.

The loan-level information for seven European countries was collected over a period of 5 years. The dataset included dynamic features about the performance of the loan that was updated at least on a quarterly basis in addition to the static features which contained the borrower's income, loan amount, and location. Below is a table that identifies the loan, borrower, and region-specific variables.

Table 3 Explanatory variables used to predict the occurrence of a default

Feature	Attribute	Type	Description
Loan-specific variables			
DTI	Static	Numeric	DTI at origination
Interest rate type	Static	Categorical	Interest rate type
Interest rate	Dynamic	Numeric	Interest rate at the pool cutoff date
LTV	Dynamic	Numeric	LTV at pool cutoff-date
Property type	Static	Categorical	Property type of the underlying asset
Seniority	Dynamic	Numeric	Loan seniority at origination (in days)
Valuation amount	Static	Numeric	Property value as at loan origination (in
Borrower-specific variables			
Borrower's employment	Static	Categorical	Employment status of the applicant at o
Income	Static	Numeric	Borrower gross annual income at origina
Regional-specific variables			
Default rate	Dynamic	Numeric	Default rate (%) by NUTS3 lagged 1 year
GDP growth	Dynamic	Numeric	GDP percentage growth by NUTS2 lagge
House Price growth	Dynamic	Numeric	House price percentage growth by NUTS
Unemployment rate growth	Dynamic	Numeric	Unemployment rate growth by NUTS2 la

Fig 1: Input Data Features [10]

The models that the paper used were quite commonly used by machine learning professionals with minor changes made to the hyperparameters. Following are the models that were used in the paper, based on their ranking:

1. Extreme Gradient Boosting Tree (69% accuracy)
2. Gradient Tree Boosting (67% accuracy)
3. Penalized Logistic Regression (63% accuracy)

Overall, most of the relevant variables that define the results were pointed directly at loan-related variables, namely, current Loan To Value (LTV) and current interest rate [10].

Although the results were quite interesting, it should be noted that this data was not continuous and was only updated quarterly. Therefore, there would be no way to forecast results in a timely manner if a major event happened that directly affected the company. However, in our project, we mitigate just that problem by feeding the real-time news from different sources whether it be numerical or textual. With the help of this data, our system is able to provide real-time warnings. Additionally, our machine learning models are tailored to process this kind of data and find correlations between them.

1.3.4 Business News Headlines and the Prophetic Vision of Bankruptcies

Appelbaum, Duan, Hu, and Sun (2021) built a machine learning model to provide predictions of whether a company will go bankrupt in 1 or 2 years [11]. In their machine learning model, they utilized news headlines, together with financial statistics and administrative properties of the company as input features, achieving an area under curve (AUC) score of 0.965 for a 1-year prediction and 0.892 for a 2-year prediction.

In their paper, Appelbaum et al. listed their data sources and input variables explicitly. They combined three data sources - Compustat and Audit Analytics and Reuters. Compustat is mainly used to provide financial statistics, and Audit Analytics is used to provide the administrative features and bankruptcy labels for the companies (as the independent variable for the learning model). From Reuters, they extract only related news headlines for further analysis, as news headlines are “more succinct” and “less repetitive” than their contents.

The obtained headlines are passed to 3 sentiment libraries, TextBlob, Flair, and VADER, to obtain sentiment scores. Scores from multiple news headlines within each year were aggregated by min, max, mean and median to provide sentiment scores for a company of that year.

With the previous steps, they obtain three types of input features as the dependent variables, which include:

1) sentiment features as polarity and subjectivity scores
2) financial statistics, like market size, annual return rate, and the debt ratio of the company;
3) administrative features, like the popularity and tenure of the auditor firm, punctuality of 10-K filings, etc.

These features are then passed to Balanced Random Forest and RUSBoost algorithms for predictions. It is notable that even though the machine learning model used is relatively simple, the result accuracy was quite good. Another observation is that financial and administrative features are useful with the fact that these figures have improved the AUC score by 0.2 when compared to a pure linguistic model.

Yet, in their model, they have used pre-trained sentiment models (TextBlob, Flair, and Vader) without fine-tuning. Those models are for general purpose texts and do not have optimal applicability to financial news. We see room for improvement by either further training the models with financial news data or by using other models specifically tuned for financial texts. Furthermore, this model predicts bankruptcy on a year-scale horizon. Further experiments have to be done to test whether this model is valid when used as a daily rolling updated indicator. Nevertheless, they have demonstrated the predictive power of news sentiment towards bankruptcy events and have given us an understanding of the diversities of non-linguistic features that can be used.

2. Methodology

2.1 Design

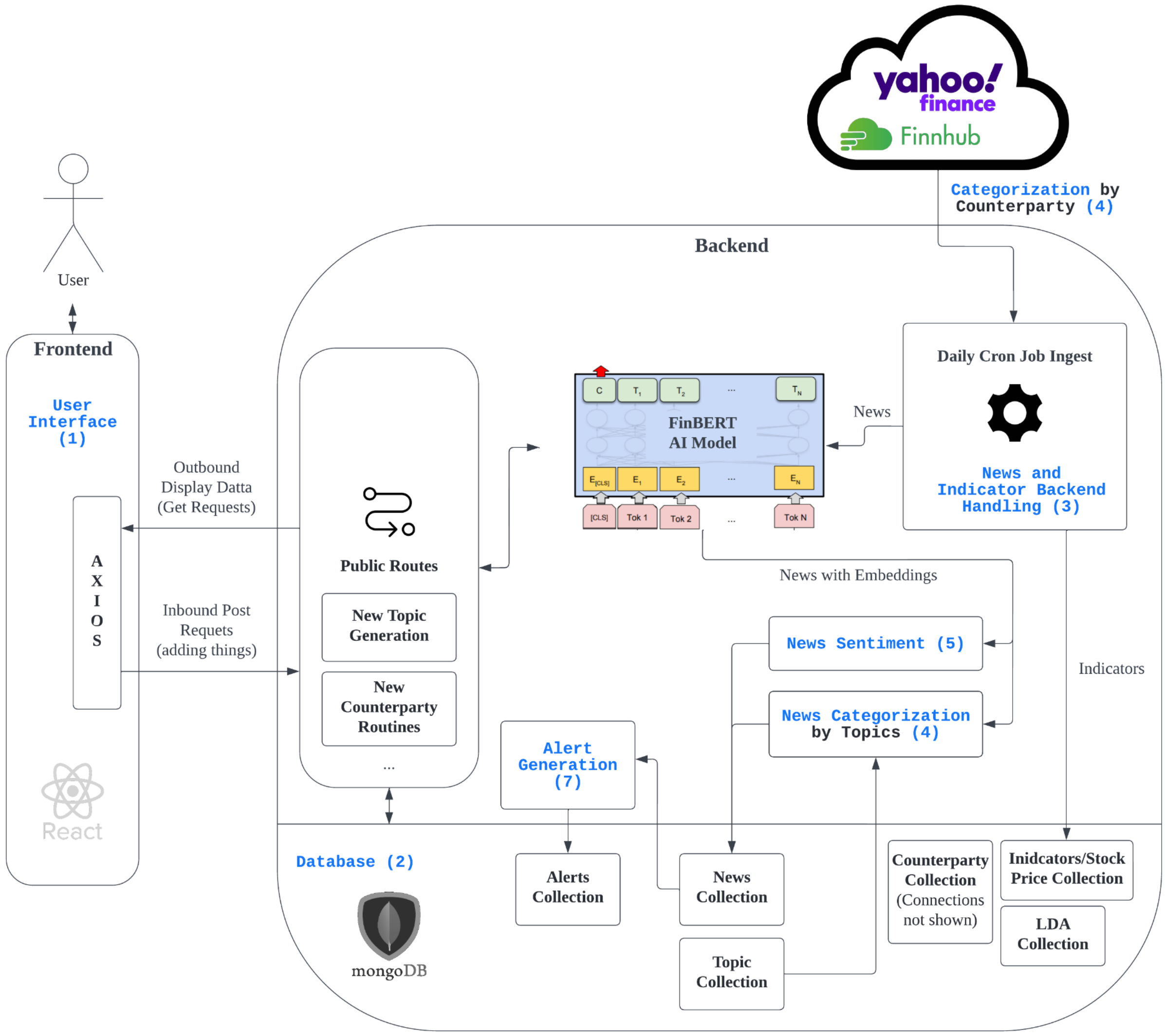


Fig 1. Simplified High Level Overview of Our Final System

(this diagram is a simplified subset of the actual system, mainly focuses on the daily corn flow)

The design of our project consists of 7 main elements, derived from the objectives captured from Société Générale in late July last year and continually iterated on throughout the project. The way each element is used in the final live system can be seen in Figure 1 above. Each design element is highlighted and numbered in the same way it is in our report. It is also listed below.

1. User Interface $_{3}$
2. Database $_{1}$
3. News/Indicator Backend Handling $_{1, 2}$
4. News Categorization $_{1}$
5. News Sentiment Model $_{1, 2}$
6. Correlation/Forecast Model $_{2}$
7. Alert Generation $_{2}$

The subscript indicates the objective(s) each part fulfills.

2.1.1 User Interface

As suggested by the UI flow in Appendix 6.2, the UI consists of

1. A dashboard as a homepage that
   - Displays an overview of the portfolio sentiment
   - Materializes the warnings generated in a list of cards.
   - Each warning card contains the counterparty name, the warning content, the sentiment score, and the key news terms.
   - Next to each warning card, options should be provided for the user indicating whether an alert is correct or not.
2. A counterparty list page that
   - consists of a table showing a list of counterparties previously inputted by the user.
   - that allows users to add new counterparties to track by clicking "New" and delete by clicking the "Delete" button
3. A detailed counterparty information page that
   - displays detailed statistics
   - visualizes the history of sentiment, key terms in charts
   - provide a list of news with their sentiments, and labeled keywords. The list of news refreshes on clicking the chart, allowing users to view news from the past on the selected day.
4. A new counterparty page that
   - Provides a search box for users to search for listed companies to add to their portfolio
5. A new topic page that
   - Provides a text input box for users to type in keywords for the topic
   - Displays LDA suggestions generated by the LDA model

2.1.2 Database

In the system, counterparties are the user inputs while news and time-series data are essential data sources for detecting a change in risk as per objective 1.1. The three types of data should be stored in three tables in the database. Models, calculation results, and alerts are derived based on the former three types of data. Although storage of derived data is not suggested by conventional database design principles, this is necessary in our case considering the hefty amount of computation cost to generate the processed results to be displayed on the frontend. Below is the visualization of the structure of the designed database.

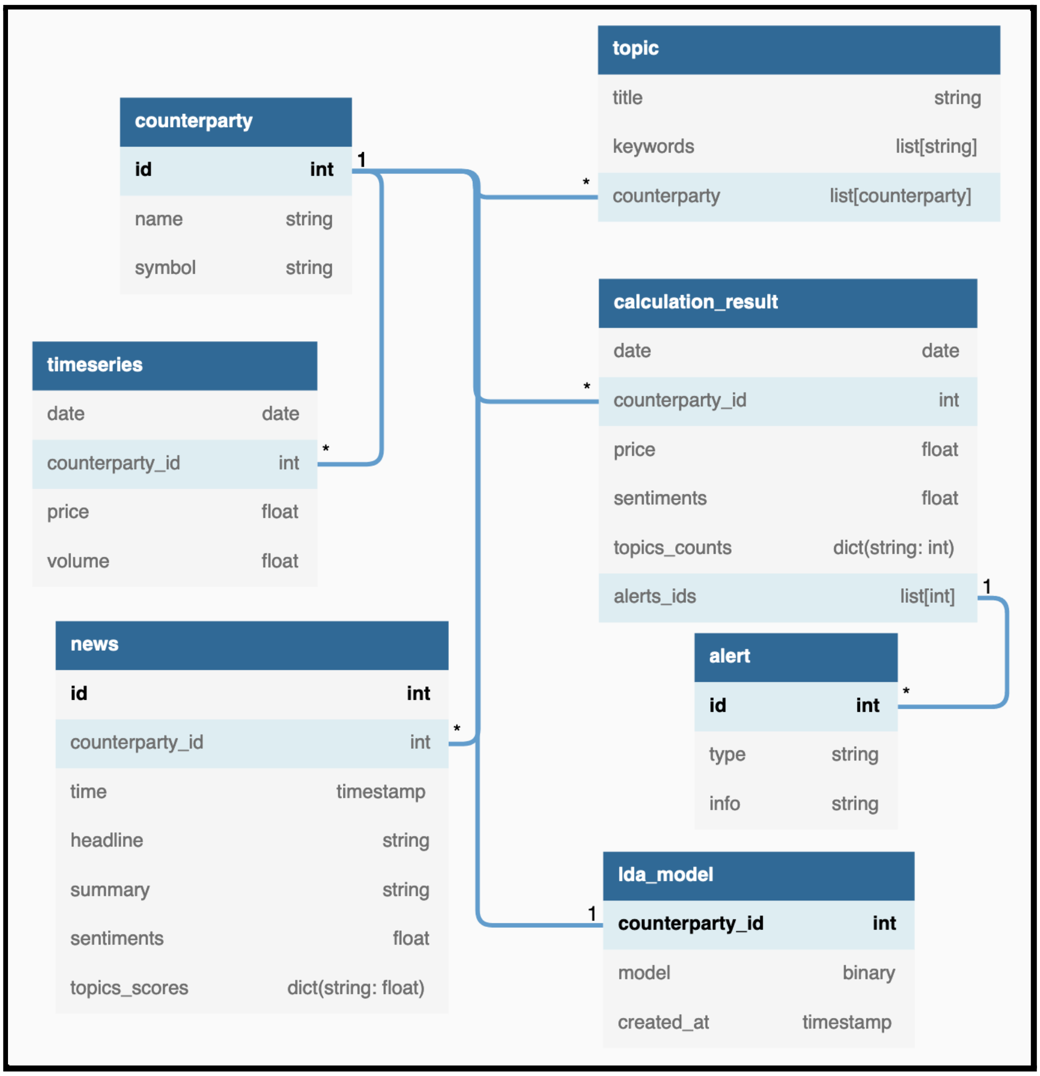


Fig 1. Database Design

2.1.3 News and Indicator Backend Handling

Since we used Python as the primary programming language for machine learning and natural language processing, we also had our backend framework to comply with it. Therefore, we used *FastAPI* as our web framework that is used for building APIs known for its speed for programmers and ease of implementation. In our system, Fast API supports different HTTP Methods (*"get", "put", "delete", and "post"*) , which can serve data from our backend database to the frontend to visualize the stored data in the form of graphs. For the data schema, please refer to 2.1.2 Fig. 1 to understand the collection structuring of the news and indicator data where indicators could be classified as *"topics"* and *"sentiment"*.

Apart from Fast API, we also used Finnhub and Yahoo Finance APIs to integrate news and time-series data sources to our backend. The backend refreshes every day to extract new content from various sources such as Yahoo, Bloomberg, Reuters et al. Since there were limitations as to the number of APIs we could use, as not all of them were available free to integrate into our system (such as Bloomberg/Reuters) despite best efforts from HKUST and SG. We finally settled on using Finnhub's API to pipeline the news from some of the other news sources. We chose Finnhub because its stability, abundance in metadata (such as summary) and wide range of news sources it scraps from (from Barons to CNN to WSJ and even foreign sites).

2.1.4 News Categorization

2.1.4.1 Categorization by Counterparty

The system groups the news articles by counterparties it relates to (as per objective 1.2). This was designed so that when the credit team is investigating the financial conditions/risk factors of a counterparty, they are able to see all the relevant news, as well as the indicators (such as sentiment, or events) generated from the news in one place. This was also important for generating alerts in case of sudden sentiment movements for a counterparty since news is tracked over time for counterparties for continual alert generation. Therefore, the final design for *news categorization by counterparty* was to tag news articles with stock symbols of related counterparties during ingest.

2.1.4.2 Categorization by Events/Topics

Once the news categorization by the counterparty is established, the news is grouped by particular events (as per objective 1.3). This allowed the credit team to track events that may pose risks to a counterparty over time, as well as allowing alerts to be generated when these events occur. At our request, the team at Société Générale gave us a list of 30+ keywords relating to events that may signify a change in the risk status of a company (such as "fraud" in Appendix 6.3). We tracked and identified news relating to these events amongst others, details are in the relevant implementation section i.e. 2.2.4.2.

In addition, we also suggest event topics/keywords for users to track via a suggestion model (LDA model discussed in the implementation section 2.2.4.2 Part 2). This is a new design element added after our proposal, as initial scanning of news for the keywords given by SG revealed very low occurrences.

2.1.5 News Sentiment Model

To evaluate a counterparty's performance, a sentiment model is put in place to check the attitude of the news which is validated by the change in the stock prices of that counterparty (objective 1.4). Since the BERT model was good for performing natural language tasks such as sentiment analysis and text generation, it quickly gained popularity amongst the NLP community [12]. The sentences in the BERT model are generated word-by-word in the pre-BERT stage. BERT allows words to be generated according to their previous and next context. It performs extremely well in all tasks like sentiment analysis or news summarization. However, BERT contains a large corpus of text and the model is too general for training specific tasks like the sentiment analysis of financial news. We would like to use FinBERT, which is developed by Y. Yang, Christopher M.C.S. Uy and A. Huang, from HKUST [13]. FinBERT uses the same architecture as the BERT developed by Google. However, the model fits with a large corpus of texts related to financial information, like using corporate reports 10K and 10Q, earnings call transcripts, and analyst reports to finetune the BERT model, This model is also optimized for sentiment analysis.

Just like BERT, it is not a fine-tuned model but it is a pretrained model. It is pretrained based on 4.9 billion tokens from financial corpora. The tokens are extracted from the most important information inside the financial sector, including corporate reports 10K and 10Q, earning call transcripts, and analyst reports. In total, the paper states that four models are provided, including FinBERT-BaseVocab, uncased/cased models, and also FinBERT-FinVocab uncased/cased models. These models are then trained until the training loss converges. In order to evaluate the models, experiments on sentiment analysis on financial datasets like Financial Phrase Bank, AnalystTone Dataset, and FiQA Dataset are conducted as sentiment analysis is a very popular task in Natural Language Processing. Sentiment analysis is also the task that we want to achieve to evaluate the financial news. The evaluation and comparisons with BERT and other models are shown in the evaluation part in 2.4.5

The API from Finnhub contains the headlines and summaries of a specific news item and the stock symbol of the company that is related to the news. Due to the slow inference time of the long summaries for the model, we decide to use headlines as the input such that it can catch up with the large amount of news obtained from Finnhub in real-time. Every time there is updated news in Finnhub, the system will automatically scrape the news and infer the title of the news to output a sentiment score (positive, negative, neutral), which is then uploaded to the database automatically.

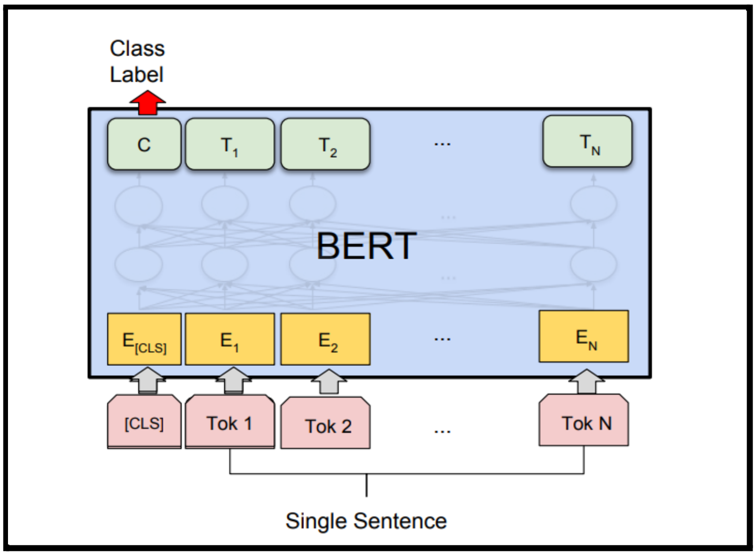


Fig 1. The structure of FinBERT/BERT in sentiment analysis

The news sentiment scores are aggregated per day and per counterparty to provide a count. The counts are further summarized into counterparty sentiment scores using Volume-Weighted Exponential Moving Average (V-EMA) techniques on a time-series graph on the counterparty analytics page for visualization (Fig. 2). To enhance understanding of the counterparty sentiment scores indicator, we further categorize scores above-median as 'positive', while scores when the count of negative news is larger than positive news (< 0) as 'negative'.

To provide an overview of the portfolio monitored, the counterparty sentiment scores are further averaged to form a portfolio sentiment score, displayed together with the categorized counterparty score on the frontend. (Fig. 3)

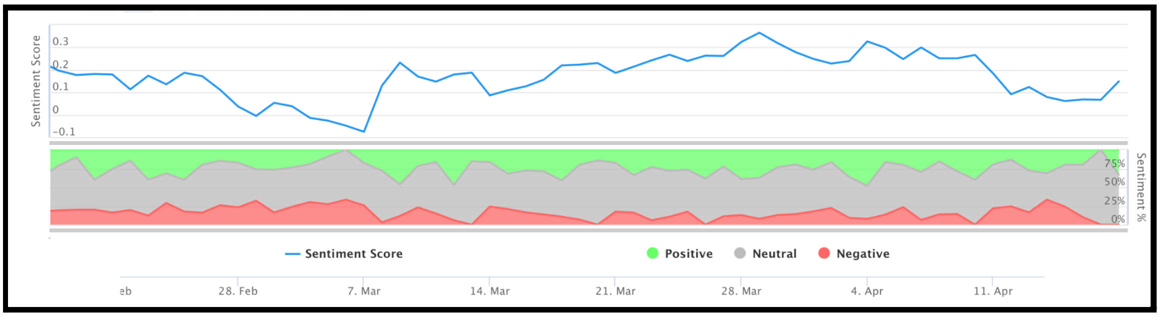


Fig 2. The Counterparty Sentiment Scores and the Aggregated Count of News Sentiment Score

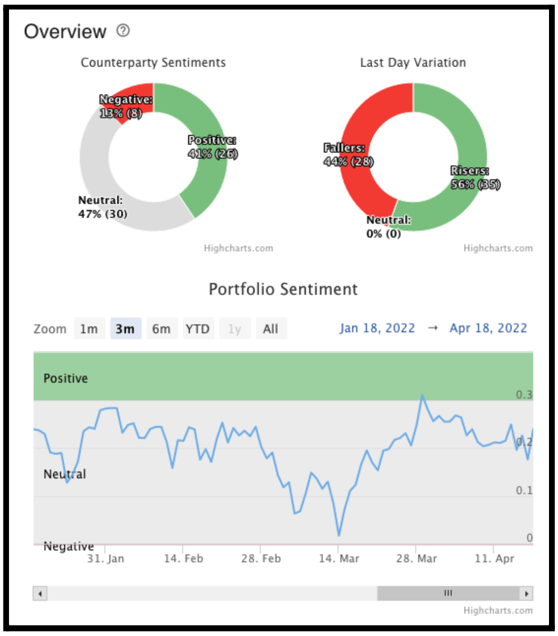


Fig 3. The average Counterparty Sentiment Scores as Portfolio Sentiment Score

2.1.6 Forecast Models

2.1.6.1 Fine-Tune FinBERT

The sentiment score of the model can tell if the financial news is positive/negative/neutral. However, to better reflect the credit risk of different counterparties, we can fine-tune the sentiment model with different sets of data. At first, we decided to use credit rating instead of sentiment score to fine-tune the FinBERT model, such that the model could infer the news of a counterparty to the credit rating of that counterparty. However, we found that credit ratings like Moody's ratings were hard to access and they were not updated frequently. Therefore, we decide to use a financial derivative i.e. credit default swap (CDS), which is relevant to each counterparty, updated daily, and exposed to the public.

A Credit Default Swap (CDS) is a credit derivative contract between two counterparties. The buyer makes periodic payments to the seller, and in return receives a payoff if an underlying financial instrument defaults or experiences a similar credit event. The CDS may refer to a specified loan or bond obligation of a "reference entity", usually a corporation or government. The spread of a CDS is the annual amount the protection buyer must pay the protection seller over the length of the contract, expressed as a percentage of the notional amount. All things being equal, at any given time, if the maturity of two credit default swaps is the same, then the CDS associated with a company with a higher CDS spread is considered more likely to default by the market since a higher fee is being charged to protect against this happening. (Wikipedia)

Therefore, a higher CDS spread of a counterparty implies a higher counterparty risk. We decide to finetune the Finbert model with the CDS spread as the target and the headline of the news as the input. However, the CDS spread of each counterparty varies a lot, therefore, the CDS spread is first converted to the percentage change of the CDS price of each counterparty across each day. To elaborate more, each news headline has a date and a counterparty originally, we will label the news headline with the percentage change of CDS spread of the corresponding date and the corresponding counterparty. The dataset (news headlines as the input and percentage change of CDS of the related counterparty as the output) are trained with the FinBERT.

As a result, by inputting the headlines of news of a specific counterparty to the model, it will output the percentage change of the CDS of that specific counterparty related to the date when the news is released. This percentage change of the CDS may better reflect the counterparty risk than just the positive/negative/neutral of the sentiment model. In summary, each news corresponding to each counterparty is considered training data. The data is then input in the FinBERT model, a percentage change in the CDS price of the counterparty on the date when the news is released.

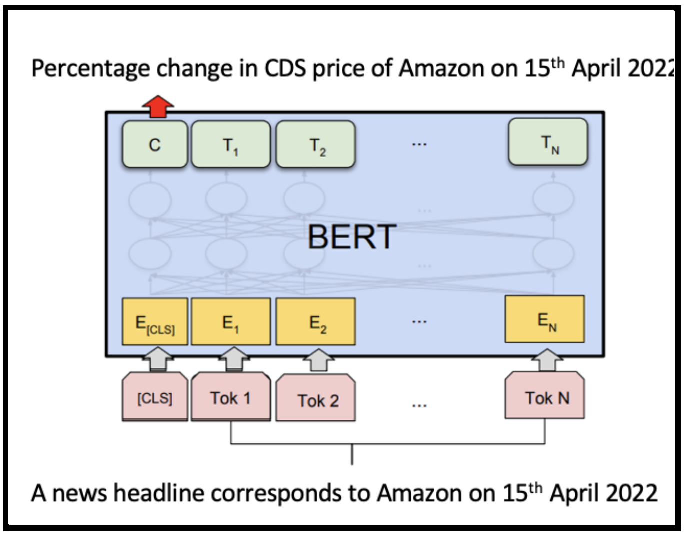


Fig 1. The structure of the fine-tuned model that outputs the percentage change in CDS price of Amazon from one single news headline on 15th April 2022 that corresponds to Amazon

However, this approach has two problems involved.

1. CDS price is updated daily, therefore, daily CDS price change should correspond to the performance of the company in the whole day. This approach only inputted single news but not the news of the whole day.

2. There will be a massive amount of data involved as there will be lots be news per day and per counterparty. Details about the training time will be mentioned more in evaluation part 2.4.6

To solve problem 1, we think of concatenating all the news correspondences of one counterparty in one single day as input data, then the output data will be the percentage change in CDS price on that day. Mr. Martin from Societe Generale said that this approach will make the training even slower as concatenating the news will make the training data .

To solve problem 2, news headlines are used as the input to the model, such that the input dimensions are much smaller. But the training can still be faster.

2.1.6.2 Using Sentences Embeddings

To solve both problems, we try to utilize sentence embeddings that FinBERT provided. When we input a news headline into FinBERT, at the last few layers with lower dimensions, it contains the semantics of the news headline, we extract the outputs of these layers as the sentence embeddings. These embeddings are vectors with smaller dimensions. Thus, the sentence embeddings of each news headline in each counterparty on the release date of the news are averaged out. Then we use a simple neural network with three layers to output the percentage change in the CDS price of the counterparty on the release date of the news.

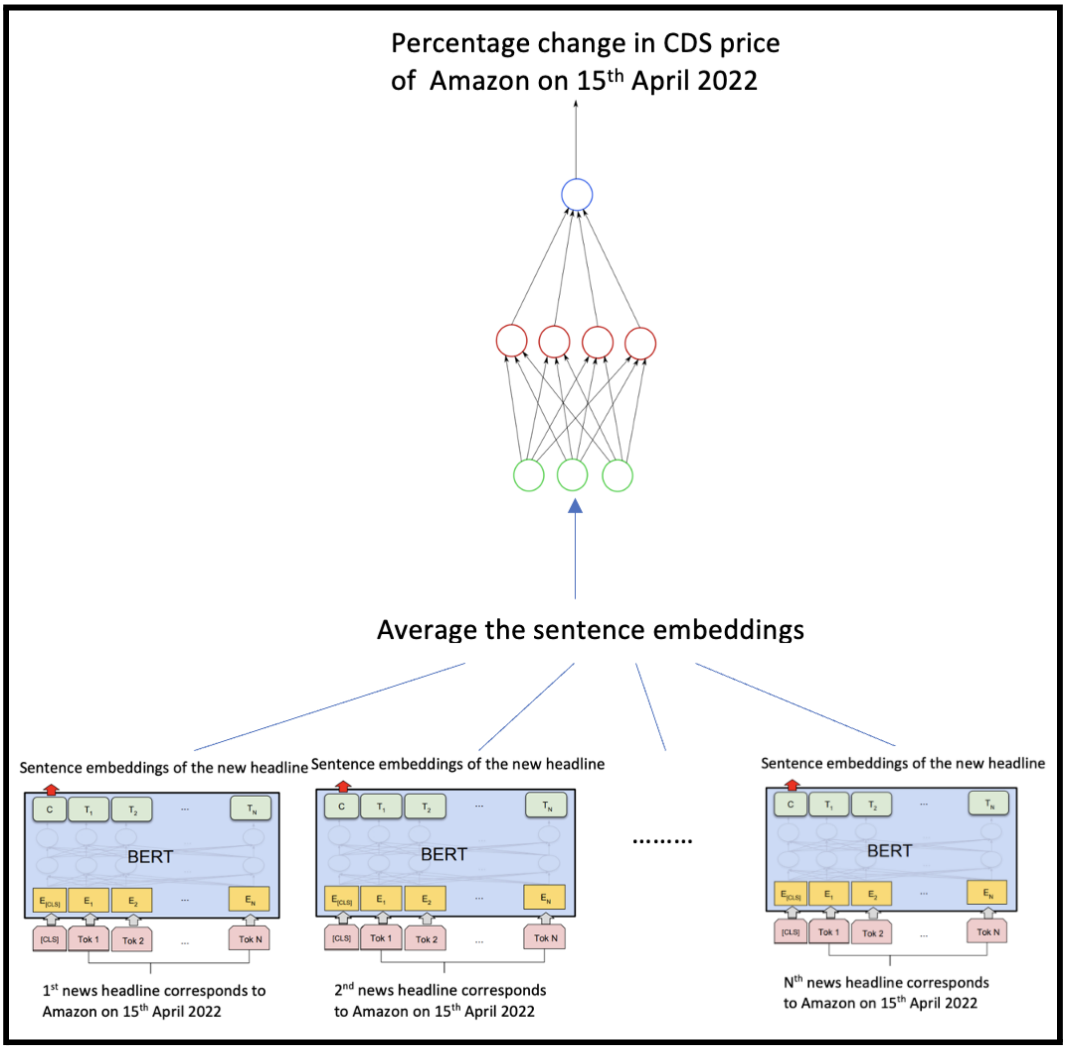


Fig 1. The structure of the forecast model that uses the sentences embeddings from FinBERT

2.1.7 Alert to counterparty

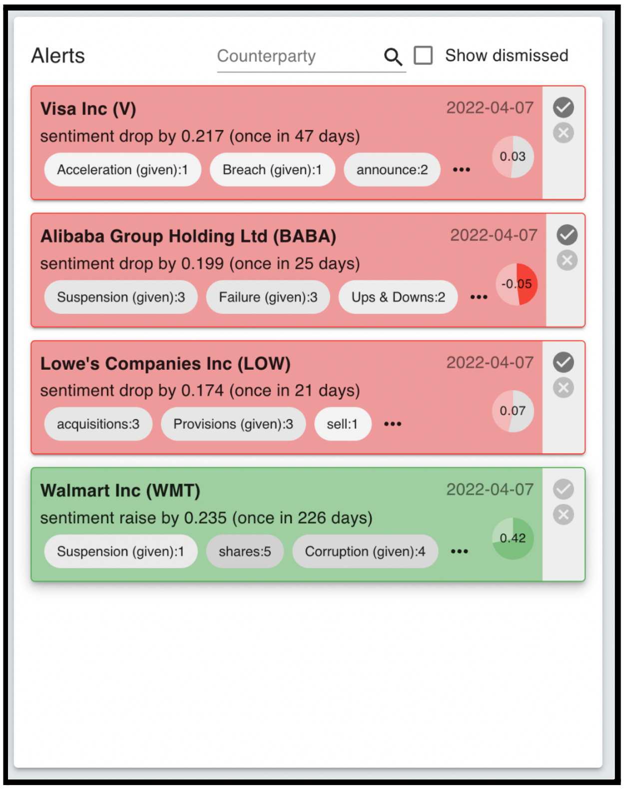


Fig 1. Alert cards on the dashboard

Fig 1. exhibit the early alert/warning dashboard for all the counterparties added to the system by the user. If a counterparty exceeds a certain sentiment threshold (defined below in 2.2.7), then an alert is shown on the dashboard indicating the sentiment drop/raise along with the flagged keywords as a notification to the user.

2.2 Implementation

2.2.1 User Interface Implementation

Based on our design, we used ReactJS as the frontend framework to develop the UI. On top of ReactJS, we used Axios as the interface with server API, fetching all necessary data to be displayed. On the other hand, Material UI is the key UI library we used, providing layout components including grids, navigation bar, buttons, and tables. Grid layout was used as the main layout model, as it supports responsive design, enabling adaptive UI in different screen sizes.

Detouring away from the original plan a bit, we replaced Recharts with Highcharts as the charting library, since the latter provides clearer visualization of multi-axis data and more sophisticated functionalities like zooming and panning than the former.

An overall picture of the frontend project structure is shown below.

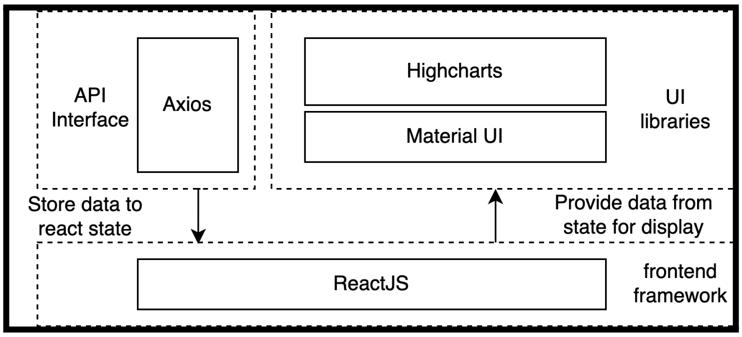


Fig 1. Libraries Used by Frontend

2.2.2 News Database Implementation

The data points for counterparties are collected from different news sources from two main sources namely Finnhub and Yahoo. Since Finnhub provides access to other sources such as Bloomberg, Reuters, et al., we did not find it necessary to purchase their APIs separately. Having said that, it is important to mention that Finnhub provides only limited data access from other sources but despite our constant struggles as mentioned in section 2.1.3 to get the free API access from HKUST Business School, we were not able to do so.

Once the news is pulled from Finnhub and Yahoo, we make use of a MongoDB that ticked all our requirements. As much as we would have liked to have all the features available for each of the data points, in practice, that could not be the case. For example, a credit risk indicator for one of the counterparties might be available but not for others. For this reason, a NoSQL database was the best for screening and saving. However, to ensure data completeness and integrity, we also developed a defined schema (refer to section 2.1.2 Fig. 1) that guarantees important features are stored correctly in the database.

2.2.3 News and Indicator Backend Handling Implementation

The backend is responsible for calling various external data source APIs, formatting the retrieved data, and saving the data to the news database. The external data source APIs included Finnhub for news collection and Yahoo finance for stock prices and time-series data.

It runs the Finbert model that takes the counterparty news to calculate whether a news article on any given day correlates to positive, neutral, or negative sentiment and finally stores it into the "calculation_result" collection.

The stock price and volume data are stored in the "time-series" collection for the users to validate the calculated sentiment against the real-time prices. For each news article, relevant event keywords are matched and tagged, keyword analysis is also done via LDA (more details on both in section 2.2.4).

To reduce latency, the calculation results are stored in the database, so that for each API call from the frontend, calculation results are obtained from the database cache instead of re-calculating from raw data. The external data retrieval and the calculation process are repeated every 24 hours through a cron job [cron.py]. Our API provides relevant calls to feed data to the front-end such as sentiment, keywords, news headlines, summaries, etc.

Below are the key indicators (discussed in detail in other segments) that the frontend displays using the data from backend database collections:

1. News Sentiment
    a. Positive
    b. Neutral
    c. Negative
2. Sentiment Weighted Moving Average (SWMA)
3. News Count - Per day
4. Stock Price
5. Keywords
6. LDA Topics

2.2.4 News Categorization Implementation

2.2.4.1 Categorization by Counterparty

The way we implemented categorizing news by counterparty was by calling an API known as Finnhub as outlined in our proposal. This API allows us to query news of a specific counterparty within a certain time period (using their stock symbol) which we cache into our database (schema outlined in section 2.1.2) with the appropriate tag. We chose this API as it was one of the best free options, with generous get request limits and the inclusion of a short summary of the article (unlike other APIs such as Yahoo Finance).

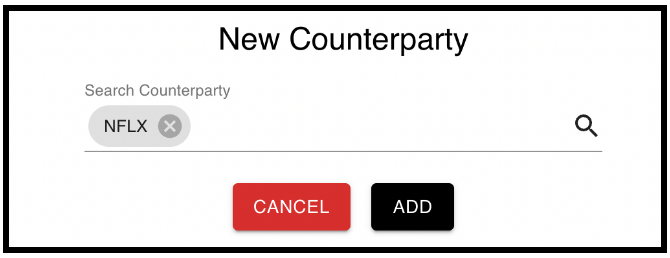


Fig 1. Adding Netflix as a New Counterparty

We ingest news of a certain counterparty whenever a user adds any counterparty for analysis. The user can do this with a simple hit and search. By specifying the stock ticker symbol of any counterparty (e.g. NFLX for Netflix Inc.), the news that counterparty in the past year is automatically extracted into our database via the Finnhub API (1 year is the limit imposed by the free tier). We then continually extract news for that counterparty as outlined in our "Backend Handling Implementation" section to keep the news up to date. The counterparty search function is done by connecting to another API route provided by Finnhub.

2.2.4.2 Categorization by Events/Topics

Part 1. Topic Identification Part

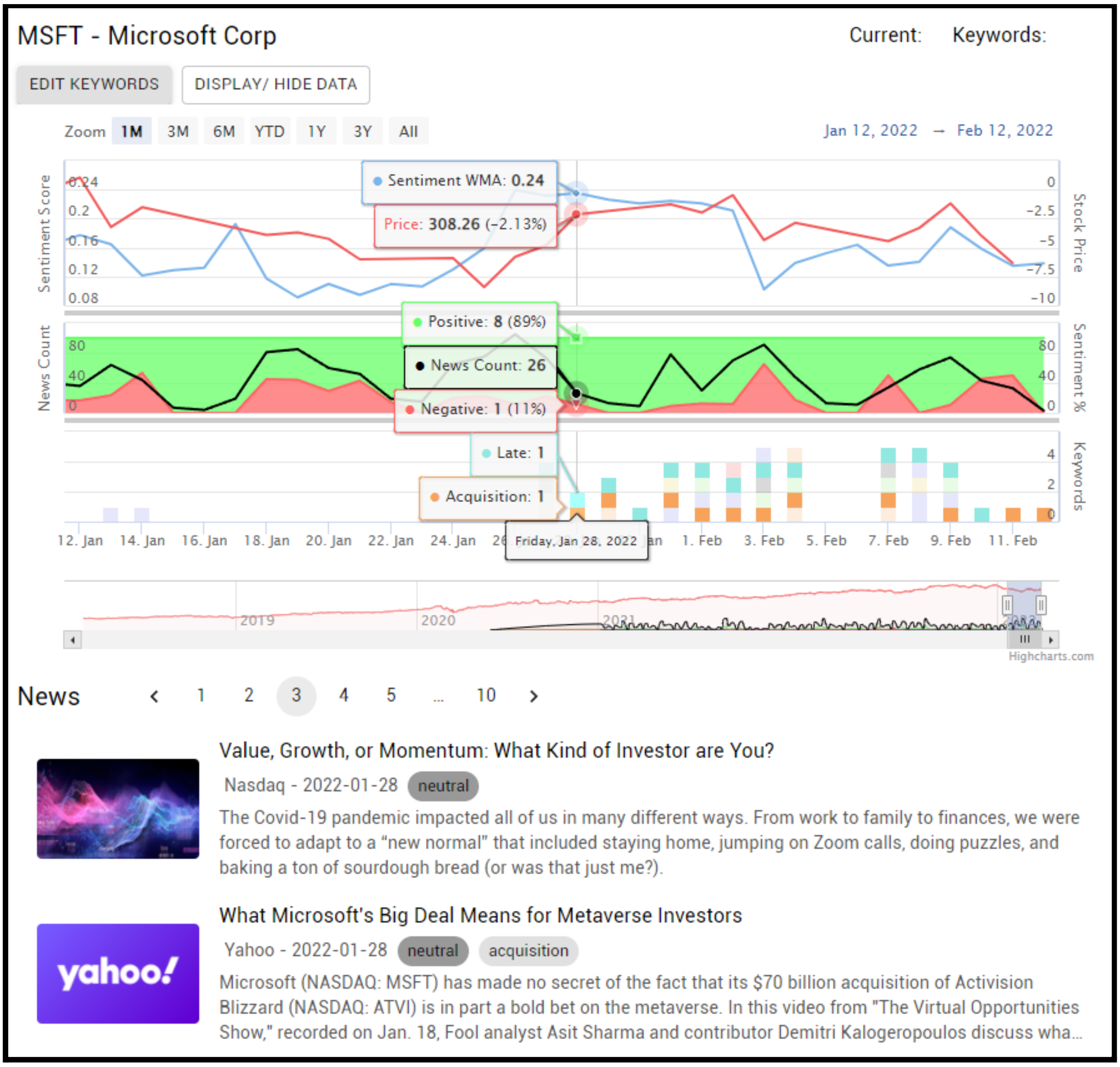


Fig 1. Screenshot of Our Counterparty Page for Microsoft for Jan 28
(with keyword “acquisition” tagged)

As mentioned in the design section 2.1.4.2, Societe Generale provided us with 30+ terms (see Appendix 6.3) that correlate to a counterparty’s credit standing and certain risk events (e.g. corruption, fraud). The role of the system is then to tag news articles with these related terms or “topics”, which we then display on the frontend as a count on the chart (as seen in Fig 1 above) amongst other uses.

Initially, we tried to tag articles via a pure keyword matching approach with the help of natural language libraries such as NLTK. We would lemmatize the keyword terms provided by SG and string match any articles containing any of the lemmatized forms of the terms (synonyms). For example, if the word bankrupt appears in the news, it will be classified into the category 'bankruptcy'.

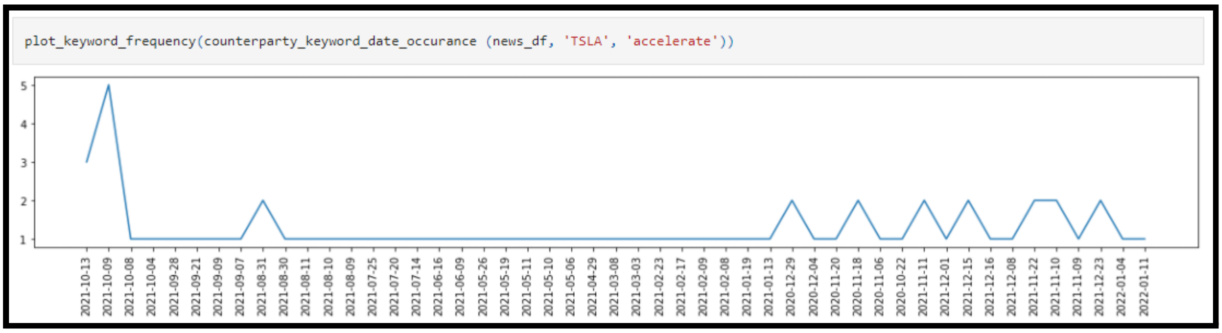


Fig 2. A Plot of the Number of Articles Tagged With Keyword "accelerate" Over Time for Tesla

However, we found that the frequency of articles being tagged with these keywords given by SG (as seen in Fig 2 above) to be actually very low (even with lemmatization enabled, which would already often overproduce not so relevant lemmas in the first place). To the point where these tags are not useful as time-series, as they cannot show any kind of trend, only blips of rare occurrences.

Therefore, through discussions with Martin Autier (a Data Scientist from SG), it was decided we should try some more advanced/less naive methods for tagging news with the concerned events. A method we experimented with was to utilize the FinBert neural network (our sentimental model, discussed further in section 2.2.5) to attempt to cluster articles similar to the ones identified by the naive method to increase the hit rate. This allows hits from the event identifier to be used as time series depending on the similarity threshold set. It also allows users to track broader relevant topics, relating to the theme/intention behind the user's keywords, instead of just direct matches. For example, if a user was tracking the keyword "sued", they probably wanted/want to know about legal events (such as lawsuits) that involve the counterparty. Instead of the user having to come up with a list of legal keywords to track, this method attempts to alleviate that work for them by tracking related articles.

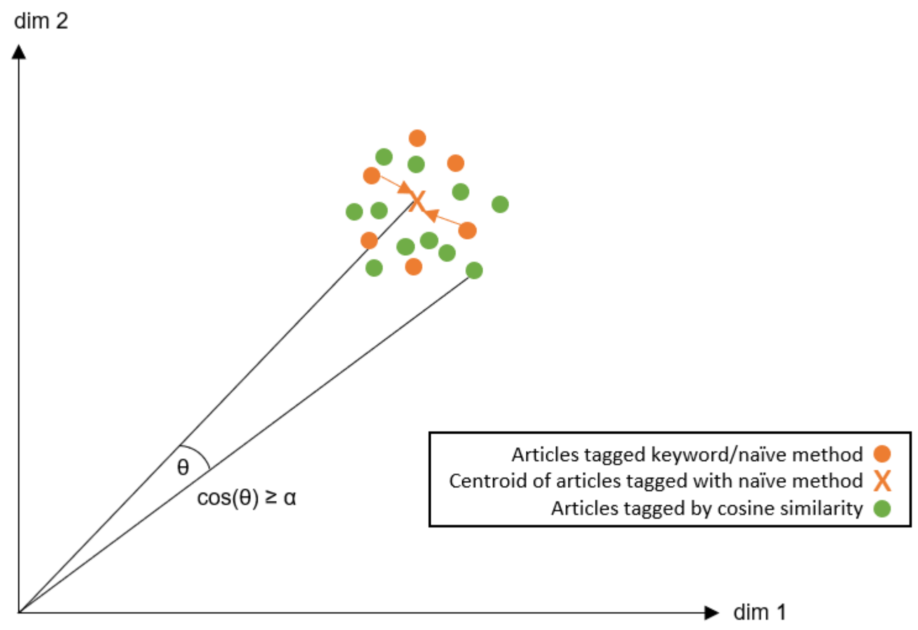


Fig 3. Simplified Visualization of the Second Method
(with vectors of just 2 dimensions, full similarity equation in Appendix 6.4)

This clustering/tracking method was achieved by feeding the articles tagged with a relevant keyword(s) for a topic (e.g. "fraud" using the naive method) to FinBert for inference. Then we take each article's last hidden layer value as an embedding, average them out across the tagged articles, to create a final centroid vector/embedding for a keyword/topic. Other articles are then compared to how close/similar (using cosine similarity) it is to a topic's centroid and given a similarity score from 0 to 1. A hyperparameter α can then be set to determine how high of a similarity score an article must have to belong to that topic (e.g. $\alpha \geq 0.7$). This is the general overview of this method. Essentially, we used FinBert to plot articles (tagged with the naive method) into a vector space, then we find the center point of the articles, and lastly, we tag articles close/similar to those center points with the topic (see above Fig 3 for simplified visual).

Parameter Experiments and Selection

On the layer choice for the embedding, we chose the last layer as our vector space to cluster with because it is linearly separable, meaning the articles that carry similar sentiment (or even meaning) should already be relatively clustered. This is thanks to the classification layer right after it in FinBert being linear (see original code). We also ran some brief tests on using other layers as our vector space. As seen from the output below (Fig 4), in other layers, the separation in similarity score from similar and different sentences is just not as pronounced as in the last layer.

```
Target Sentence: "growth is strong and we have plenty of liquidity"
Sentence with Similar Meaning: "growth is strong and we have plenty of liquidity now"
Sentence with Different Meaning: "growth is weak and we have plenty of liquidity"

Layer -1
[0.96982855] (Similar Sentence Similarity Score)
[0.43475103] (Different Sentence Similarity Score)

Layer -2
[0.96304333] (Similar Sentence Similarity Score)
[0.6460768] (Different Sentence Similarity Score)

Layer -3
[0.96068895] (Similar Sentence Similarity Score)
[0.836292] (Different Sentence Similarity Score)

Layer 0
[0.97492874] (Similar Sentence Similarity Score)
[0.99377394] (Different Sentence Similarity Score)

Layer 1
[0.977111] (Similar Sentence Similarity Score)
[0.9914181] (Different Sentence Similarity Score)
```

Fig 4. The Output of Cosine Similarity Testing on Different FinBert Layers

Moving on to the choice of dimensionality, in this experiment, we chose to set our FinBert model to have a max token length of 64. This was done for performance reasons as more than 99% of our headlines had 32 words or less. Reducing the token length and therefore vector dimensions allows for faster inference, faster similarity computation, and less memory to be used during runtime and potential storage.

Implementation in Live System

For the final implementation in our system we do not use the full 49152 (768*64 tokens) long vector (or hidden layer) to do the cosine similarity topic matching/scoring. Instead, we average out the vector by token and save a 768 long vector instead, which is the “average” of all 64 tokens. Basically, we use a shorter embedding that represents the average token, instead of one that “spells” out the whole sentence by concatenating each token into an extremely long embedding. This shorter vector/embedding allows us to save space when storing it into the database as well as reduce the computation cost when computing cosine similarity for topic tagging (since now you are only comparing 768 numbers).

These storage and computation savings allows us much more flexibility on what we can achieve with these embeddings. Firstly, it allows us to store all of the computed embeddings of all the news articles in the database directly without much concern (the final news collection of 350k news including summary and other metadata was only 3.4 GB). This then allows us to compute new topic embeddings and other visualizations in near-real-time (a few seconds) on our fairly weak dual-core server CPU. Since creating a new topic embedding just means querying articles with the correct keywords (via regex) and averaging their already computed embedding. This faster responsiveness helps enhances the user experience and allows them to experiment with different topics.

Without these optimizations, we would not be able to store the embeddings directly on our storage-limited database, meaning every time anything in our system interacts with the news embedding we basically have to run the expensive FinBert inference again. This is particularly infeasible when it comes to tagging old articles with a new topic embedding because it would mean having to infer your whole database through the FinBert AI again. Our smaller 768 vector allows us to store the computation result in our database as a cache and also reduces the number of values to compare in the cosine similarity calculation.

In terms of quality degradation from averaging the vector from 49152 values to 768 for topic matching, it was very minimal and captured most types of articles the original matched, albeit a bit less accurate. As a team, we thought it is a reasonable tradeoff given the massive computation savings generated. We also tried other ways of reducing the vector size such as max pooling but found the result to be much worse. Lastly, we also tried to include the articles containing synonyms generated from NLTK into the topic embedding generation. However, we found that the increased articles into the topic embedding calculator just introduced more input noise rather than improving output quality so it was forgone as an option. A sample of the outputs of the various attempted configurations is included in Appendix 6.11.

A further extension that could be done once the system is deployed and amasses a significant amount of users is to train an attention-based mechanism (instead of averaging) for the vector length reduction based on collected user feedback regarding the topic tagging. The general idea would be to collect negative and positive match signals from the user and then backpropagate that signal back to the reduction system where the weighting mapping from 49152 to 768 will be adjusted. The result map will have non-equal weights (unlike normal averaging) and should improve performance.

Part 2. Topic Suggestion Part

In order to provide a quick summary of the news and save users' time, we implemented a topic suggestion model via Latent-Dirchilest Allocation aka LDA model as mentioned in the design section.

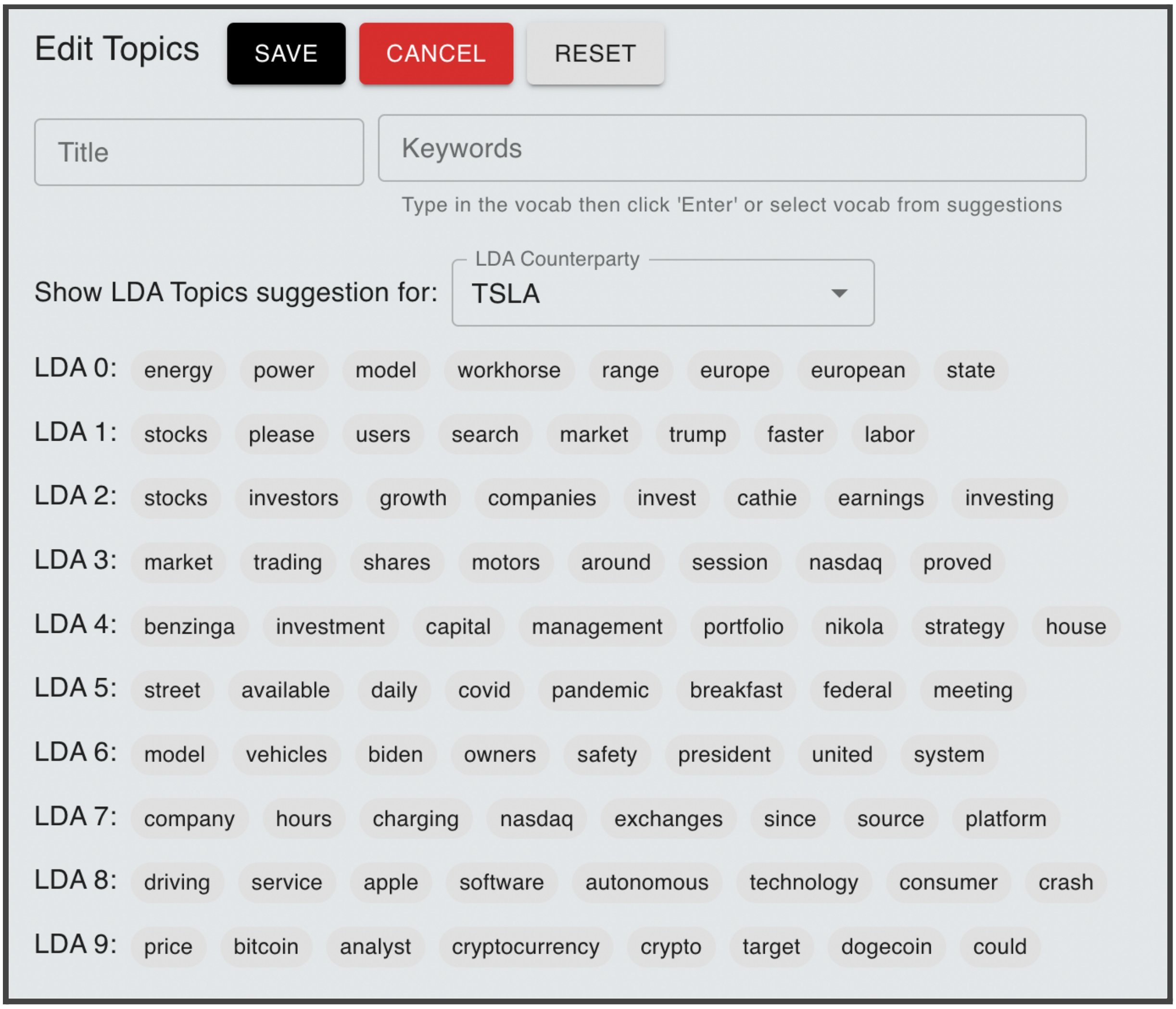


Fig 1. LDA suggestive keyword topics for TSLA (LDA 9 suggests "crypto")

To analyze a counterparty's news, a summary of the news is produced with the help of 10 LDA suggestive topics defined by the Latent Dirichlet allocation (LDA) algorithm. For each topic, similar news articles are clustered together and the most common keywords appearing in that cluster are combined together to form a topic that is later to be deciphered by the user. The precise mathematics is presented in Appendix 6.5.

As an example, after introducing the news segment for the counterparty, Tesla, to the LDA model, it produced a combination of keywords that best describe the topic of the news articles. The number of topics can be custom picked by the user themselves. Although this functionality is only available in our backend system since we have set a hard limit of 10 topics per counterparty. Fig.1, Topic 9, and Fig. 2 Topic 2 generated for Tesla on 20 Apr 2022 and 2 Jan 2022 suggest a correlation between Tesla and cryptocurrencies such as bitcoin and dogecoin. Anyone merely interested in Tesla's CEO, Elon Musk, would know he is highly involved in the crypto world and the LDA model is clearly able to point this out. Based on this and the news sentiment on those articles, users can map the negative or positive relationship between the two i.e. Crypto and Tesla. This is a familiar scenario to indicate the direct relevance and importance of keywords produced by LDA.

Echoing on the above statement, similar to Tesla, users can also look at niche or bizarre counterparties, and with the help of the LDA topic modeling, they can find the correlation of that counterparty with other industries and if that will impact it positively or negatively.

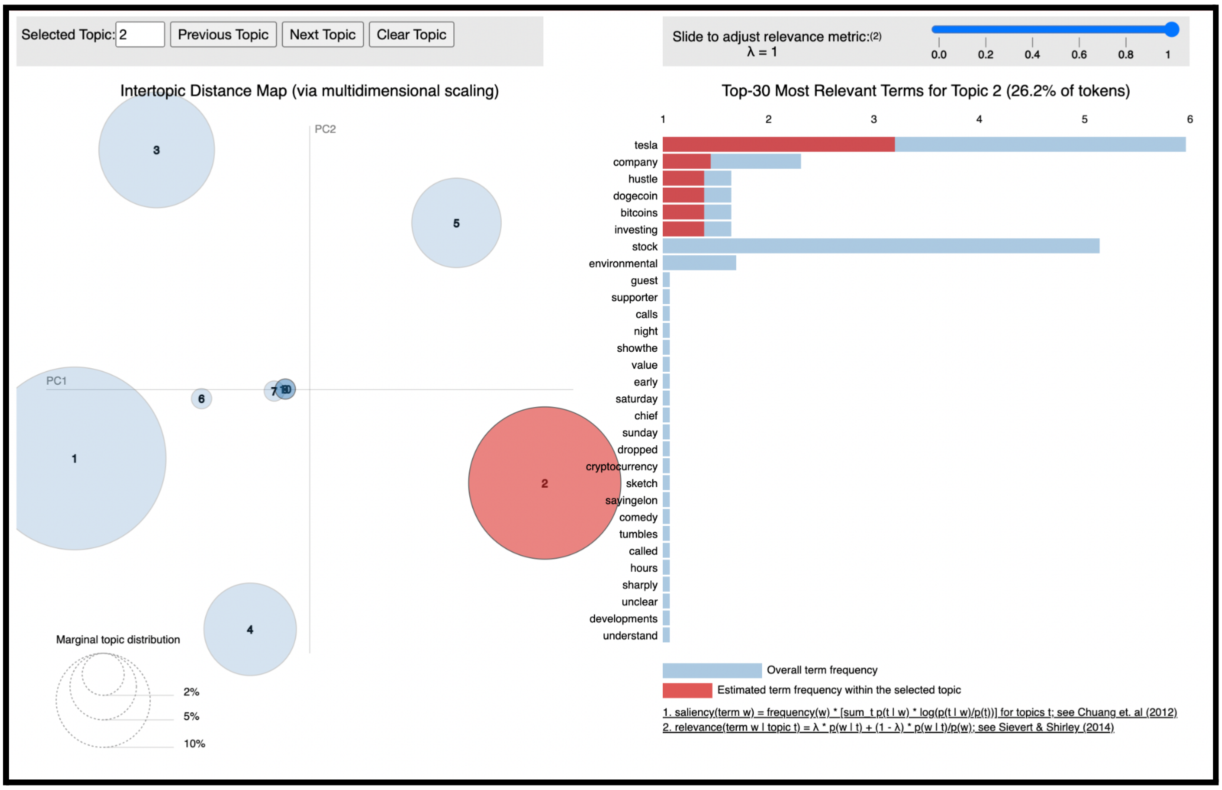


Fig 2. LDA Model on Tesla with 8 topics
[Topic 2: "Tesla company hustle dogecoin bitcoins investing stock environmental"]

Fig 2 depicts the importance of the keywords and their frequency in any given topic on the right panel whilst annotating the difference between different topics on the Intertopic Distance Map. The right panel depicts a horizontal bar chart whose bars represent the individual terms that are the most useful for interpreting the currently selected topic on the left. A pair of overlaid bars represent both the corpus-wide frequency of a given term as well as the topic-specific frequency of the term.

The left panel visualizes the topics as circles in the two-dimensional plane whose centers are determined by computing the *Jensen–Shannon divergence between topics* [Appendix 6.6], and then by u*sing multidimensional scaling* [Appendix 6.7] to project the inter-topic distances onto two dimensions. Each topic's overall prevalence is encoded using the areas of the circles.

2.2.5 News Sentiment Model Implementation

There is a Jupyter notebook in the repository for the FinBERT developed by Y. Yang, M.C.S. Uy, and A. Huang, we converted the code inside the Jupyter notebook to the code in a python file. We integrate the codes and the model (fine_tuned.pth) file to the backend. Every time there is updated news, cron.py will automatically scrape the news from Finnhub and automatically infer the news headline to output the sentiment. The sentiment of the news is uploaded to the database, and displayed in the frontend of the system.

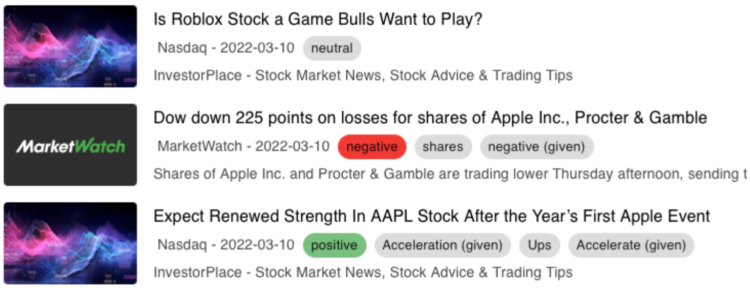


Fig 1. News Displayed as 'Positive', 'Neutral' and 'Negative' Respectively in Frontend

We cannot directly display the sentiment score of the news in one single graph, as there are large amounts of financial news and it is infeasible for the credit risk team of Societe Generale to monitor the sentiment score of each news. We group the news sentiments by each counterparty and by each date, obtaining a counterparty sentiment score through averaging.

We tried different averaging methods, including simple average and Exponential Moving Average (EMA). Yet, simple average over-fluctuates and is especially unreliable on days with few news, whereas EMA is over-smooth and unable to capture abrupt changes. To capture 'sudden events', there is a need to emphasize days with high news count. Referencing the volume weighted techniques commonly used in stock price calculation, we discovered a rarely used indicator, Volume Weighted Exponential Weighted Moving Average (V-EMA).

V-EMA is a combination of EMA and VWMA, incorporating the time decay weight (a^{T-t}) from the former and the volume weight (V_t) from the latter. In our case, we substitute the volume weight with the daily news count. The formula of V-EMA sentiment score is described below,

$$S_T = \frac{\sum_{t=0}^{T} \alpha^{T-t} V_t s_t}{\sum_{t=0}^{T} \alpha^{T-t} V_t}$$

α : time decay factor
s_t: average score at day t
V: volume (news count)

As shown in Figure 2, the result generated by V-EMA is smoother in normal days and hence more interpretable than simple average. While compared to EMA, V-EMA preserves the ability to capture abrupt change in sentiments during days that news count surge (marked in red in Figure 2) and hence being able to trigger alerts.

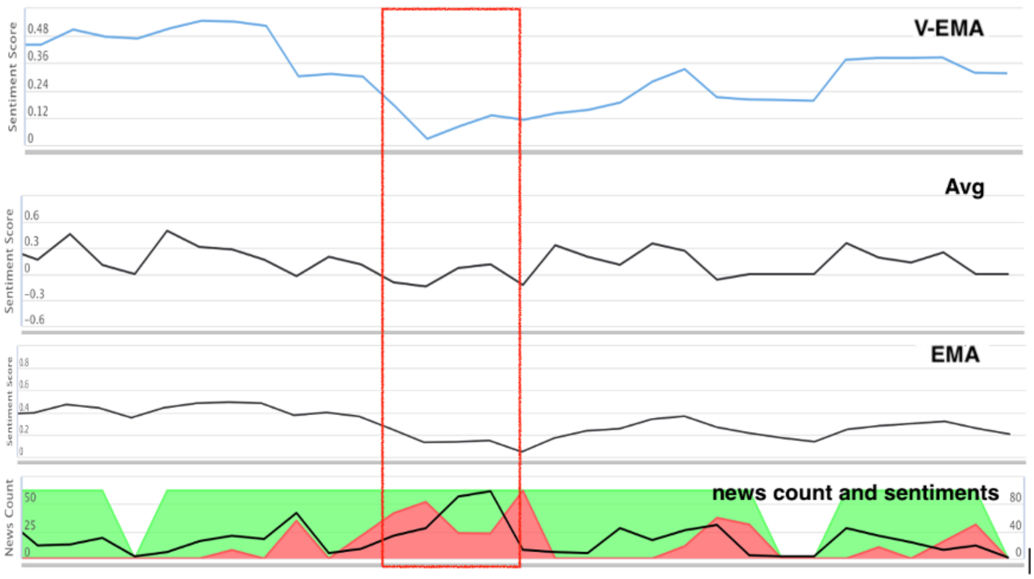


Fig. 2. Comparison between V-EMA, simple averages and EMA

The implementation of V-EMA sentiment aggregation method is available in ***'calculations/aggregate.py'***.

2.2.6 Forecast Model Development

2.2.6.1 Fine-Tune FinBERT

Dataset

First, we need to decide on the news dataset and the CDS dataset for the model training. The news in Finnhub has a limited period of time, only recent news can be scraped. Therefore, we found a dataset of Daily Financial News with 6000+ Stocks (2009-2020), with 4 million articles contained in it. For the dataset of the CDS, we found another dataset called Credit Default Swap (CDS) Prices in Kaggle, with more than 600 counterparties and in a time series of 6 years (2015-2021). Then, we need to use python to calculate the percentage change of CDS for the dataset. We then merged the news datasets and the CDS datasets by the counterparties and the date. A new dataset with news headlines and percentage change in CDS is formed.

The merged dataset has 104981 rows of data, and the data is collected over the period of January 2015 till June 2020. This dataset is then split into training sets, validation sets, and testing sets according to time. The training set is under the period from January 2015 to June 2019, with 80797 rows of data. The validation set is under the period from June 2019 to April 2020, with 24184 rows of data. The testing set is under the period from April 2020 to June 2020, with 9043 rows of data. Due to the limited power of the GPU and the large scale of the model, we only take parts of the data to train. We used 10000 rows out of the training data to train and 2000 rows out of the validation data to validate the model. We used 1000 rows out of the testing data to test the performance of the model. The data inside each dataset is then shuffled to make it fair.

For each row of training data, we have the headline of a news, and the percentage change in CDS price on the date of the news released of the counterparty

title	date	Percentage change in CDS	stock	Company
Jim Cramer Gives His Opinion On Humana, Alibab...	2015-08-06	0.001042	AET	Aetna Inc
U.S. Army Gives Green Light to General Dynamic...	2015-06-08	0.034086	GD	General Dynamics Corp
Shares of several technology companies are tra...	2019-05-29	-0.064325	IBM	International Business Machines Corp
Google Cloud Platform Signs Home Depot as Clie...	2016-03-22	0.005032	HD	Home Depot Inc The
The Lucky Beneficiaries Of Tax Day	2017-04-20	-0.001756	NOC	Northrop Grumman Corp

Fig 1: The merged dataset with news title as the input and percentage change in CDS as the output for the forecast model

Implementation of model structure

To fine-tune the FinBERT model with the news headlines as the input and percentage change in CDS price, we need to change the last linear layer of the FinBERT model to only 1 output target (Percentage change in CDS) instead of 3 output targets originally (positive, negative, neutral), and change it to a regression model from a classification model. This can be done by changing the loss function to mean squared error. To optimize the performance of the training, we used Adam optimizer to perform the training of the model.

For the training part of the model, in the first 5 epochs, the validation loss drops significantly from 0.06 to 0.02, in the next 5 epochs of the training, the drop of the loss remains saturated at around 0.02. Therefore, we trained for 10 epochs with around 4 minutes per epoch.

2.2.6.2 Using sentence embeddings

Dataset

We utilized the dataset made from 2.2.6.1. There are 104981 rows in total, each row has the news headline, counterparty, date, percentage change in CDS price. We feed all the news headlines to the sentiment model, then the model outputs a sentence embedding. We average out all the sentence embeddings relevant to a specific counterparty on a specific date. Each sentence. Each averaged sentence embedding will act as an input and the percentage change of CDS price relevant to the counterparty on the date when the news release is the output.

Similarly, the dataset is splitted into training dataset, validation dataset and also testing dataset. We use the dates to segments between the datasets as it prevents the model from forecasting the result when doing the validation. The training data is from January 2015 to January 2019, it contains 36632 rows. The validation data is from January 2019 to September 2019, which contains 8284 rows of data. The testing data is from September 2019 to September 2020, which contains 12875 rows of data.

Implementation of model structure

Comparing the approach in FineTuning the FinBERT (2.2.6.1), there is less amount of training data as we average out the embeddings. These embeddings are put into a neural network for training. Each sentence embedding has a size 768. We use three layers of neural networks: layer 1 has dimension of 768X128, layer 2 has dimension of 128X32, the last layer has a dimension of 32X1. The final output will be the percentage change in CDS price of the counterparty relevant to the news and the date when the news is released.
The training is much faster then before as the embeddings now only need to go through a neural network. As this is a regression problem, we used mean squared loss and the learning rate of the model is 0.001. As the training time is slow with around 1 second per epoch in GPU, we trained for 100 epochs, in total around 100 seconds.

2.2.7 Alert Generation

We used quantile score as a metric for our alert categorization system. The inputs are the daily sentiment score obtained with the V-EMA method (denoted as $\boldsymbol{S}$) and the differential mean score (denoted as $\boldsymbol{D}$), computed as the difference between the scores of the current day corresponding to an average score of past 1, 2, and 4 days.

$$D_t = S_t - \frac{S_{t-1} + S_{t-2} + S_{t-3}}{3}$$

Considering the potentially non-normal distribution of $\boldsymbol{S}$ and $\boldsymbol{D}$, we use the non-parametric order statistics method (formulas provided in Appendix 6.8) and obtain the quantile of the above-computed array ($\boldsymbol{q_D}$) and quantile of that day's sentiment score ($\boldsymbol{q_S}$) with respect to previous days. With the computed quantiles, two-sided alerts (sentiment rise/ sentiment drop) are generated if the quantile score locates at two tails of the distribution, with a cutoff of differential score (α_D) as 0.05 as the cutoff of sentiment score (α_S) as 0.1.

Sentiment Drop Alert if,

$$q_D < \alpha_D = 0.05$$
$$q_S < \alpha_S = 0.1$$

Sentiment Rise Reminder if,

$$q_D > 1 - \alpha_D = 0.95$$
$$q_S > 1 - \alpha_S = 0.9$$

The implementation of the alert generation model is available on '***calculations/alertGenerator.py***'.

2.3 Testing

During the development process, unit testing was done to ensure modules were functioning correctly. Integration testing of the various components of the system was also done on a regular basis, since we adopted an agile methodology and would deploy different versions of our application live unto our deployment server and website (https://fyp.cslix1.ml/) on major feature changes.

The following components will be/have been tried and tested by a human manually:

1. User Interface Testing
2. Database Testing
3. News and Indicator Backend Handling
4. News Categorization Testing
5. News Sentiment Model Testing
6. Forecast Models Testing
7. Alert Testing
8. User Acceptance Testing

2.3.1 UI Testing

To test the User Interface, we

1. confirmed the UX behavior is in accordance with the designed UI flow
2. confirmed the data displayed is consistent with the backend
3. confirmed the UI displays error messages when the backend server returns certain unsuccessful responses
4. confirmed the layout is usable on a mobile screen

2.3.2 Database Testing

To test if the database is working as required, we have confirmed the data is properly stored and extracted in the database. This includes,

1. confirm that all news contains the required fields in the correct data type
2. confirm that news are not duplicated in the database
3. confirm that all time-series data contains the required fields in the correct data type
4. confirm that time series data are not duplicated in the database
5. confirm the calculation result stored in the database are updated and integrated

2.2.3 News and Indicator Backend Handling

To test the backend handling, we have been checking the correctness by monitoring the server logs on a regular basis. We have confirmed that

1. Counterparties can be added or deleted correctly
2. News and time series data are fetched automatically and correctly after the addition of new counterparties
3. News and time series data are updated as designed everyday through the cron job
4. Calculations are refreshed without error through the cron job

2.3.4 News Categorization Testing

For testing counterparty categorization, since Finnhub already provides the counterparties relevant to the news (why we chose it as a ingest source in the first place), not much testing should be needed. However, we did do online due diligence research on the quality of the categorization, as well as monitor the quality of the categorization in our day-to-day usage and during demos to SG. Furthermore, to fully make sure the categorization is up to standard, we sampled our news database (see Appendix 6.11) and try to score how accurate the counterparty tagging was from Finnhub. The results of the above tests will be discussed in relevant evaluation section 2.4.4.

For testing the performance of topic categorization, specifically the new clustering method, we made a Google form (https://forms.gle/ZjtoV8xBsMWyoEwu9) to ask users for the degree of relevance between the keywords and different news headlines of various cosine similarity scores. This is done to test if there is a real correlation between cosine similarity scores produced by our method with "human-perceived" similarity, which should the "optimal baseline". The survey asks participants how similar headlines are to a specific keywords from a 1 to 5 scale, which are them mapped back to the similarity scores for evaluation. A 5 point Likert scale was chosen because of its wide use in psychology survey research [17], and its ability to ensure distances between each choice is equal. Results, details and evaluation of this survey will also be discussed in relevant evaluation section 2.4.4.

2.3.5 News Sentiment Model Testing

For the testing of the FinBERT model, it was already conducted thoroughly by the authors of the model, Y. Yang , M.C.S. Uy, and A. Huang experimented on three different datasets that are prepared for financial sentiment classification. The first one is Financial Phrase Bank, which is a public dataset containing 4840 sentences and labeled by researchers in the financial market. The second dataset is the Analyst Tone Dataset, which is extracted from Kaggle and contains 1000 sentences. The FiQA dataset is an open challenge for financial sentiment analysis with 1111 text sentences. The dataset is randomly split into 90% training and 10% testing. It is then being compared with BERT. The result of the testing is put in 2.4.5 in the evaluation part.

2.3.6 Forecast Models Testing

Unlike classification problems, regression problems like this use mean square error as the loss functions; there are no metrics like f1-score to test the performance of the model. On top of the mean square error loss, we also introduce mean absolute error to track the performance of the model. To visualize how well the model predicts the percentage change in the CDS price of the sentence, we plotted two graphs to compare the prediction and the target across a period of time. The result of the testing is put in 2.4.6 in the evaluation part.

2.3.7 Alert Testing

This part had been subjectively analyzed by the team to estimate its accuracy and the results were presented to Mr. Martin Autier and Mr. Ray Wong from Société Générale. After the satisfactory demonstration, the results from the Alert Generation model as defined in section 2.2.7 were passed. A further illustration of the results through examples can be found under section 2.3.7. Those demonstrated alerts in section 2.3.7 indicate the directly proportional relationship between a counterparty's stock price in conjunction with a major incident and counterparty's sentiment drop or rise as shown in the alert. To our estimate, 3 in 5 alerts are backed with a strong relationship of sentiment drop/rise which is caused by a major market event.

2.3.8 User Acceptance Testing

We made a google form to ask questions regarding how well the system performed (https://docs.google.com/forms/d/1yKO9zduF6haIQU9_I3XCtmF7Z49WSWk7JTsueoUAb6A/viewform?edit_requested=true). This is the link to the survey. As the users of the system are the credit risk team, the questions are optimized for the user experience for them.

We asked the following questions:

1. What is your occupation?
2. Is the system user-friendly and not confusing?
3. How would you rate the charts that display the sentiment scores, keywords and also stock price?
4. Is the alert feature clearly shown?
5. Can the system improve the efficiency of the market risk team?
6. What would you like to see in the system that is not there?
7. Any improvements on the system?

For this user acceptance testing, we asked users coming from different backgrounds, including people working in Société Générale, software engineers in financial firms, computer science students, and business students from HKUST. As many of them don't have enough financial knowledge, we explained the system to them before they took the survey. The evaluation of the collected feedback is shown in 2.4.8.

2.4 Evaluation

After we have finished all the testing, we will ask people from Societe Generale's Data Science Team and their credit risk team for their feedback on the following questions. Since we do not have direct interactions with the end-user, we will also get feedback from ordinary people with no knowledge of Computer Science and NLP to check the feasibility of our design and implementation.

1. Is the UI distinctive enough to remind the credit team of important alert events?
2. Is the UI intuitive enough for users to input and modify their desired counterparty?
3. Is the database properly loaded and pipelined with the back and front end?
4. Is the API able to autonomously load data from web sources without server errors?
5. Can the system accurately sort news by the counterparty and events?
6. Is the sentiment score of the sentiment model accurate to human intuition?
7. Does the correlation model provide useful information for managing a portfolio?
8. Is the system able to alert Risk Analysts to potential risk-inducing events that may affect their portfolios?

We have been conducting regular meetings with people from the Société Générale Data Science team and collecting their feedback on these questions. This feedback has been positive and has been great help at continually evaluating and improving our system.

2.4.4 News Categorization Evaluation

As mentioned in testing section 2.3.4, we sampled news from our database to test & evaluate the quality of the counterparty categorization. We found that ≥85% of the counterparties tagged were mentioned in the article as key subjects and ≥65% were the sole focus of the articles. This shows that the counterparty categorization source we chose is of high quality and should have low enough noise to produce good results further down in the pipeline (which we have seen).

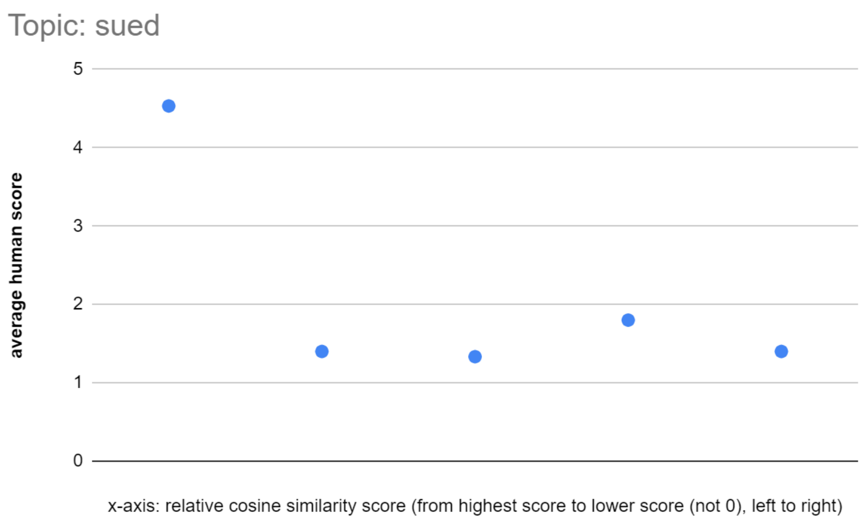


Fig 1. Plot of Average Human Score vs. Cosine Similarity Score for Topic "sued"
Notice the big separation in human score for high cosine similarity which we used to categorize topics.

Also, mentioned in testing section 2.3.4, we did a survey consisting of 5 randomly selected topics and a total of 75 headlines in an attempt to figure out if there was a correlation between the concise similarity score (used for our topic categorization) and "human-perceived" similarity. The results showed a clear downwards trend of scores given by the humans as cosine similarity dropped, confirming a correlation between the concise similarity generated from our method and "human-perceived" similarity. The complete results of the survey with all the data is in Appendix 6.12 and a plot of one of the topics is shown above in Figure 1. Overall, the results confirm our method for topic categorization, it shows that it is sound and of good quality, as well as grounded in human truth. It also shows there is an emergence of topic understanding in the model without the need for explicit topic labels to train on, which is extracted with our method.

2.4.5 News Sentiment Model Evaluation

For all the three datasets suggested above (PhraseBank, FiQA, Analyst Tone) FinBERT performs much better in terms of the accuracy of the model, with FinVocab or BaseVocab, cased or uncased. Overall, FinBERT in FinVocab and Uncased have the best results in all three datasets, with an accuracy of 87.2% in PhraseBank, an accuracy of 84.4% in FiQA, and also an accuracy of 88.7% for the AnalystTone Dataset.

	BERT		FinBERT-BaseVocab		FinBERT-FinVocab	
	cased	uncased	cased	uncased	cased	uncased
PhraseBank	0.755	0.835	0.856	0.870	0.864	**0.872**
FiQA	0.653	0.730	0.767	0.796	0.814	**0.844**
AnalystTone	0.840	0.850	0.872	0.880	0.876	**0.887**

Table 1. compares the accuracies of the BERT model and the FinBERT model in all three financial sentiments dataset

Next, we tested the correlation of the counterparty sentiment score generated by E-WMA method with the stock price of the corresponding company. We compare the mean change of sentiment score in 1, 2, and 4 days (denoted as $D = S_t - \frac{S_{t-1}+S_{t-2}+S_{t-4}}{3}$) with the intraday price change of stock price (denoted as ΔP). The sample size consists of 73 counterparties with in total of 27068 days of results.

As provided in the first confusion matrix in Table 2, we discovered that only 56.94% of days that the sentiment score changes in the same direction as the stock price, indicating a low overall correlation of the two-time series data at an intraday level. Yet, if we consider the days when stock price move signification (exceeding the quantile level of 0.05), it is discovered that 84.37% of cases the sentiment score follows in the dropping case and 83.6% of cases the sentiment score rises together in the rising case (shown in second and third confusion matrix of Table 2 respectively).

	$D \geq 0$	$D < 0$
$\Delta P \geq 0$	0.2889	0.2092
$\Delta P < 0$	0.2214	0.2805

	$q_D \geq 0.05$	$q_D < 0.05$
$q_{\Delta P} \geq 0.05$	0.9423	0.0078
$q_{\Delta P} < 0.05$	0.0078	0.0421

	$q_D \geq 0.95$	$q_D < 0.95$
$q_{\Delta P} \geq 0.95$	0.0418	0.0082
$q_{\Delta P} < 0.95$	0.0082	0.9418

Table 2. normalized confusion matrix of stock price and sentiment score in signed, lower tail and upper tail cases (left to right)

The observation would suggest that during most of the days, the sentiment score is fairly independent of the stock price. The two indicators present high correlations only during days with sudden events that cause both to move significantly. It matches our impression that the sentiment score of many counterparties tracked in our system shares similar trends with stock price (Fig.1)

Fig 1. Comparison of past 6-month sentiment score with the stock price of Apple, Intel and Bank of America (top to bottom)

We would like to conclude our sentiment score indicator to be efficient, with the fact that it shows appropriate independence from the stock price in normal days (otherwise would mean that the score is redundant) while at the same time able to capture important price fluctuations. The evaluation result also justified our choice of parameters in our alert generation model (2.2.7).

2.4.6 Forecast Model Evaluation

2.4.6.1 Fine-Tune FinBERT

For this model, the validation mean square error drops from 0.06 to 0.02 through 10 epochs of training. As each headline is mapped to a percentage change in CDS, we barely cannot compare it to the real target of daily percentage change in CDS. In order to visualize the result, we average out the prediction of percentage change in CDS price inferred from all the sentences and compare it to the target CDS percentage change daily. Figure 1 compares our result between the prediction and the target of Amazon in May. We observe that the CDS percentage change is not as stable as the stock price. On 30th May 2020, there is a sudden surge that the percentage change in CDS increases by 15%. Our fine-tuned model is not able to capture this surge. Overall, our prediction is too stable and does not provide not enough fluctuation.

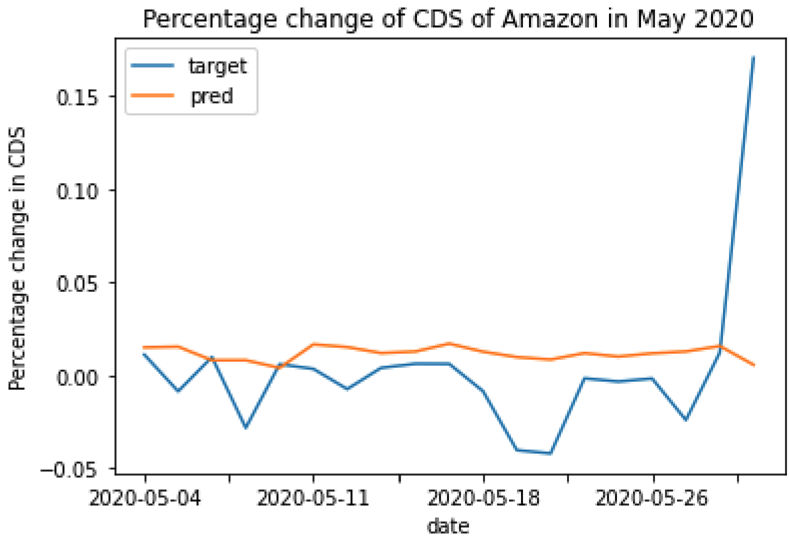


Fig. 1. shows the comparisons of our prediction and the real result on the percentage change of CDS in May using Fine-Tune FinBert Approach

2.4.6.2 Using sentence embeddings

Unlike Fine-Tuning FinBERT, using sentence embeddings will be much faster to do training. Through 20 epochs of training, the mean square error drops from 0.25 to 0.17. We compute the same graph to compare the prediction and target of the percentage change of CDS of Amazon in May 2020. This cannot capture the surge on May 30. However, this model fits the target better and follows the general trend more than the Fine-Tune FinBERT approach.

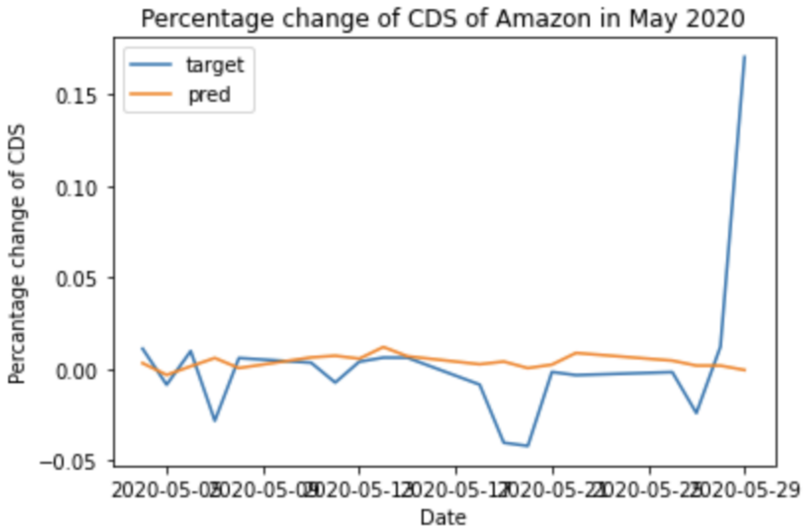


Fig. 2. shows the comparisons of our prediction and the real result on the percentage change of CDS in May using the sentence embeddings approach

As both models are too stable and provide not enough fluctuations, we did not integrate these models into our system at the end.

A further exploration that could be potentially done in the future is to extend the complexity of the model (proposed in 2.1.6.2) trained on the various sentence embeddings. Such as an LSTM model with an awareness of time and can ingest much more data (such as embeddings from the past days, past CDS/stock price and etc.).

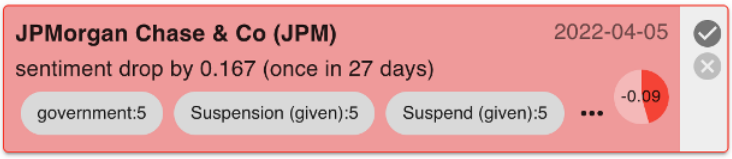
2.4.7 Alerts Model Evaluation

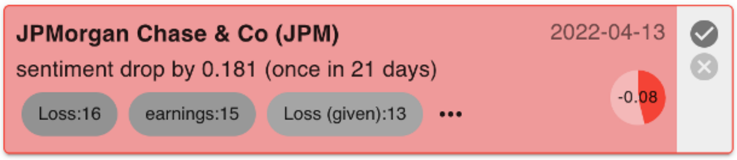
Real-world examples are examined as an evaluation of our alert generation model. In this section, we provide detailed explanations of the examples of JPMorgan, Netflix, and Twitter. More examples are enlisted in Appendix 6.9.

2.7.4.1 J.P. Morgan portfolio bond default and negative earnings surprise (April 2022)

CEO Jamie Dimon of J.P. Morgan warned of the bank's potential losses on $1 billion on Russian bonds on 4th April. [14] Despite stock price not reflecting the news, our system captures a sentiment drop of 0.31 (with quantile score 0.0046), and generates an alert correspondingly.

On 13th April, J.P. Morgan released a sub-expected 2022 Q1 earnings report, with quarterly profits slumped 42%. Stock of JPM is traded 3.2% down in reaction to the earning surprise. Another alert is captured in our system, with key topics identified as 'Loss' and 'Earnings'





2.7.4.2 Netflix's negative outlook (21st January 2022)

Netflix releases its 2021 Q4 earning report on 21st January 2022. Despite beating analysts' expectations marginally, Netflix shows signs of stagnation with a flattening of active users count and a sub-expected outlook. [15] The stock price of Netflix plummets by 21.79%, creating the hugest historical intraday drop as a large-cap company. Our system captures two pre-incident alerts on 3 days (18th January) and 1 day (20th January) before the plunge.

2.7.4.3 Twitter's acquisition by Elon Musk (4th April 2022)

Elon Musk is revealed to have taken Twitter's shares valued at \$3.7 billion on 4th April, becoming the largest stakeholder of the social media company. [16]

Reacting to the news, Twitter is traded up 26%. A reminder is generated by our system, indicating the counterparty is experiencing a hype in sentiment (with quantile score 0.012) with key topics 'Admin Change' and 'Acquisition' captured.

More examples of real-world news events captured by our alert system are provided in Appendix 6.9.

2.4.8 User Acceptance Evaluation

Overall, we got eight students from computer science and business schools to take our survey. We also have software engineers working in financial industries and people working in markets to take our survey. The responses are positive. All people agree that our system is user-friendly and not confusing. 9 out of 10 users gave a rating of 4 out of 5 to our charts. Users strongly agree that our alert feature is clearly shown inside the system. 9 out of 10 users agree that the system can improve the efficiency of the market risk team. There is also some feedback given by the users. We already fixed some of the useful ones like making more instructions and headers for each feature, and adding links to the article in the news section. For more details about the user acceptance testing, please view Appendix 6.10.

3. Discussion

We started Portfolio Early Alert Intelligence from scratch with no knowledge of Finance or Risk Management. Since this was an open-ended project, we had to brainstorm more than we could code in the initial phase of the project to fit the best and most optimal features in our system. In the brainstorming period, we defined three clear objectives as aforementioned and further received approval from the product users i.e. Societe Generale team.

Once we attained the minimum requirement of the project i.e. three objectives, we further explored multiple datasets such as Credit Default Swap, Geographic Stock-Based such as FTSE, topic modeling via LDA, and topic inference or similarity with fine-tuned FinBERT model. Some of these explorations led to excellent results, however, other times, despite our constant efforts, we have not achieved the desired outcome.

Further enhancements that we would have hoped to make had time permitted:

1. Feedback loop to the model on sentiment tagging: With this, we were trying to achieve the evolution of the definition of sentiments. Although scientific measures were taken when labeling the dataset, however, the sentiment tags i.e. “Positive”, “Neutral” and “Negative” are still to an extent subjective. Therefore, in the future, if a user finds discrepancies in our model’s output, they can feed the correct output to the model by labeling those discrepancies themselves and retrain the model.
2. Adding SocGen’s positions per counterparty and giving a custom recommendation based on the severity of the stakes i.e. Long $500,000,000 vs Long $5,000.
3. Based on the weighted sentiment, predicting the best portfolio one can make from the selected counterparties given a fund.

Limitations and challenges that we encountered during the project:

1. Supervisors: During the course of the project, many supervisors from SG had come and left not just our project but the firm for greener pastures. Because of these changes, there was a lack of supervision and it was left upon us students to come up with a workable idea and thereby deliver it at the end. There were motivational issues during the struggling period of the project, especially because everything was online, however, the team overcame those challenges and was stronger than ever.

2. Data: The financial data was not enough for our research and we would have potentially wanted to fetch more data including full articles of news from Bloomberg Terminal when this project began. We tried to gain their API access from HKUST Business School, however, we were no allowed.
3. GPUs: After gaining access to finnhub, we had access to plenty of data even though more is always better. We had prepared multiple models for experimentation, however, training these models from scratch was a tedious job. The LDA model which was to generate the topics for the articles would alone take a lot of time. Therefore, having access to GPU clusters other than our personal computer would have been great. Neither SocGen nor HKUST were able to help us with these problems.

Descriptive summary of the three major objectives achieved:

1. **Find ways to detect a sudden change in risk based on the news:**
   The Market Risk Department at Société Générale relies majorly on news to make important decisions. Since they are majorly concerned about the sentiment defined by the keywords present in the news rather than the content of the news itself, we took it upon ourselves to explore and come up with ways to give a summary of the news and events using natural language processing tools. After investigating different approaches such as LSTM, BERT, FinBERT, et al, we found that since it is optimized on the financial datasets and includes corpora that demonstrate finance lingo, it produces the best results.

   In addition, we investigated different ways backed by data science techniques to detect sudden changes in counterparties' news and report a non-appetible risk in the form of alerts. With the help of the FinBERT model, we created alerts based on quantile scores as opposed to sticking with trivial techniques such as a hard-defined threshold. The alerts were then evaluated and found to have produced more than satisfactory results as discussed in 2.4.7,

2. Investigate relationships between different times-series data on counterparties for potential correlation or forecasting

In terms of the correlation models, the stock price correlates well with the average sentiment score generated by FinBERT and V-EMA method, according to 2.4.5. The model is highly responsive and correlated during days with significant price changes. Such features make it an efficient alert indicator.

In terms of forecasting models, there is still room for improvements. Though we improved our original approach from directly fine-tuning the FinBert model to using word embeddings to complete our tasks. The prediction and the target percentage change in CDS price still has some instability. As daily news may not directly affect the percentage change in CDS price, more inputs like corporate reports and stock price can act as inputs to the model. We did not integrate these two models into the system.

3. Provide a web app that displays information and risks

The final component that brings our work to life clearly has 3 main components:

1. *Dashboard*:
   This component gives an overview of all the user-added counterparties along with the daily alerts if any.
2. *Counterparties:*
   This gives a separate analyzing space to the user to lookup particular counterparties, add/delete counterparties in the portfolio, sort counterparties based on their sentiment, keyword count, etc. Most of all, it serves as an intermediary to the counterparty analysis page that provides a visual summary through graphs, news tags, and past alerts.
3. *Topics:*
   Lastly, the topics component uses Latent-Dirichlet Allocation for topic suggestion and cosine similarity to find the top-most counterparties that include articles that correlate with user-defined topics.

4. Conclusion

To conclude, our system encompasses many different aspects of computer science and risk management such as finance, artificial intelligence, and web programming. After careful scrutiny of our users' demands, we as a team unanimously decided to put only the key features that we saw would be relevant to Société Générale's Risk Management Team. The portfolio in our system is constructed from the counterparties added to the system. These counterparties are thereafter tracked on a daily basis and the users are alerted of any suspicious activities early on so they can perform their due diligence in a timely fashion.

As for the objectives, by setting high expectations and precise objectives during the initial phase of the project, we were able to tick all the objectives successfully and surpass Société Générale's expectations at the end of the project. HKUST student team proudly created a fully functioning counterparty portfolio early alert system by integrating news ingest from various sources, natural language processing state-of-the-art fine-tuned FinBERT model along with cosine-similarity and Latent-Dirchilest Allocation topic modeling algorithms with advanced analytics bundled in the form of a React Web Application.

5. References

[1] S. Hamilton, “Key risk indicator categories for bankers,” Key Risk Indicator Categories for Bankers. https://www.360factors.com/blog/key-risk-indicator-for-bankers/ (accessed: 15-Sept-2021)

[2] CFI, “What is credit rating,” https://corporatefinanceinstitute.com/resources/knowledge/finance/credit-rating/

[3] StockSnips, “StockSnips - Linkedin Profile,” *Linkedin.* https://www.linkedin.com/company/stocksnips/.

[4] StockSnips, “Stocksnips - Apps on Google Play,” *Google*. [Online]. Available: https://play.google.com/store/apps/details?id=com.osage.stock.snips.

[5] Crunchbase, “StockSnips - Funding, financials, valuation & investors,”. https://www.crunchbase.com/organization/stocksnips/company_financials.

[6] StockSnips, “Stocksnips 2017 v 3,” *YouTube*, 11-May-2017. https://www.youtube.com/watch?v=38Ey7yD0jPM

[7] Z. E. Research, “CrowdStrike (CRWD) Q2 Earnings & Revenues Top Estimates, Up Y/Y,” 01-Sep-2021. [Online]. Available: https://www.zacks.com/stock/news/1789718/crowdstrike-crwd-q2-earnings-revenues-top-estimates-up-yy.

[8] Faculty, “Applying NLP to news articles to trace business sentiment,” 25-Jul-2021. https://faculty.ai/ourwork/applying-nlp-to-news-articles-to-trace-business-sentiment/.

[9] S&P, “Creating an early warning system for potential credit risks,” *S&P Global Market Intelligence*, 03-Jun-2019. https://www.spglobal.com/marketintelligence/en/news-insights/blog/creating-an-early-w

arning-system-for-potential-credit-risks.

[10] L. Barbaglia, S. Manzan, and E. Tosetti, "Forecasting loan default in Europe with MACHINE LEARNING*," *OUP Academic*, 03-Jul-2021.

[11] D. Appelbaum, H. K. Duan, T. Sun, and H. Hu, "Business news headlines and the prophetic vision of BANKRUPTCIES: An application of natural language processing," *SSRN Electronic Journal*, 2021.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding.", 11-Oct-2018. https://arxiv.org/abs/1810.04805

[13] Y. Yang , M.C.S. Uy, and A. Huang, "FinBERT: A Pretrained Language Model for Financial Communications", 15-Jun-2020. https://arxiv.org/abs/2006.08097

[14] Reuters. "JPMorgan's Dimon warns of possible $1 billion Russia loss.", April 5, 2022. https://www.reuters.com/business/finance/jpmorgans-dimon-warns-potential-1-bln-loss-russia-exposure-2022-04-04/

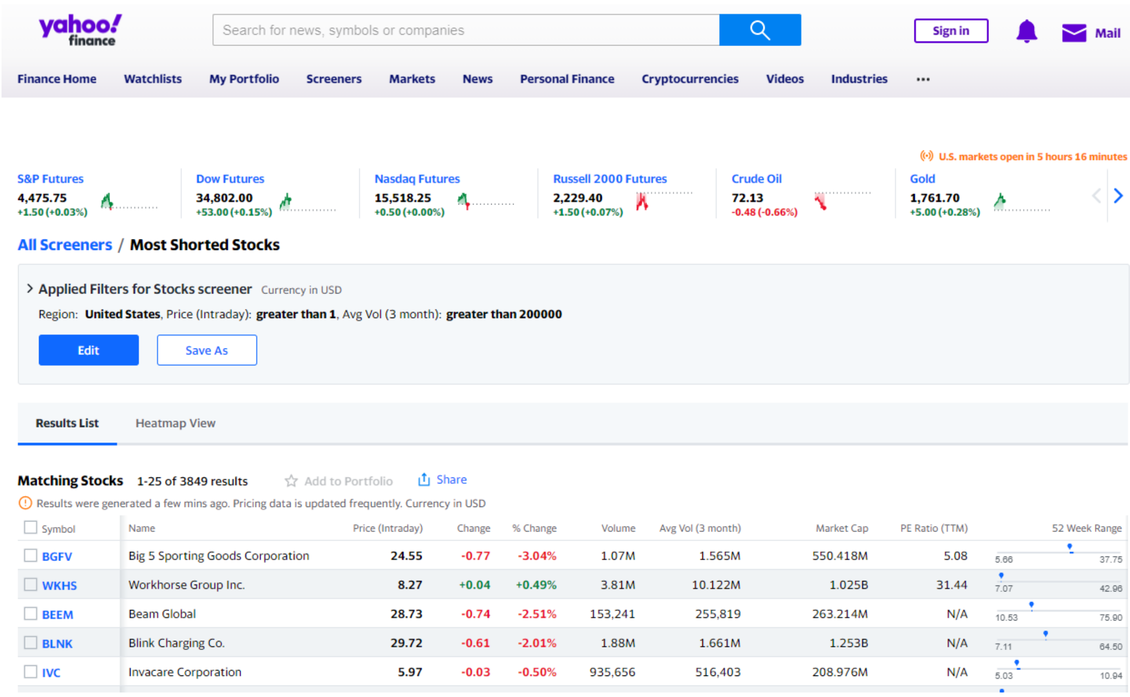
[15] Hayes D. , "Netflix Stock Gets Pummeled, Closing At Lowest Level Since April 2020 After Disappointing Earnings & Wall Street Downgrades – Update." January 21, 2022. https://deadline.com/2022/01/netflix-stock-pummeled-disappointing-earnings-wall-street-streaming-subscribers-1234917209/

[16] Yahoo Finance. "Twitter worth $8.5 billion more after Elon Musk's investment sends stock on record run", April 5, 2022. https://finance.yahoo.com/m/c6f4d8e3-60d2-3de4-96cd-92f5ad9d2071/twitter-stock-rockets-toward.html

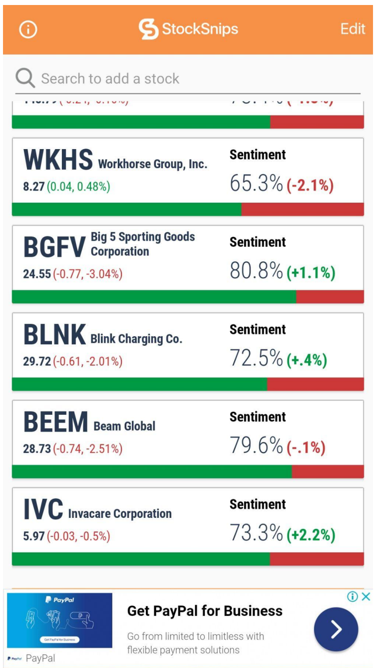
[17] Wikipedia. "Likert scale", April 18, 2022. https://en.wikipedia.org/wiki/Likert_scale

6 Appendix

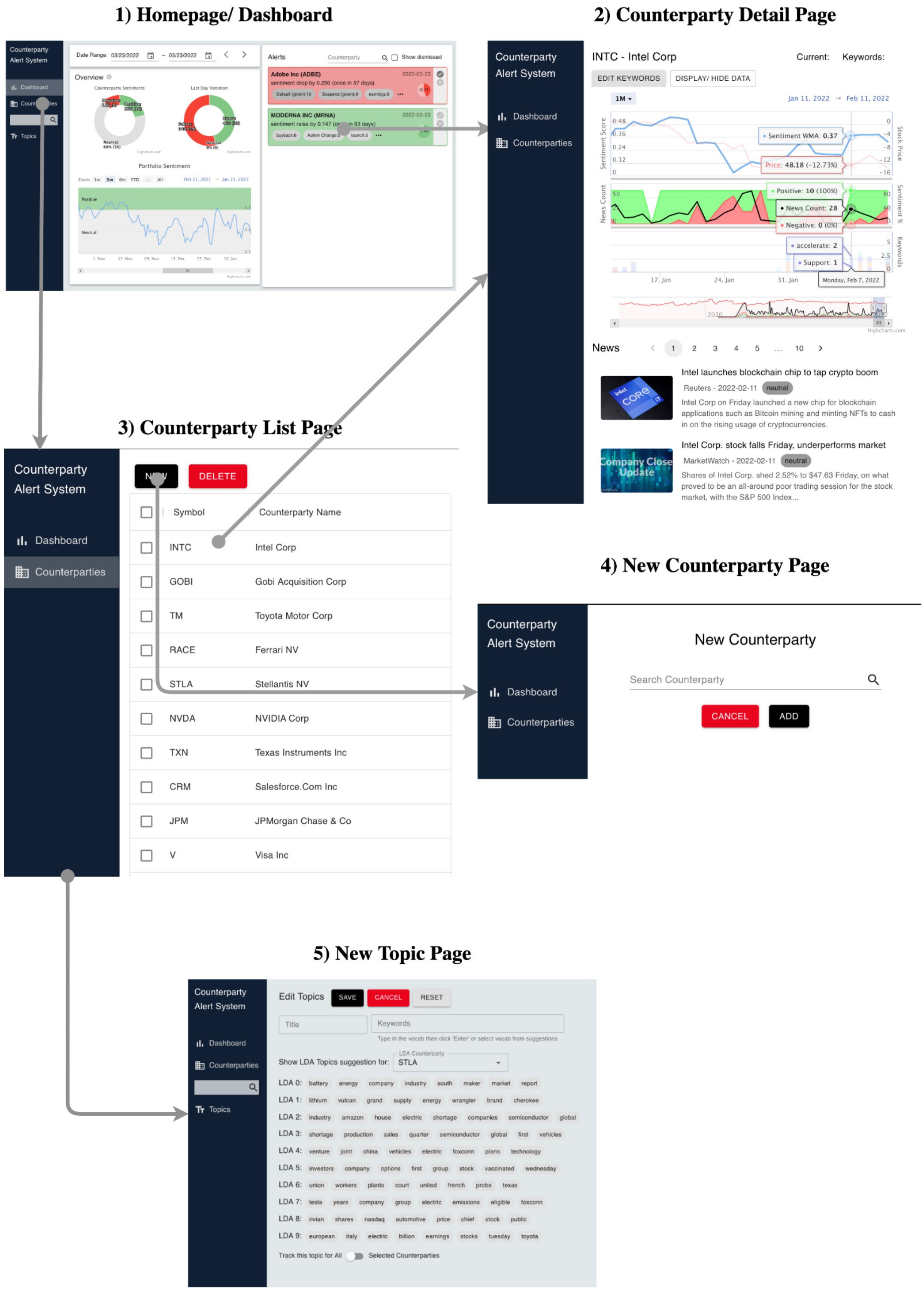
6.1 Most Shorted Stocks vs. StockSnips Sentiment



Screenshot of https://finance.yahoo.com/screener/predefined/most_shorted_stocks/



6.2 UI Flow


1) Homepage/ Dashboard
2) Counterparty Detail Page
3) Counterparty List Page
4) New Counterparty Page
5) New Topic Page
Counterparty Alert System
Dashboard
Counterparties
Topics
Overview
Alerts
Show dismissed
Adobe Inc (ADBE)
MODERNA INC (MRNA)
Portfolio Sentiment
INTC - Intel Corp
Current:
Keywords:
EDIT KEYWORDS
DISPLAY/ HIDE DATA
Jan 11, 2022 → Feb 11, 2022
Sentiment WMA: 0.37
Price: 48.18 (−12.73%)
Positive: 10 (100%)
News Count: 28
Negative: 0 (0%)
accelerate: 2
Support: 1
Monday, Feb 7, 2022
News
Intel launches blockchain chip to tap crypto boom
Reuters - 2022-02-11
neutral
Intel Corp on Friday launched a new chip for blockchain applications such as Bitcoin mining and minting NFTs to cash in on the rising usage of cryptocurrencies.
Intel Corp. stock falls Friday, underperforms market
MarketWatch - 2022-02-11
neutral
Shares of Intel Corp. shed 2.52% to $47.63 Friday, on what proved to be an all-around poor trading session for the stock market, with the S&P 500 Index...
NEW
DELETE
Symbol
Counterparty Name
INTC
Intel Corp
GOBI
Gobi Acquisition Corp
TM
Toyota Motor Corp
RACE
Ferrari NV
STLA
Stellantis NV
NVDA
NVIDIA Corp
TXN
Texas Instruments Inc
CRM
Salesforce.Com Inc
JPM
JPMorgan Chase & Co
V
Visa Inc
New Counterparty
Search Counterparty
CANCEL
ADD
Edit Topics
SAVE
CANCEL
RESET
Title
Keywords
Show LDA Topics suggestion for:
STLA
LDA 0: battery energy company industry south maker market report
LDA 1: lithium vulcan grand supply energy wrangler brand cherokee
LDA 2: industry amazon house electric shortage companies semiconductor global
LDA 3: shortage production sales quarter semiconductor global first vehicles
LDA 4: venture joint china vehicles electric foxconn plans technology
LDA 5: investors company options first group stock vaccinated wednesday
LDA 6: union workers plants court united french probe texas
LDA 7: tesla years company group electric emissions eligible foxconn
LDA 8: rivian shares nasdaq automotive price chief stock public
LDA 9: european italy electric billion earnings stocks tuesday toyota
Track this topic for All
Selected Counterparties

6.3 Cleaned Keywords from Société Générale

```
keywords = ["Ownership Change", "Change of Control", "Acceleration",
"Accelerate", "Default", "Insolvency", "Insolvent", "Delay", "Late",
"Failure", "Fail", "Dispute", "Liquidation", "Liquidator", "Margin call",
"Haircut", "Bank Run", "Termination", "Moratorium", "Suspension", "Suspend",
"Fraud", "Misrepresentation", "Fine", "Sanction", "Breach", "Reschedule",
"Restructuring", "Restructure", "Credit Event", "Losses", "Loss", "Bailout",
"Bailing", "Bankrupt", "Receivership", "Receiver", "Judicial Management",
"Judicial Manager", "Administration", "Administrator", "Sequestrate",
"Sequestration", "Support", "Capital call", "Liquidity Event", "Negative
trends", "Price changes", "Board Infighting", "Corruption", "Inappropriate or
ultra vires dealings", "Negative working capital", "Acquisition", "LBO",
"Qualified audit opinion", "Regulatory Breach", "Non-performing Assets",
"Provisions", "Force majeure", "Distress", "Frozen", "Delisted", "Sued",
"Suit", "Arrested", "Disappeared", "Uncontactable"]
```

6.4 Similarity Equation used in the System

$$\textit{similarity}(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Fig 5. Cosine Similarity Formula Between A and B Vectors
(Relatively common & fast method for similarity, other metrics may be used later based on performance.)

6.5 Latent-Dirchilest Allocation (LDA)

In a corpus with M documents, each document consists of N words, the documents consists of Kunknown topics, V vocabulary in total.

Denote

$W \in R^{M\times N}$, where $W_{i,j} \in [1, V]$ be the vocab of j[th] word of i[th] document

$Z \in R^{M\times N}$, $where\ Z_{i,j} \in [1, K]$ as the topic of j[th] word of i[th] document.

$\theta_i = \left[\theta_{i,1}, \theta_{i,2}, ..., \theta_{i,K}\right]$ as the topic distribution of document i

$\varphi_k = [\varphi_{k,1}, \varphi_{k,2}, ..., \varphi_{k,V}]$ as the vocab distribution of topic k

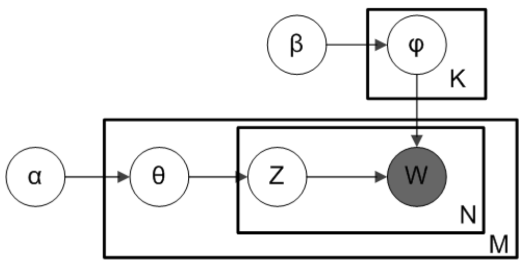
In LDA, θ_i and φ_k are modeled by Dirichlet distribution,

$\theta_i \sim Dir(\alpha)$, where $\alpha = [\alpha_1, ..., \alpha_K]$ be the prior distribution of topic *(num_topics)*

$\varphi_k \sim Dir(\beta)$, where $\beta = [\beta_1, ..., \beta_V]$ be the prior distribution of vocab *(id2word)*

The Bayesian inference process of LDA is then described as

1. Input parameters are α and β
2. θ_i and φ_k is generated by Dirichlet distribution of α and β respectively
3. Each of $Z_{i,j}$ is generated by the multinomial distribution of vector θ_i
4. Each of $W_{i,j}$ is generated by multinomial distribution of $\varphi_{Z_{i,j}}$



The Bayesian probability of the model is given by

$$P(W, Z, \Theta; \alpha, \beta) = \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j)\, P(W_{j,t} | \varphi_{Z_{j,t}})$$

The task is to find the parameter α, β such that the Bayesian probability is maximized. In a computer program, it can be estimated by Gibbs sampling, described as following

1. Randomly select $W_{i,j}$
2. Update $Z_{i,j}$ to $Z_{i,j}'$ by $\varphi_{k, W_{i,j}}\ \theta_{i,k}$
3. Update θ_i' and φ_k' according to new $Z_{i,j}$

4. Fit θ_i' and φ_k' with new distribution $\theta_i' \sim Dir(\alpha')$, $\varphi_k \sim Dir(\beta')$

With new document to be identified, according to the topic distribution of vocab in Θ, we can infer the topic of each word in the new document. This word-by-word inference is a weighted sum by TF-IDF score to get the inference of the document as a whole. TF-IDF is defined as

$$w_{i,j} = tf_{i,j} \times log(\frac{N}{df_i})$$

Where $tf_{i,j}$ is the number of occurrences of the word i in document j, df_i is the number of documents containing word i, N is the total number of documents.

```python
def get_lda(sentences, num_topics=20):
    sentences = sentences.map(lambda x: re.sub(r'[,\.!?]', '', x))
    docs = list(sents_to_words(sentences))

    docs = remove_stopwords(docs)
    id2word = corpora.Dictionary(docs)

    id2word.filter_extremes(no_above=0.25)
    corpus = [id2word.doc2bow(doc) for doc in docs]

    lda_model = LdaModel(corpus=corpus,
                        id2word=id2word,
                        num_topics=num_topics,
                        passes=10)
    return lda_model
```

Fig.1 Normal LDA without Fine-Tuned TDidVectorizer

```python
def tokenize (articles):
    stemmer = WordNetLemmatizer()
    tokens = [word for word in nltk.word_tokenize(articles) if (len(word) > 4) ]
    stems = [stemmer.lemmatize(item) for item in tokens]
    return stems

def advanced_lda_model (sentences, num_topic_words = 10, num_topics = 10):

    vectorizer_tf = TfidfVectorizer(tokenizer = tokenize, stop_words = 'english',\
                                max_features = 1000, use_idf = False, norm = None)
    tf_vectors = vectorizer_tf.fit_transform(sentences)

    lda = decomposition.LatentDirichletAllocation(n_components = num_topics,\
                                        max_iter = 3, learning_method = 'online',\
                                        learning_offset = 50, n_jobs = -1, random_state=4201)

    W1 = lda.fit_transform(tf_vectors)
    H1 = lda.components_
    vocab = np.array(vectorizer_tf.get_feature_names())

    top_words = lambda t: [vocab[i] for i in np.argsort(t)[:-num_topic_words-1:-1]]
    topic_words = ([top_words(t) for t in H1])
    topics = [' '.join(t) for t in topic_words]

    return topics
```

Fig.2 Advanced LDA with Fine-Tuned TDidVectorizer

6.6 Jensen-Shannon Distance

The Jensen-Shannon distance measures the difference between two probability distributions. For example, suppose P = [0.36, 0.48, 0.16] and Q = [0.30, 0.50, 0.20]. The Jenson-Shannon distance between the two probability distributions is 0.0508. If two distributions are the same, the Jensen-Shannon distance between them is 0.

Jensen-Shannon distance is based on the Kullback-Leibler divergence. In other words, to compute Jensen-Shannon between P and Q, one must first compute M as the average of P and Q and then Jensen-Shannon is the square root of the average of KL(P, M) and KL(Q, M). In symbols:

$JS(P, Q) \; = \; \sqrt{\frac{[KL(P,M) + KL(Q,M)]}{2}}$, where $M \; = \; \frac{P+Q}{2}$ and

$P \; \& \; Q \; are \; probability \; distributions.$

$$KL(P, Q) \; = \; \sum_{i=0}^{k} (p[i] \; * \; ln(p[i] \; * \; q[i])),$$

where $p[i] \; and \; q[i] \; are \; ith \; elements \; in \; P \; and \; Q \; respectively.$

```
P = [0.36, 0.48, 0.16]
Q = [0.30, 0.50, 0.20]

M = 1/2 * (P + Q)
  = [0.33, 0.49, 0.18]

KL(P,M) = 0.36 * ln(0.36 / 0.33) +
          0.48 * ln(0.48 / 0.49) +
          0.16 * ln(0.16 / 0.18)
        = 0.002582

KL(Q,M) = 0.30 * ln(0.30 / 0.33) +
          0.50 * ln(0.50 / 0.49) +
          0.20 * ln(0.20 / 0.18)
        = 0.0002580

JS(P,Q) = sqrt[ (KL(P,M) + KL(Q,M)) / 2 ]
        = sqrt[ (0.002582 + 0.0002580) / 2 ]
        = 0.050803
```

Fig 4. Worked Example for computing Jensen Shannon Distance

6.7 Multidimensional Scaling/Principal Component Analysis

Multidimensional scaling (MDS) is a means of visualizing the level of similarity of individual cases of a dataset. MDS is used to translate "information about the pairwise 'distances' among a set of n objects or individuals" into a configuration of n points mapped into an abstract Cartesian space. There are various types of metrics used to calculate the difference such as Euclidean, L1, or Chebyshev's, however, in this case, we use an iterative method.

The iterative method is more general and can be applied when distances (dissimilarities) are not Euclidean. The iterative method consists of minimizing a cost function, which is defined as followed:

$$Stress_D(x_1,\ x_2...,\ x_N = sqrt(\sum_{i\neq j=1}^{N}(d_{ij} - ||x_i - x_j||^2)$$

where $d_{ij} = distance\ between\ ith\ and\ jth\ object\ \approx ||x_i - x_j||$

6.8 Quantile Score Metrics used in Alert Generation

Denote a set of data in ascending order as $\{X_{(1)},\ X_{(2)}, \dots, X_{(n)}\}$, then by order statistics, the empirical distribution of X is

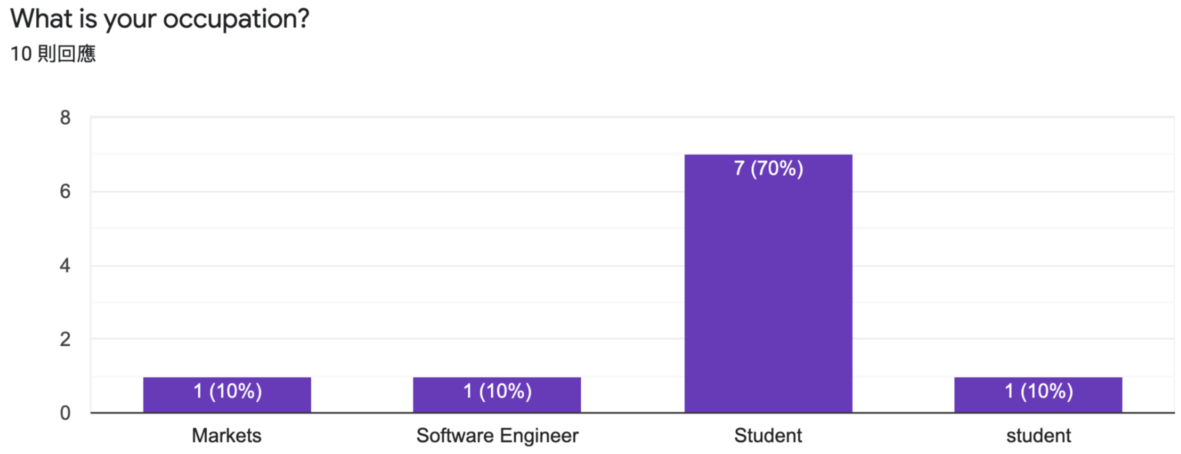
$$F_X(x) = \begin{cases} 0, & if\ x < X_{(1)} \\ \frac{k}{n}, & if\ X_{(k)} \leq x < X_{(k+1)} \\ 1, & if\ x \geq X_{(n)} \end{cases}$$

6.9 Example Events Captured in Alert Generation System

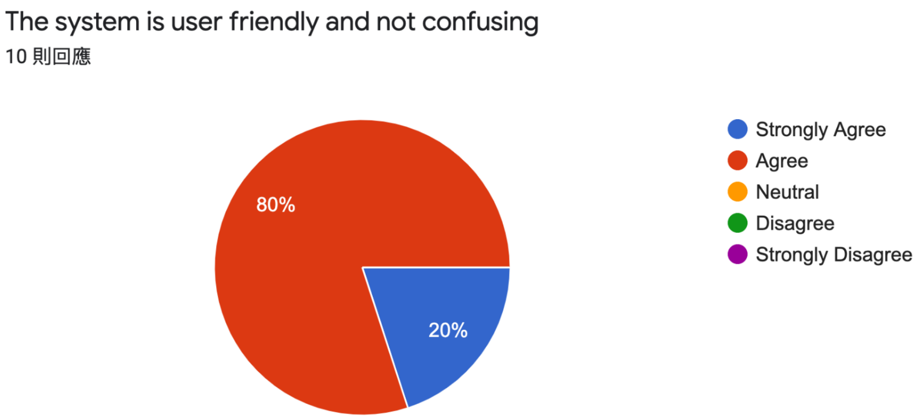
Company	Date	events	alert/ reminder captured
Charles Schwab Corp (SCHW)	2022-04-18	Negative Earnings Surprises Stock price drop of 9.44%	Alert (2022-04-18) Topics Identified: Loss, Up& Downs, Price
Alibaba Group Holding Ltd. (BABA)	2022-04-07	Chinese companies threatened to suspend listing in US by regulations	Alert (2022-04-07) Key topics: Lawsuit, Suspension
Salesforce.Com Inc (CRM)	2022-04-05	Earning surprises with positive outlook; Increasing demand of cybersecurity due to the outbreak of Russian-Ukraine war	Reminder (2022-04-05) Topics Identified: Accelerate, Ups, Support
Adobe Inc. (ADBE)	2022-03-23	Negative outlook surprises (-9.26%)	Alert (2022-03-23) Alert (2022-03-24)
Honda Motor Co. Ltd. (HMC)	2022-03–02	Reports of sales declines; Suspension of automobile exports to Russia	Alert (2022-03-02) Topics Identified: Suspend, government
Meta Platforms Inc. (FB)	2022-02-03	Negative earnings surprises with historical intraday drop (-26.39%)	Reminder (2022-02-01) Alert (2022-02-03) Topics Identified: Loss, Negative, Earnings
HSBC Hodlings (HSBC)	2021-12-17	Fined due to failing UK's anti-money laundering controls	Alert (2021-12-17) Topics Identified: Suspension, Fined
JD.com Inc. (JD)	2021-07	Delisting threats of Chinese stock from Nasdaq exchange	Alerts: (2021-07-06~08) Topics Identified: Fine, Breach

6.10 User Acceptance Testing Results

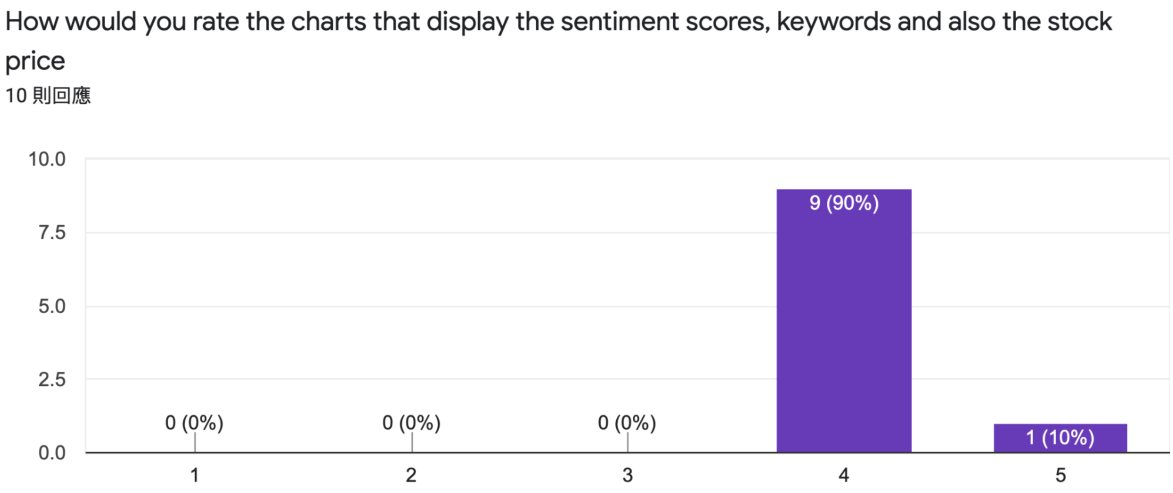
1. What's your occupation?



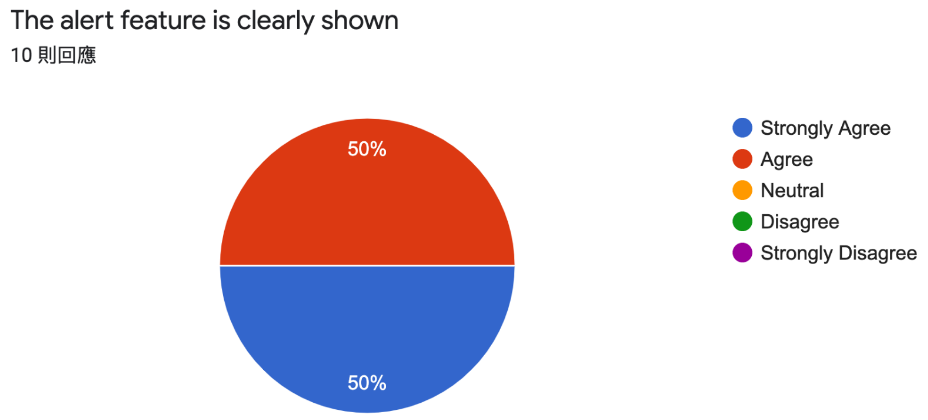
2. Is the system user-friendly and not confusing?



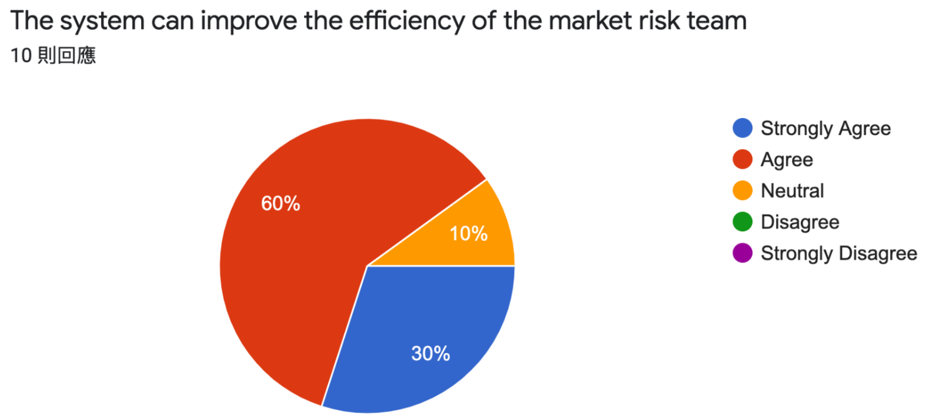
3. How would you rate the charts that display the sentiment scores, keywords and also stock price?



4. Is the alert feature clearly shown?



5. Can the system improve the efficiency of the market risk team?



6 users strongly agree that the system can enhance the efficiency of the market risk team. 3 users agree that the system can enhance the efficiency of the market risk team, while one remains neutral.

6. What would you like to see in the system that is not there?

what would you like to see in the system that is not there?

10 則回應

N/A

I'm satisfied

Addition of Position with a counterparty in the Portfolio

instructions

Maybe a link to the article in the news section

Including stock price, volume data in the visualization port; Can add correlation about news count, sentiment score, and volume

stock recommendations

No

The system has been improved after receiving the aforestated feedback from the test users. Following are our responses to the feedback:

a. **Addition of Position with a counterparty in the Portfolio**

   This feature has already been implemented.

b. **Instructions**

   More headers to each section have been added to make things clearer. Further instructions may be added to a guide page if time allows.

c. **Maybe a link to the article in the news section**

   Each post is linked to the origin of the article after the feedback.

d. **Including stock price, and volume data in the visualization port; Can add correlation about news count, sentiment score, and volume**

We have added the stock price inside the visualization part. Along with that, there is a relationship that can be seen with the stock pricke from the sentiment trends.

e. Stock recommendations

Stock recommendations would be interesting, however, that is outside the scope of the project. This has been taken into account in the discussion section and, perhaps the next iteration of the system would include that. For now, there is an overview included in the dashboard that indicates the “Failures” and “Risers” percentages for visualization.

7. Any improvements to the system that you would like to see?

Any improvements on the system?

10 則回應

N/A

NO

NA

instructions for functions

Looks ok to a layman

Including stock price, volume data in the visualization port; Can add correlation about news count, sentiment score, and volume

looks good. perhaps a page which provides an overview of the features and how to use it (i.e. a video) would be useful.

No

We improved our system after we received the survey. Here are our responses to the feedback:

a. **Instructions for functions**

   More headers to each section have been added to make things clearer. Further instructions may be added to a guide page if time allows.

b. **Looks good. Perhaps a page that provides an overview of the features and how to use it (i.e. a video) would be useful.**

   There will be a video trailer including the instructions on using the system and it’s features.

6.11 Samples used in Testing of News Counterparty Categorization


```
[{
  "_id": {
    "$oid": "61d7d7fd09341707b4cc9439"
  },
  "headline": " Britain approves Pfizer COVID-19 vaccine for teenagers",
  "summary": "Britain's medicines agency has approved an extension of Pfizer
and partner BioNTech's COVID-19 vaccine for use in children as young as 12, it
said on Friday, a week after a similar clearance was given by European
authorities. (Reporting by Pushkala Aripaka in Bengaluru;...",
  "url":
"https://finnhub.io/api/news?id=3d6fe1394d77f2696c739cb2c81b30042e7add24a1918d
649c0be0c900a152c6",
  "counterparty": "PFE",
  "date": "2021-06-04"
},{
  "_id": {
    "$oid": "61bad5e1350f0214a8ef5ffc"
  },
  "headline": " FT names Elon Musk as 'Person of the Year' ",
  "summary": "Tesla Inc Chief Executive Officer Elon Musk has been named the
Financial Times newspaper's \"Person of the Year\", two days after getting a
similar recognition from the Time magazine. The FT said it picked Musk for
triggering a historic shift in the world's auto industry towards...",
  "url":
"https://finnhub.io/api/news?id=f5a82d2582b60b330822c42f531ce4d9d5818860f89173
ef7268b89b11eb2aec",
  "counterparty": "TSLA",
  "date": "2021-12-15"
},{
  "_id": {
    "$oid": "6233a1ddc475d78a84b74c38"
  },
  "counterparty": "NFLX",
  "headline": " Netflix tests sharing accounts outside household",
  "date": "2022-03-16",
```

```
  "summary": "Netflix Inc is testing features including one that will allow
accounts to be shared outside members' household at an extra cost, the
streaming pioneer said on Wednesday.",
  "url":
"https://finnhub.io/api/news?id=c3ba6009f75bef8922f91a83193779fb1746ee68fcbdae
775c69888ddb108a54"
},{
  "_id": {
    "$oid": "61d6859a09341707b43b1430"
  },
  "headline": " Tesla rival Lucid plans to launch in Europe this year",
  "summary": "Lucid has announced that it plans to start expanding into
European markets this year as demand for electric vehicles soars. ",
  "url":
"https://finnhub.io/api/news?id=b1088ea25af89d3c3e30ce9c7fbe04b5a6858cf64fee23
cc3dc87869ad49d67d",
  "counterparty": "F",
  "date": "2022-01-05"
},{
  "_id": {
    "$oid": "624d03f7c475d78a84ce9940"
  },
  "counterparty": "TWTR",
  "headline": " Tesla rival Lucid plans to launch in Europe this year",
  "date": "2022-01-05",
  "summary": "Lucid has announced that it plans to start expanding into
European markets this year as demand for electric vehicles soars. ",
  "url":
"https://finnhub.io/api/news?id=b1088ea25af89d3c3e30ce9c7fbe04b5a6858cf64fee23
cc3dc87869ad49d67d"
},{
  "_id": {
    "$oid": "616a9215d6b0969db79ff673"
  },
  "headline": " Tesla's Brandenburg factory becomes festival site for
'Giga-Fest'",
  "summary": "From flashing lights and booming speakers to sprawling stages
and a Ferris wheel, Tesla's factory near Berlin has been transformed into a
festival site for a one-day county fair on Saturday, hosted by CEO Elon
Musk.",
```

```
  "url":
"https://finnhub.io/api/news?id=f1c5d986d87aeb1504ed2154581e6bb6cf5ddb163688f3
d9e43db8c808baa77f",
  "counterparty": "TSLA",
  "date": "2021-10-09"
},{
  "_id": {
    "$oid": "61cd4b2e350f0214a8b9064f"
  },
  "headline": " Toyota Q4 operating profit nearly doubles from year-ago ",
  "summary": "Toyota Motor Corp reported on Wednesday its fourth-quarter
operating profit jumped 92%, nearly doubling since last year as it shrugged
off chip-supply issues and as its sales recovered from the impact of the
coronavirus pandemic.",
  "url":
"https://finnhub.io/api/news?id=028e5b6a6c578bd0835dfa86fd69711ed1ab029f7e2013
d55c96bfebcbb1d7d0",
  "counterparty": "TM",
  "date": "2021-05-12"
},{
  "_id": {
    "$oid": "61cd4b75350f0214a8b949aa"
  },
  "headline": "\"Acredito que ainda vale a pena investir no Brasil\", diz
presidente da General Motors na América do Sul",
  "summary": "A GM retomou, no início do mês, o plano de aplicar R$ 10 bilhões
ao longo de cinco anos anunciado em 2019",
  "url":
"https://finnhub.io/api/news?id=72d7f781096bbcdfca4028bb94d03cdb4b711fbb4ce25e
8133cee84fbaf37bcb",
  "counterparty": "GM",
  "date": "2021-01-15"
},{
  "_id": {
    "$oid": "61cd501c350f0214a8bd518b"
  },
  "headline": "\"Alexa, what can I ask Bianca Andreescu?\"",
  "summary": "NASDAQ (AMZN) - Today, Amazon Canada announced a collaboration
with Canadian tennis champion Bianca Andreescu to celebrate community this
summer. No matter your level of sports knowledge, Alexa has you covered. Fans
are invited behind the scenes to adopt the same Match Day routine as Bianca.
```

```
Alexa will play crowds cheering you on, give you a compliment, and start a
meditation skill when you say \"Alexa, start Bianca's routine.\" Click on this
Routine, follow the steps, save and the routine will be",
  "url":
"https://finnhub.io/api/news?id=3c48a295e685516f8cbbc70f5ef7294a10c039ccd55e78
9c066f9c6ece9c3120",
  "counterparty": "AMZN",
  "date": "2021-07-27"
},{
  "_id": {
    "$oid": "616a92cad6b0969db7a1e2ce"
  },
  "headline": "\"Apple must loosen app payment system\" - judge in Epic Games
case",
  "summary": "Looking for stock market analysis and research with proves
results? Zacks.com offers in-depth financial research with over 30years of
proven results.",
  "url":
"https://finnhub.io/api/news?id=df24b64acb03e55bd4d2c624c4195117b2680377e69c96
e3f4e479575d895965",
  "counterparty": "AAPL",
  "date": "2021-09-10"
},{
  "_id": {
    "$oid": "61cd501f350f0214a8bd6b40"
  },
  "headline": "\"Ayudadme a buscar vuelos a otro universo\", el drama de un
joven cuyas fotos desnudo se cuelan en la tele de sus padres",
  "summary": "¿Dónde guardas tus fotos? Seguro que, además de guardar tus
archivos en la memoria de tu smartphone, tienes copias almacenadas en lo que
se conoce como ‘la nube’,",
  "url":
"https://finnhub.io/api/news?id=b496e5c35008e2fa54349246a9b20cbce8a4a48b66be03
b92efd2899c2461d66",
  "counterparty": "AMZN",
  "date": "2021-03-23"
},{
  "_id": {
    "$oid": "62227e370f24882f462c8f3f"
  },
  "counterparty": "NFLX",
```

```
  "headline": "\"Bridgerton\" producer Rhimes, make-up artist McGrath heralded
with Barbie dolls",
  "date": "2022-03-04",
  "summary": "U.S. TV producer and writer Shonda Rhimes, British make-up
artist Pat McGrath and French blogger Lena Mahfouf are being honoured with
Barbie dolls in their likeness as part of a celebration of female business
founders.",
  "url":
"https://finnhub.io/api/news?id=335609d344401a074e938c4a5972fb548342ec88af1411
7ab8f7cf46ee4244aa"
},{
  "_id": {
    "$oid": "623e2d56c475d78a84022e57"
  },
  "counterparty": "AAPL",
  "headline": "\"Chip Rally,\" NATO Results, Drive Market Indexes",
  "date": "2022-03-24",
  "summary": "Today's chip rally also owes a debt of gratitude to the NATO
meetings in Brussels just wrapping up today.",
  "url":
"https://finnhub.io/api/news?id=90f8c250fd57854df9cf40c5a39a24021d3bd637647833
b36401c2847acec64b"
},{
  "_id": {
    "$oid": "61f84dd38de3c145e9b7c803"
  },
  "counterparty": "GOBI",
  "date": "2022-01-31",
  "headline": "\"Crash-Proof\" Your Portfolio With These Tax-Free 5%
Dividends",
  "summary": "There's a group of 7%+ dividends out there that are perfect for
today's market. They're far less volatile than \"regular\" stocks, their
payouts are tax-free, and (for now) you can get them for a steal--as cheap as
93 cents on the dollar!",
  "url":
"https://finnhub.io/api/news?id=d1996a88bca853ad4570e7f2d82b0e5ec9fa67f473a186
e7b61f7337f4dfe40f"
},{
  "_id": {
    "$oid": "61cd4e50350f0214a8bbd33b"
  },
```

```
  "headline": "\"Destabilisiert das Land\": Burmas Militär lässt Facebook...",
  "summary": "Seit dem Putsch haben sich Aufrufe zum zivilen Ungehorsam und
Videos von Protestaktionen über Facebook verbreitet. Die Plattform trage zur
Destabilisierung des Landes bei, hieß es nun in der...",
  "url":
"https://finnhub.io/api/news?id=c35c2f4cbda2b48e1c0aa934200658040726beaf019d94
18c74b704da9a25e72",
  "counterparty": "FB",
  "date": "2021-02-04"
},{
  "_id": {
    "$oid": "616a9438d6b0969db7a3edb8"
  },
  "headline": "\"El fin de la propiedad\": en 10 años nadie comprará coches o
iPhones, según el CEO de una empresa valorada en más de 1.000 millones",
  "summary": "Marcas como Volkswagen y Apple apuestan por estrategias de
suscripción al ver como la era del fin de la propiedad golpea la industria
automotriz y de los móviles.",
  "url":
"https://finnhub.io/api/news?id=8fd50ac37c69f2bb905ae5b6354aaa2fdbf65769b5d9a9
fceb3817e4867065fe",
  "counterparty": "AAPL",
  "date": "2020-12-01"
},{
  "_id": {
    "$oid": "61cd4eb2350f0214a8bc1923"
  },
  "headline": "\"Elf on the Shelf\" founder sees excited but 'careful' crowd
turnout for live events",
  "summary": "Christa Pitts, founder and CEO of The Lumistella Company, which
owns the \"Elf on a Shelf\" line of books and properties, shares her views on
expanding during a pandemic, the return to in-person events, and a new deal
for shows on Netflix.",
  "url":
"https://finnhub.io/api/news?id=6b6c3105bc2b133a0c20ba6cb5c96cfa02464d50f2c7ba
639bb9c0dd0bf724a5",
  "counterparty": "NFLX",
  "date": "2021-12-10"
},{
  "_id": {
    "$oid": "61f3087c8de3c145e907325c"
```

```
  },
  "counterparty": "GOOG",
  "date": "2022-01-26",
  "headline": "\"Grand Theft Auto\" Comes to \"FarmVille\" as Take-Two Buys
Zynga",
  "summary": "Shares of Lululemon Athletica (NASDAQ: LULU) fall as the
retailer lowers expectations for its next quarterly report. Asit Sharma
analyzes those stories, the latest innovation from Deere & Co.(NYSE: DE), and
why he's focusing on both capital-light and capital-heavy businesses",
  "url":
"https://finnhub.io/api/news?id=4952cc07b01b07e7426d859d0438e7c2b626ca6c0b083f
054d86a75046652f2c"
},{
  "_id": {
    "$oid": "61f308868de3c145e907365e"
  },
  "counterparty": "HD",
  "date": "2022-01-26",
  "headline": "\"Grand Theft Auto\" Comes to \"FarmVille\" as Take-Two Buys
Zynga",
  "summary": "Shares of Lululemon Athletica (NASDAQ: LULU) fall as the
retailer lowers expectations for its next quarterly report. Asit Sharma
analyzes those stories, the latest innovation from Deere & Co.(NYSE: DE), and
why he's focusing on both capital-light and capital-heavy businesses",
  "url":
"https://finnhub.io/api/news?id=4952cc07b01b07e7426d859d0438e7c2b626ca6c0b083f
054d86a75046652f2c"
},{
  "_id": {
    "$oid": "61f308ad8de3c145e907467e"
  },
  "counterparty": "MS",
  "date": "2022-01-26",
  "headline": "\"Grand Theft Auto\" Comes to \"FarmVille\" as Take-Two Buys
Zynga",
  "summary": "Shares of Lululemon Athletica (NASDAQ: LULU) fall as the
retailer lowers expectations for its next quarterly report. Asit Sharma
analyzes those stories, the latest innovation from Deere & Co.(NYSE: DE), and
why he's focusing on both capital-light and capital-heavy businesses",
```

```
  "url":
"https://finnhub.io/api/news?id=4952cc07b01b07e7426d859d0438e7c2b626ca6c0b083f054d86a75046652f2c"
}]
```


Highly relevant: 13

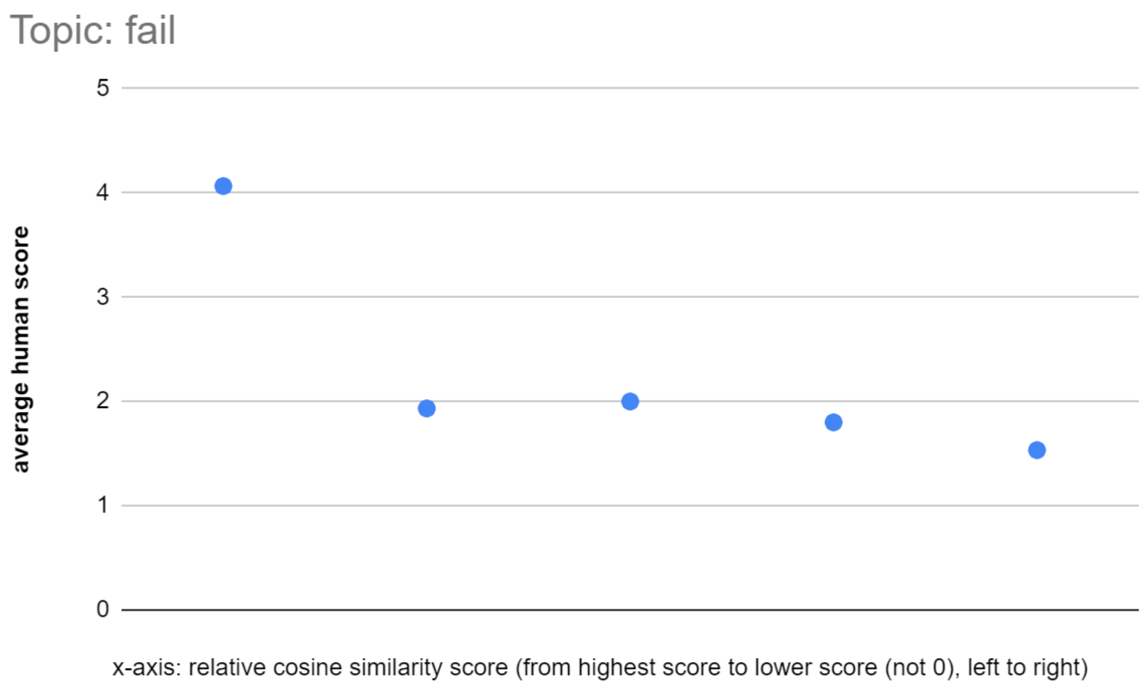
Mentioned: 4

Not relevant: 3

6.12 Survey Results for FinBert-based Topic Categorization

Top x% Percentile of News Sorted by Sim Score vs. "Human-perceived" Similarity

Topic	Top 20% Percentile	Top 40% Percentile	Top 60% Percentile	Top 80% Percentile	Top 100% Percentile
administration	2.933333333	2.266666667	2.666666667	2.333333333	2.266666667
sued	4.533333333	1.4	1.333333333	1.8	1.4
Ibo	2.2	2	2.266666667	1.866666667	1.466666667
provision	2.333333333	2	1.8	1.733333333	1.8
fail	4.066666667	1.933333333	2	1.8	1.533333333



Full data can be found here:
https://github.com/SGHKUST-FYP-Portfolio-Alert/portfolio-early-alert-intelligent-backend/tree/main/survey

6.13 GitHub Repository / Notion

GitHub Org:

https://github.com/SGHKUST-FYP-Portfolio-Alert

Frontend GitHub Repo:

https://github.com/SGHKUST-FYP-Portfolio-Alert/portfolio-early-alert-intelligent-frontend

Backend GitHub Repo (Public):

https://github.com/SGHKUST-FYP-Portfolio-Alert/portfolio-early-alert-intelligent-backend

Notion:

https://www.notion.so/cbcc2d5a8e8d4833b92783ad71507951?v=d1c8f713543b4008be686cd d7128ab97

To view the repository or the notion, send an email to (shipaa@connect.ust.hk or walkerchow@pm.me) such that we can add you as collaborators.

7. Meeting Minutes

7.1 Minutes of the 1st Project Meeting

Date: July 13, 2021
Time: 4:00 pm
Place: Zoom
Present: Walker, Abhishek, Ka Fong, Si Hou, Mr. Ray Wong, Mr. Bao Yang and Dr. Li
Absent: None
Recorder: Ka Fong

1. Approval of minutes
   This was the first formal group meeting, so there were no minutes to approve.

2. Report on progress
   2.1 All team members have been notified that Soc Gen is offering another project (Portfolio Early Alert Intelligent) beside the original one (NLP Market News Analysis) so that two project groups will be working on different topics.
   2.2 Walker has shared with us some information he obtained verbally from Ray, that the new project is a product that helps the Soc Gen credit team to monitor the status of its debtor company.
   2.3 All team members have come to a brief agreement on prioritizing the new project (Portfolio Early Alert Intelligent), since the new project offers higher flexibility and customizability.

3. Discussion items
   3.1 Supervisor Ray Wong introduced the two types of risk SocGen is facing - Market Risk and Counterparty Risk. Market Risk refers to the risk of the market as a whole, like interest risk, currency risk, etc. , while counterparty risk refers to the risk of their clients, whom they lend money to.
   3.2 Ray re-introduced the details of Project1 - NLP Market News Analysis. As written in the project description, this project mainly involves extracting global market news, analysis of sentiments and classification of news events to generate alerts.

   3.3 Ray introduced to us the details of the new Project2 - Portfolio Early Alert Intelligent. This project focuses on Counterparty risk - risk of individual clients. Besides using the NLP news approach, this project encourages the use of more types of features and this project is more data science orientated.

   3.4 Dr. Li suggested Ray re-providing the project description of two projects
   3.5 Dr. Li suggested that two teams should indicate their preference and provide a brief proposal to Ray so that a decision can be made.

3.6 Dr.. Li reminds us of the important date of the FYP project, as listed in the CS FYP 2021-2022 page.

4. Goals for the coming week
   4.1 Ray will provide a new project description.
   4.2 All members will study the new project description, discuss and re-confirm their preference.
   4.3 All members will prepare a brief planning and provide the preference to Ray by Sunday.
   4.4 All members will research different text sentiment models.

5. Meeting adjournment and next meeting
   The meeting was adjourned at 5:00 pm.
   The next meeting will be at 11:00 am on August 6th on Zoom.

7.2 Minutes of the 2nd Project Meeting

Date: August 6th 2021
Time: 2:00pm
Place: Zoom
Present: Walker, Abhishek, Ka Fong, Si Hou, Mr. Ray Wong, Mr. Bao Yang
Absent: Dr. Li
Recorder: Ka Fong

1. Approval of minutes
   The minutes of the last meeting were approved without amendment.

2. Report on progress
   2.1 It is confirmed that our team will be working on a new project - Portfolio Early Alert Intelligent.
   2.2 All members have researched some sentiment models.

3. Discussion items
   3.1 Ray and Bao talk about the data source, to evaluate risk of counterparty, three types of inputs are considered
      a) News from Google, Yahoo, Bloomberg, Reuters
      b) external credit rating - S&P, Moody and Fitch.
      c) internal transaction data
   3.2 Based on three types of inputs, they suggested some usage:
      a) external credit rating change alerts
      b) new sentiment alerts
      c) cash flow/ liquidity alerts
      d) business relationship alerts
      e) key person indicator alerts (eg, CEO resigns)
      f) specific industry indicators

   3.3 Walker and Abhi ask about the list of counterparties, Ray said they will provide one.
   3.4 Walker suggests that we shall be using FinBERT as our text sentiment model. It is a model trained by professors from HKUST Business School, that have fine-tuned Google BERT with 10000 financial sentences.

4. Goals for the coming week
   4.1 All members will research to gain an intuition of the methodologies of credit rating. Abhi will specifically research on S&P methods; Walker will do Moody, Ka Fong will do Fitch and Si Hou will research some other methodologies.
   4.2 All members will come up with some specific indicators.
   4.3 Ray will provide a list of counterparties.
   4.4 Walker will seek chances to meet the credit team of SocGen, as this project will mainly be used by the credit team.

5. Meeting adjournment and next meeting
   The meeting was adjourned at 3:00 pm.
   The next meeting will be at 2:00 pm on August 13th via Zoom.

7.3 Minutes of the 3rd Project Meeting

Date: August 13th, 2021
Time: 2:00 pm
Place: Zoom
Present: Walker, Abhishek, Ka Fong, Si Hou, Mr. Bao Yang
Absent: Mr. Ray Wong, Dr. Li
Recorder: Ka Fong

1. Approval of minutes
   The minutes of the last meeting were approved without amendment.

2. Report on progress
   2.3. All members have researched the credit rating methodologies of their assigned institutions from the last meeting.
   2.4. Ray provided a counterparty list consisting of two companies.

3. Discussion items
   3.3. Si Hou questions that a counterparty list consisting of just 2 companies is insufficient. Bao replied that due to internal policy, it is difficult to disclose more. Yet, using a made up list is possible, like we can target the constituent companies of the Hang Seng Index.
   3.4. Bao reported that this project is not an "automated credit rating system", so predicting the credit rating is not the utmost objective of this project. Yet, it can be adopted as one of the features.
   3.5. Bao stated that as Soc Gen have already built a Reuters headlines scrapper, we should not be spending too much time on data extraction
   3.6. Bao suggested a most basic project model pipeline:
   Step 1: scrape news about the counterparty list
   Step 2: perform NLP to the news
      - use sentiment model to give a sentiment score
      - use NER to retrieve key words
   Step 3: alert user in case significant sentiment score change occurs, with key words displayed
   3.7. Bao suggested two useful sources: Google Alert API and risk.net
   3.8. Walker suggests that we should start typing the proposals, especially the objective part. He states that it is hard to get anything confirmed without a clear paper work.
   3.9. Ka Fong disagrees on the necessity of a written objective. He thinks that deliverables in chart forms, like UI flow chart, backend schemas and model flowchart, are better than deliverables in text forms.

4. Goals for the coming week

4.3. All members will start typing up ideas on the objective sessions of the proposal report. Walker will be responsible for guiding it.
4.4. Ka Fong will start building a prototype UI to demonstrate the deliverables

5. Meeting adjournment and next meeting
The meeting was adjourned at 3:05 pm.
The next meeting will be at 03:00 on August 20^{th} via Zoom.

7.4 Minutes of the 4th Project Meeting

Date: August 20th, 2021
Time: 2:00 pm
Place: Zoom
Present: Walker, Abhishek, Ka Fong, Si Hou, Mr. Ray Wong, Mr. Bao
Absent: Dr. Li
Recorder: Ka Fong

1. Approval of minutes

   The minutes of the last meeting were approved without amendment.

2. Report on progress
   - 2.3. Ka Fong constructed a prototype UI as a demonstration of the primitive objectives and what the team obtained from the last meeting.
   - 2.4. All team members have discussed and filled the objective part of the proposal report with some point form ideas.

3. Discussion items
   - 3.3. Bao reviews the prototype UI and agrees that the UI demonstrates what they want to achieve in the large. Yet, Bao argues that more complicated and sophisticated warnings should be included. An example included would be like "GDP is expected to drop, yield rate is expected to rise, currency exchange rate is expected to drop, so alert risk of default since bond price drops".
   - 3.4. Si Hou argues the feasibility of having such complex warnings. It is already difficult to justify the chain of relationships financially, so it lacks a solid group to automate these types of warnings.
   - 3.5. Ray suggests adopting a model to identify the risk of default or bankruptcy. To achieve this, we should manually label some news as 'bankruptcy related' or 'not bankruptcy related'.
   - 3.6. Abhi disagrees with the approach to manually labelling data. It would take too much of the project time to achieve this. Semi-supervised approach might help, but still, verifying the label one by one is still exhausting.
   - 3.7. Ray suggests that except for an alert generating system, a news ranking system should be included. Ranking the news according to its relationship with bankruptcy would be a choice.
   - 3.8. Team members agree that the bankruptcy labelling model needs further discussion.

4. Goals for the coming week
   - 4.3. All members should finish the introduction part of the proposal report by Sunday.
   - 4.4. All members should finish the objectives of the proposal by Sunday.
   - 4.5. All members should start writing reviews on the products or papers they have reviewed.

5. Meeting adjournment and next meeting
   The meeting was adjourned at 3:10 pm.
   The next meeting will be at 2:00pm on August 27th via Zoom.

7.5 Minutes of the 5th Project Meeting

Date: August 27th, 2021
Time: 2:00 pm
Place: Zoom
Present: Walker, Abhishek, Ka Fong, Si Hou, Mr. Ray Wong, Mr. Bao
Absent: Dr. Li
Recorder: Ka Fong

1. Approval of minutes
   The minutes of the last meeting were approved without amendment.

2. Report on progress
   2.3. Introductions and objectives are drafted.
   2.4. Literature reviews are in progress.
   2.5. Members have discussed that the "bankruptcy news model" is not very feasible currently.

3. Discussion items
   3.3. As labelling "bankruptcy news" is too exhausting, Walker suggests returning to a primitive plan, that is to use "stock price change" as the label first, since a significant drop in stock price could be an alert for debt default or bankruptcy.
   3.4. Ray and Bao agree it is reasonable to use price change as a target at this stage.
   3.5. Ray emphasizes on our second deliverables - "to investigate the relationship between different indicators and risk of bankruptcy and default", thinking this is important and we should spend more effort on designing it.
   3.6. We discuss the choice of implementation tools. Ka Fong suggests using ReactJS as he is familiar with it. Walker suggests using FastAPI as the backend framework as it is light-weight, easy to implement with Python and it aligns with SocGen practice. Si Hou proposes to use SQLite initially due to its light-weight but Abhi suggests using a noSQL database like MongoDB is more suitable to get rid of rigorous data type and schema requirements.
   3.7. Ray and Bao agree on our choice of implementation tools.
   3.8. Si Hou suggests using Finhub API as backup in case Reuters and Bloomberg news datasets are not feasible. It allows backward retrieval of historial news, which cannot be achieved by a lot of other API like Yahoo and Google API.

4. Goals for the coming week
   4.3. All members shall finish the literview reviews by 4th September.

5. Meeting adjournment and next meeting
   The meeting was adjourned at 2:50 pm.

The next meeting will be at 2:00pm on September 3rd via Zoom.

7.6 Minutes of the 6th Project Meeting

Date: September 3rd, 2021
Time: 2:00 pm
Place: Zoom
Present: Walker, Abhishek, Ka Fong, Si Hou, Mr. Ray Wong, Mr. Bao
Absent: Dr. Li
Recorder: Ka Fong

1. Approval of minutes
   The minutes of the last meeting were approved without amendment.

2. Report on progress
   2.3. The introduction and objectives of the project are drafted.
   2.4. The literature review part is completed.

3. Discussion items
   3.3. Ray and Bao notice that we finished the literature review part and do not have any comments on that.
   3.4. Ray and Bao review and comment on the deliverables and prototype UI. It would be good that an option is provided to users to "rate" the warning and correct the potential mistakes, which in turn feedback to the model and generate more similar warnings of the highly rated type, basically like a recommendation algorithm.
   3.5. Abhi argued the effectiveness of having "supervised learning" that allows users to feedback and improve the model. In fact, to make a difference, at least thousands of ratings would be needed that this project would not likely to be escalated to this scale.
   3.6. All team members agree that this feedback system should only be done after basic functionalities of the projects are completed. It is too early to include these advanced features in our project scope.
   3.7. Ka Fong suggested for this stage, let just make the rating as a frontend feature such that the low rated warning will be simply dismissed and hidden, but not necessarily used as data point to feedback and improve the learning model.
   3.8. Work distribution is discussed, with a brief agreement. Yet, more detailed distribution should be discussed

4. Goals for the coming week
   4.3. Ka Fong will update the prototype UI and parts of the objectives to include the new expectations mentioned in 1.2 and 1.4.
   4.4. The Methodologies part will be completed by the 8th. Specifically, the UI part will be finished by Ka Fong, the API part will be finished by Abhi while the model part will be finished by Si Hou and Walker.

5. Meeting adjournment and next meeting
   The meeting was adjourned at 3:00 pm.

The next meeting will be at 2:00pm on September 9^{th} at Dr. Li's office.

7.7 Minutes of the 7th Project Meeting

Date: September 9th, 2021
Time: 2:00 pm
Place: Dr. Li's Office
Present: Walker, Abhishek, Ka Fong, Si Hou, Dr. Li
Absent: Mr. Wong, Mr. Yang
Recorder: Ka Fong

1. Approval of minutes
   The minutes of the last meeting were approved without amendment.

2. Report on progress
   2.3. The design and implementation part of the proposal report have been completed.
   2.4. Detailed work distribution has been discussed but yet to be formatted to the proposal report. Ka Fong will be responsible for UI; Walker will be responsible for the AI pipeline and correlation model; Si Hou will be responsible for the sentiment model and the event and counterparty identification model; and Abhishek will be responsible for data extraction, database and backend.
   2.5. Tentative work phase deadlines have been discussed, but not yet to be formatted to the proposal report.

3. Discussion items
   3.3. Team members present the progress of the proposal report and the prototype UI to Dr. Li.
   3.4. Dr. Li inquires about the work distribution. After hearing our primitive distributions as mentioned in 1.2, Dr. Li wishes Ka Fong will be responsible for me. Hence, Ka Fong will also be co-responsible for the correlation model and backend API.
   3.5. Dr. Li asked about the resources that SocGen could provide. Walker replies that getting hardware and data resources from SocGen would be impossible. What SocGen could provide is just high-level objectives and ideas. Dr. Li wishes that the industrial supervisors could provide financial knowledge that we computer science background students lack.
   3.6. Dr. Li suggested team members start building prototypes soon so that the industrial supervisor can give more specific, pin-pointed feedback instead of just general and broad suggestions.
   3.7. Team members ask Dr. Li about the GPU resources. Dr. Li replies GPU resources should be available for FYP projects and she will ask for it.
   3.8. Team members discuss with Dr. Li regarding the news data source and reach an agreement that we will contact the Library about this immediately after the meetings.

4. Goals for the coming week
   4.3. Walker will contact the Library regarding the news data source.
   4.4. Dr. Li will ask about the GPU resources.

4.5. All members will complete the work distribution and resource requirement part of the proposal report by the end of next day (10th September).

5. Meeting adjournment and next meeting
   The meeting was adjourned at 3:00 pm.
   The next meeting will be at 6:00pm on September 14th at Glue, Wan Chai.

7.8 Minutes of the 8th Project Meeting

Date: September 14th, 2021
Time: 6:00 pm
Place: Glue, Wan Chai
Present: Walker, Abhishek, Ka Fong, Si Hou, Mr. Ray Wong, Mr. Bao Yang
Absent: Dr. Li
Recorder: Ka Fong

1. Approval of minutes
   The minutes of the last meeting were approved without amendment.

2. Report on progress
   2.1 The work distribution and hardware part of the proposal has been completed. Overall, drafting of the proposal reports has been finished.
   2.2 The prototype UI has been updated to adapt to a new requirement as suggested in the last meeting with the industrial advisor. (options provided to user to give feedback to the alert generated)
   2.3 Getting news from the library Bloomberg or Refinitiv terminal is confirmed to be impossible since licensing issues.

3. Discussion items
   3.1 Ray mentioned that another supervisor Martin, who is a member of the credit rating team, will be able to join our project after quarantine.
   3.2 Bao reconfirmed the objectives in our proposal aligned with their expectation. However, he said that lower level objectives and deliverables would be needed in the design phase.
   3.3 Abhi asks for opinions of news data sources. Ray stated that Reuters and Bloomberg are the most preferable, CaiXin consists of a lot of useful Chinese news.
   3.4 Bao suggested not using too much time on data extraction. Even using fake dataset is possible
   3.5 Abhi disagrees with the idea of using fake dataset, as there is no way to verify the genuity and effectiveness of the warning.
   3.6 Walker restates our primitive plan of using Finhub as our news api. Ray and Bao agree that it might already be sufficient.
   3.7 Ray clarifies that provision of hardware and data resources are difficult, yet, financial knowledge is what they are able to provide.
   3.8 Ray suggested some financial approach:
   - associate sentiment score to credit rating
   - use yield rate as a proxy. Track the yield rate, investigate their effect on corporate bond price.
   - use of currency exchange rate as a proxy. Some companies might be affected when CNY rises or drops.

3.9 Ray raised the questions of the new meeting schedule as semester has started. It is confirmed we will do bi-weekly face-to-face meetings. Yet, we fail to arrive at a conclusion of a long term venue. The workspace today is too costly, while UST is too far away from their office location.

4. Goals for the coming month
   4.1 All members shall touch up and proofread the proposal before submission.
   4.2 The meeting time and long term meeting venue have to be confirmed by WhatsApp or Email.
   4.3 Data extraction shall be done by the end of this month.
   4.4 All members shall discuss to provide a deliverable deadline other than the Gantt Chart date.

5. Meeting adjournment and next meeting
   The meeting was adjourned at 7:00 pm.

7.9 Minutes of the 9th Project Meeting

Date: October 21st, 2021
Time: 6:00 pm
Place: Zoom
Present: Walker, Abhishek, Ka Fong, Si Hou, Mr. Ray, Mr. Bao, Mr. Martin
Absent: Dr. Li
Recorder: Ka Fong

6. Approval of minutes
   The minutes of the last meeting were approved without amendment.

2. Report on progress
   2.1 The news collection pipeline from Finnhub to the database is completed.
   2.2 Savio has integrated the Finbert sentiment model to the system.
   2.3 Richard has modified the frontend to display average sentiments and recent news.

3. Discussion items
   3.1 The project group welcome a new supervisor, Mr. Martin from SG credit risk team.
   3.2 Bao reintroduced the three objectives to Martin, which is i) provide sudden news alert, ii) find relationships between time series data, iii) data visualization
   3.3 After hearing our methodology, Martin provides the following suggestions
      3.3.1 Keyword matching can be done using nltk synonym
      3.3.2 Besides using Finbert, the project team may evaluate the performance of VADER and compare it with Finbert.
      3.3.3 To classify the news to different counterparties, POS tagging model can be used to identify Proper Noun
      3.3.4 Topic modeling can be done with LDA model
   3.4 Savio explain Finbert is optimized for financial news, using VADER is unnecessary
   3.5 Richard explains the Finbert news has already been sorted by counterparties, there is hence no need to use POS.

4. Goals for the coming month
   4.1 Abishek will research more on LDA topic modeling
   4.2 Si Hou will do more experiment with fine tuning the sentiment model
   4.3 Walker will construct a pipeline to collect price data to database
   4.4 Ka Fong will work on improvements on frontend and bugfix on backend.
   4.5 Martin will provide a more complete list of counterparties, including large and small companies from multi-industries.

5. Meeting adjournment and next meeting
   The meeting was adjourned at 7:00 pm.

7.10 Minutes of the 10th Project Meeting

Date: November 11th, 2021
Time: 6:00 pm
Place: Zoom
Present: Walker, Abhishek, Ka Fong, Si Hou, Mr. Ray, Mr. Bao, Mr. Martin
Absent: Dr. Li
Recorder: Ka Fong

6. Approval of minutes
   The minutes of the last meeting were approved without amendment.

7. Report on progress
   7.1 Walker has completed the ingestion pipeline of stock price to the database.
   7.2 Ka Fong has modified the frontend to display stock price and fix some crashing issues of the server.

8. Discussion items
   8.1 Si Hou expresses the concern that the sentiment model does not provide insight. Hence, he suggests an alternate approach to fine tune the sentiments with CDS price instead.
   8.2 Walker suggests the CDS price prediction should be generated from the embedding layer instead of the sentiment output.
   8.3 Ray and Martin consider Walker's attempt to be better.

9. Goals for the coming month
   9.1 Walker will input the new counterparty list provided by Martin to the system
   9.2 Si Hou will work on the new CDS sentiment model
   9.3 Abhishek will perform experiments with LDA model.

10. Meeting adjournment and next meeting
    The meeting was adjourned at 7:00 pm.

7.11 Minutes of the 11th Project Meeting

Date: December 2nd, 2021
Time: 6:00 pm
Place: Zoom
Present: Walker, Abhishek, Ka Fong, Si Hou, Mr. Ray, Mr. Martin, Mr Bao, Ms. Yiming
Absent: Dr. Li
Recorder: Ka Fong

1. Approval of minutes
   The minutes of the last meeting were approved without amendment.

2. Report on progress
   2.1 Abhishek received preliminary results on LDA.
   2.2 Ka Fong integrates naive keyword counting to the system frontend and backend.
   2.3 Walker inputted the new counterparty list to the system

3. Discussion items
   3.1 Bao will be leaving the supervisor team and a new supervisor Ms Yiming will be in charge of the project
   3.2 Abhishek share his results on LDA - the result is very noisy yet and a lot of topics is not useful.
   3.3 Walker suggests more cleaning to be done before the LDA.
   3.4 As the final exam is approaching, hence the project team is expected to make less progress these few weeks.

4. Goals for the coming month
   4.1 Abishek will research more on LDA topic modeling
   4.2 Si Hou will do more experiment with fine tuning the sentiment model

5. Meeting adjournment and next meeting
   The meeting was adjourned at 7:00 pm.

7.12 Minutes of the 12th Project Meeting

Date: December 23rd, 2021
Time: 6:00 pm
Place: Zoom
Present: Walker, Abhishek, Ka Fong, Si Hou, Mr. Ray, Mr. Martin, Ms. Yiming
Absent: Dr. Li
Recorder: Ka Fong

1. Approval of minutes
   The minutes of the last meeting were approved without amendment.

2. Report on progress
   2.1 Abhishek attempted cleaning the news before LDA with lemmatization/ stop word filtering
   2.2 Richard found a longer term kaggle news dataset for sentiment model.

3. Discussion items
   3.1 The group had made little progress due to the final exam. With the start of winter break, the project group is expected to perform more work and meet more frequently.
   3.2 Abhishek observes some preliminary patterns on LDA. Ray suggests he might start trying to name some topics.
   3.3 Regarding the keyword matching, Martin suggests the project team could try experimenting with TFIDF and cosine-similarity.
   3.4 Yiming comments that the scaling of charts is not appropriate now as the sentiment score is compressed.

4. Goals for the coming month
   4.1 Abishek will work on naming LDA topics
   4.2 Si Hou will do more experiment with FinBERT-CDS price model
   4.3 Walker will experiment on a document similarity model.
   4.4 Richard will continue to integrate teammate’s experiment result to the system and improve the charting functionalities.

5. Meeting adjournment and next meeting
   The meeting was adjourned at 7:00 pm.

7.13 Minutes of the 13th Project Meeting

Date: January 6th, 2021
Time: 6:00 pm
Place: Zoom
Present: Walker, Abhishek, Ka Fong, Si Hou, Mr. Ray, Mr. Martin, Ms. Yiming
Absent: Dr. Li
Recorder: Abhishek

1. Approval of minutes
   The minutes of the last meeting were approved without amendment.

2. Report on progress
   2.1 Ka Fong implements the counterparty addition function. On the charts, keyword frequencies are now displayed.

3. Discussion items
   3.1 Walker raised that the frontend still needs further update because the sentiments is currently hidden
   3.2 Ka Fong suggests that the keyword display is still too vague and uncategorized. Similar keywords should be grouped.
   3.3 Si Hou still needs more time to train more results with CDS.

4. Goals for the coming month
   4.1 Abishek will continue experimenting LDA topics
   4.2 Si Hou will do more experiment with FinBERT-CDS price model
   4.3 Walker will experiment on a document similarity model.
   4.4 Richard will continue to integrate teammate's experiment result to the system and improve the charting functionalities.

5. Meeting adjournment and next meeting
   The meeting was adjourned at 7:00 pm.

7.14 Minutes of the 14th Project Meeting

Date: January 18th, 2021
Time: 4:00 pm
Place: Zoom
Present: Walker, Abhishek, Ka Fong, Si Hou, Mr. Ray, Mr. Martin, Ms. Yiming
Absent: Dr. Li
Recorder: Ka Fong

1. Approval of minutes
   The minutes of the last meeting were approved without amendment.

2. Report on progress
   2.1 Walker researched using FinBERT embeddings to compare similarities between topics and new articles. Results are satisfactory for specific keywords like sue, administration and lawsuits.
   2.2 Ka Fong switch the charting library from ReCharts to HighCharts, enabling smooth panning of x-axis and displaying/ hiding data.
   2.3 Si Hou researched on matching CDS price with news. Validation loss is shown to converge slightly.

3. Discussion items
   3.1 Martin suggests the need to determine a threshold likelihood for classification. The result also need to be tested on accuracy in progress report.
   3.2 As Martin suggests, the sentiment will be put into percent stack for clearer visualization. Keywords should be collected in categories as there are too many similar keywords now. Yiming suggests that topic modeling update to be integrated to frontend.
   3.3 Si Hou express the slow training issues with his CDS model, Martin suggests only the last few layers of the model should be trained.
   3.4 Ray suggests us to use some time to organize our research and settle what to put in the progress report. A table of content should be decided.

4. Goals for the coming month
   4.1 Ka Fong will make the sentiments display to percentage stack instead
   4.2 Si Hou will train the CDS with more counterparties.
   4.3 The team will prepare a table of content for the progress report.

5. Meeting adjournment and next meeting
   The meeting was adjourned at 5:00 pm.

7.15 Minutes of the 15th Project Meeting

Date: January 27th, 2022
Time: 3:30 pm
Place: Zoom
Present: Walker, Abhishek, Si Hou, Mr. Ray, Mr. Martin, Ms. Yiming
Absent: Ka Fong, Dr. Li
Recorder: Si Hou

1. Approval of minutes
   The minutes of the last meeting were approved without amendment.

2. Report on progress
   2.1 Ka Fong made plenty of progress on the frontend, which includes
      2.1.1 Improve chart visuals by plotting sentiments in percentage stack, display stock price in accum. percentage change
      2.1.2 Displaying sentiments and keywords in news list
      2.1.3 Interaction between chart and news list
      2.1.4 Making side-menu mobile friendly
   2.2 Abhishek discover LDA results work much better on individual counterparties than on the whole corpus
   2.3 The team starts working on progress report (mainly the methodology)

3. Discussion items
   3.1 Yiming raises that the stock price plotting might be erroneous as the mapping keep changing
   3.2 Regarding the FinBERT model, Ray suggests besides using MAE as evaluation, Si Hou should also plot it out on a line graph.
   3.3 Walker and Abhishek raise a suggestion to use the LDA model as a keyword suggestion system and output the difference of topics outputted over time.

4. Goals for the coming month
   4.1 Ka Fong & Abhi to integrate LDA keyword suggestions to the system backend and frontend.
   4.2 Si Hou will evaluate the FinBERT-CDS model by plotting the graph.
   4.3 Walker will code the code to word documents.
   4.4 The team will continue to work on progress report

5. Meeting adjournment and next meeting
   The meeting was adjourned at 5:00 pm

7.16 Minutes of the 16th Project Meeting

Date: February 10th, 2022
Time: 3:30 pm
Place: Zoom
Present: Ka Fong, Walker, Abhishek, Si Hou, Dr. Li
Absent: Mr. Ray
Recorder: Si Hou

1. Approval of minutes
   The minutes of the last meeting were approved without amendment.

2. Report on progress
   2.1. Ka Fong made plenty of progress in the UI, including:
      2.1.1. Improve chart visuals by plotting sentiments in percentage stack, display stock price in accum. percentage change
      2.1.2. Displaying sentiments and keywords in news list
      2.1.3. Interaction between chart and news list
      2.1.4. Making side-menu mobile friendly
   2.2. Abhi made progress in his research on the LDA model and report its usage to Dr. Li
   2.3. Savio reported his reserach in CDS price model to Dr. Li
   2.4. Walker reported his research in consine similarity to Dr. Li

3. Discussion item
   3.1. Discussed the improvement of the UI, and how the user can manage the system
   3.2. Discussed how the progress report can be improved
   3.3. Discussed how to better present to attract the readers

4. Goals for the coming month
   4.1 The team should make the system as a story to tell the purpose of the system
   4.2 Wrap up the tasks together and connect all the components as a story

5. Meeting adjournment and next meeting
   No meeting is being scheduled for now

7.17 Minutes of the 17th Project Meeting

Date: 21th March, 2022
Time: 3:30 pm
Place: Zoom
Present: Ka Fong, Walker, Abhishek, Si Hou, Mr. Martin
Absent: Mr. Ray
Recorder: Si Hou, Walker

1. Approval of minutes
   The minutes of the last meeting were approved without amendment.

2. Report on progress
   2.1. Ka Fong updated the UI. The charts are easier to visualized and the keywords feature are now added
   2.2. Abhishek implements the alert feature using z-score
   2.3. Walker made his research and insights in embeddings and FinBERT model
   2.4. Savio reported his research in the model predicting credit default swap price

3. Discussion item
   3.1. How to make deeper research in the CDS price model and the topic modeling
   3.2. How to visually describe the model pipeline
   3.3. How to make the system clearer to users

4. Goals for the coming month
   4.1. Cover what each curve stands for in the graph and make it easier to visualize
   4.2. Make further reserach in coefficient matrix of different things with price changes
   4.3. Make further research on the sentiment model

5. Meeting adjustment and next meeting
   Schedule a meeting with Ray before submitting a final report

7.18 Minutes of the 18th Project Meeting

Date: 11th April, 2022
Time: 3:30 pm
Place: Zoom
Present: Ka Fong, Walker, Abhishek, Si Hou, Mr. Ray
Absent: Dr. Li
Recorder: Si Hou

1. Approval of minutes
   The minutes of the last meeting were approved without amendment.

2. Report on progress
   2.1. Ka Fong implemented the alert feature and users can determine whether the alert is accurate or not inside the system
   2.2. Keywords can now be added and tracked per counterparty and also globally

3. Discussion item
   3.1. How to make the UI clearer and make the users clearly know what is the whole system about
   3.2. How to do a better presentation

4. Goals for the coming month
   4.1. Prepare a story to better present the whole system to the users
   4.2. Add clearer instructions inside the system such that users know clearly what the whole system is about

5. Meeting adjustment and next meeting
   No further meeting has been scheduled yet

7.19 Minutes of the 19th Project Meeting

Date: 19th April, 2022
Time: 2:00 pm
Place: Zoom
Present: Ka Fong, Walker, Abhishek, Si Hou, Dr. Li
Absent: Mr. Ray
Recorder: Abhishek

1. Approval of minutes
    a. The minutes of the last meeting were approved without amendment.

2. Report on progress
    a. Team demonstrates the progress and final draft of the report to Cindy.
    b. Ka Fong demonstrates the key events based on the alerts on the counterparties such as JP Morgan, and Tesla.

3. Discussion item
    a. How to further enhance our story telling technique from the existing UI
    b. Dos and Don’ts in the presentation
    c. Walker asked Cindy if she remembers any issues with our progress report so that we could act upon those issues in our final report.

4. Goals on the coming month
    a. Submit the final report.
    b. Report the UAT results.
    c. Prepare a story to better present the whole system to the users.

5. Meeting adjustment and next meeting
    No further meeting has been scheduled yet

7.20 Minutes of the 20th Project Meeting

Date: 19th April, 2022
Time: 4:30 pm
Place: Zoom
Present: Ka Fong, Walker, Abhishek, Si Hou, Ted Spaeth
Absent: Mr. Ray, Dr. Li
Recorder: Abhishek

1. Approval of minutes
    a. The minutes of the last meeting were approved without amendment.

2. Report on progress
    a. Team demonstrates the progress and final draft of the report to Ted.

3. Discussion item
    a. Walker enquires on the ways team can make the report more comprehensive since we lost points in that area from the second reader in the progress report.
    b. Grammatical and complex sentences error correction
    c. How to further enhance our story telling technique from the existing UI
    d. Dos and Don'ts in the report and presentation

4. Goals on the coming month
    a. Submit the final report.
    b. Report the UAT results.
    c. Prepare a story to better present the whole system to the users.

5. Meeting adjustment and next meeting
    a. No further meeting has been scheduled yet

8. Acknowledgments

Ray from Societe Generale for proposing the partnership with HKUST and being the main liaison for this project.
Professor Cindy Li for being our advisor, guiding us through the project and providing counsel at critical times to put our worries at ease. Her feedback and confidence in our project was also critical for our success.
Martin Autier from Societe Generale for providing his invaluable Data Science expertise and input, many of which turned into features or implementation methods.
Lastly, various current and pat SG employees which helped guide us through our project at various times and provide invaluable insight on the needs of an investment bank.