SentiVol

A self service risk analytics platform that computes and leverages X (formerly Twitter) sentiment to forecast the VIX (US volatility index).

A project proposal report by:

Andrei Rizea
SID: 220300153

Table of Contents

Abstract

This proposal is for a project that will explore whether sentiment aggregated from tweets and comments from X (formerly Twitter) can be used as a signal for US stock market volatility and specifically, forecasting VIX. Building on similar work performed by Bollen et al. (2011), this proposal outlines a scalable data collection and analysis pipeline that combines advanced natural language processing (NLP), machine learning, and statistical tests as well as the infrastructure required to make this app operable. Bollen et al. (2011) applied two different sentiment analysis tools. The first tool (OpinionFinder) gave an overall positive or negative score, while the second tool (GPOMS) measured several mood dimensions like calm and happiness. To capture the complex relationship between mood and market fluctuations, they employed a Self-Organizing Fuzzy Neural Network (SOFNN) as it is specialised in capturing non linear trends and relations such as those seen in text and speech.

The platform proposed will be a self service application, where users can select custom date ranges and keywords to filter X tweets to be analysed. This will trigger automated machine learning models which will display interactive visualisations of sentiment trends, volatility metrics, and predictive forecasts of the VIX. The aim is to offer a tool that not only forecasts volatility spikes but also makes complex data analysis accessible to all, ranging from sentiment trends to market forecasts.

This proposal dives into the main underlying aims of SentiVol, where the idea came from as well as assessing research conducted by Bollen et al. (2011) which is highly relevant for this project. In addition a plan to gather requirements will be outlined as well as an implementation plan providing an overview of the steps and tech stack that SentiVol will use.

Glossary

Social Media Sentiment: The overall mood for a given topic expressed by users on social media platforms.

X (formerly Twitter): A social media application used by 650 million users every month.

Volatility: A measure of the degree of variation in a stock price over time. Higher volatility signifies larger price swings indicating instability.

VIX (Volatility Index): It's a real-time market index representing the market's expectations for volatility over the coming 30 days used to measure the level of risk, fear, or stress in the market when making investment decisions.

Sentiment Analysis: Processing and analysis to determine the overall mood or trend of a block of text. For example a simple sentiment analysis model groups text into positive and negative buckets.

Granger Causality: A statistical test used to determine whether there is a casual relation between 2 variables. In this context, does social media posts directly cause a shift in market volatility or is it just by chance.

Supabase: Supabase functions as an open-source backend-as-a-service platform which generates PostgreSQL databases automatically while providing real-time subscriptions and authentication features and storage capabilities to support developers in building scalable applications.

Natural Language Processing (NLP): A subfield of artificial intelligence which enables computers to process and generate as well as understand human language.

Backtesting: Testing a trading strategy or model against historical market data to evaluate its potential performance and identify areas for improvement.

OpinionFinder: A sentiment analysis tool OpinionFinder uses a lexicon-based system to analyze text documents by identifying word polarity which results in a document-level sentiment score.

GPOMS (Google-Profile of Mood States): The GPOMS sentiment analysis system evaluates textual data to measure multiple mood states including calmness and alertness and happiness.

SOFNN (Self-Organizing Fuzzy Neural Network): The Self-Organizing Fuzzy Neural Network uses neural network learning and fuzzy logic to process uncertain data while modeling complex non-linear relationships in information.

Greenfield: A new piece of software developed from the ground up using no existing or legacy code.

Project Aims

Scalable data pipeline

SentiVol must have a scalable data gathering module leveraging X and yFinance APIs, with raw data storage in a Supabase Postgres database that is generic enough to replace the sources of the data easily. This module will be designed to run continuously during market hours to ensure SentiVol has access to the latest data. It will be architectured to horizontally scale leveraging AWS Kubernetes containers ensuring that increased data loads do not affect performance.

Robust data processing

Data preprocessing and sentiment analysis modules will be implemented that will clean and aggregate sentiment. This module will handle cleaning, filtering, and normalizing tweets to remove noise. Lexicon based methods such as VADER will then be implemented to calculate overall scores for tweets based on the words used. As sentiment analysis relies on natural language processing, the chance of errors are high so robust error handling and validation steps will be added to maintain data quality.

Reliable predictive modeling

A Self Organizing Fuzzy Neural Network (SOFNN) will be implemented to forecast market volatility using the computed X sentiment scores prior. Due to the unpredictable and black box nature of machine learning, its performance should be compared with Random Forests and ARIMAX models to reinforce SOFNN outputs and hence, establish clear causal links between sentiment and market volatility. The results will determine if the forecasts are consistent across different market conditions as well as calculating if certain parameters from the sentiment analysis module do in fact improve forecasting abilities. Without reliable forecasts, risk calculations will be harder and more unpredictable.

Intuitive user interface

A UI will be the gateway between the machine learning forecast model and the user. It should be simple, intuitive and lightweight and should allow users to select custom time ranges and filters such as entering specific keywords or hashtags they would like to be included in the sentiment analysis. The platform will trigger forecasting and subsequently display interactive visualisations such as charts, graphs and tables within an isolated dashboard page. In addition, useful statistics and metrics will be shown to users that will allow them to draw their own conclusions if necessary such as sentiment trends and VIX performance indicators. The interface will also ensure that both technical and non-technical users can easily access, interpret, and act on the information output by SentiVol.

Comprehensive testing framework

An end-to-end (E2E) testing module must be implemented using a range of unit, integration and usability tests leveraging industry standard libraries such as Pytest, Vitest and Cypress (SE5). This

will ensure data processing accuracy as well as overall app health by checking that all features and modules are working as load and functionality increases.

Problem Statement

Forecasting stock market volatility is still a challenge not only for GS but also investors and traders globally. Traditional models that rely on historical price data and economic indicators fall short when it comes to anticipating sudden market swings as those indicators are in fact reactionary to a physical change such as inflation figures being released or tariff announcements. This is because the stock market can be seen as a game of trust where if a CEO of a fortune 500 company announces some unfavorable changes to the company structure, shareholders may begin to sell off their shares causing the price to dip. Arcuri et al., (2023) looked at 149 fake news events between 2009 and 2017 and found that there was a 90% significance one day after the fake news was released to indicate a significant market swing, especially in the short term. This evidence shows that there is some sort of causal relation between the media and the markets illustrating that media posts can cause severe swings in markets. However, currently there is no established method of computing and predicting this relationship between sentiment and market movement, especially for the VIX. This leaves a gap wide open that, if solved, will ensure companies have better risk calculations allowing for more informed decisions in the markets.

This introduces the question of whether or not social media platforms like X, that allows for thousands of daily tweets which capture public sentiment, can be used to predict a reactionary spike in the VIX index. Despite there being millions of available tweets, extracting a reliable sentiment signal is difficult. This is due to the noise in these tweets hiding the true sentiment of the writer from unstructured and random content as well as spam content which may bias the model. However, previous studies, such as Bollen et al. (2011), have shown that Twitter sentiment may have some predictive value for market direction but they have left open the question of whether sentiment can specifically forecast volatility spikes.

The aim of SentiVol is to determine whether a real time measure of X sentiment can act as a strong indicator for market volatility. If sentiment shifts can be shown to predate significant market movements, this information would be beneficial in risk detection and extremely useful for investors and traders.

Furthermore, many investment banks and hedge funds already leverage quantitative analysis through data pipelines, machine learning and statistical analysis providing leverage and differentiation between their competitors (Investopedia). Therefore, once SentiVol is released, it will provide new leverages to improve risk calculation algorithms and ensure stakeholders maintain favourable positions in markets.

Paper Assessment

Paper assessed: Bollen, J., Mao, H. and Zeng, X. (2011), “Twitter mood predicts the stock market.”

Bollen et al.’s work is widely recognised as a pioneering study in linking social media sentiment with financial market behavior. The writers collected millions of tweets during a turbulent market period in 2008, employed sentiment analysis tools, specifically OpinionFinder, GPOMS and SOFNN and also used statistical tests like Granger causality to demonstrate that certain mood statistics could predict the direction of the Dow Jones Industrial Average (DJIA).

Quality

The study has a large number of citations, nearly 4000 (Science Direct), meaning it is a well received and ground breaking study. It introduced a new dimension of possible market indicators, suggesting that Twitter mood can be used to improve the prediction of market swings. It was peer reviewed by the Journal of Computational Science ensuring its validity and quality.

A Self Organizing Fuzzy Neural Network (SOFNN) was used for the first time to establish the connection between moods and market swings. SOFNN deals with the non-linear and uncertain, and often noisy, aspect of social media posts. Also, the application of statistical tests strengthened its claims, and it was found that some moods can actually be used to predict movements. Specifically, calm sentiment tweets were found to be 87.6% accurate in predicting DJIA movements.

Research methods

Bollen et al. (2011) employed a multi step approach in linking Twitter mood with market behavior. They started by gathering and cleaning millions of tweets during 2008. Two sentiment analysis tools were applied: OpinionFinder, which gives a general positive or negative score, and GPOMS, which measures six distinct mood dimensions from calm and happy to anger. These produced daily indices that captured the combined sentiment of Twitter users. To examine the predictive relationships, they performed Granger causality tests to see if historical mood data could predict changes in the DJIA. Finally, they employed a SOFNN to model the nonlinear dynamics between the aggregated indices and market movements to evidence the predictive power.

However, some argue that the study is linked to a particular time frame in 2008 which may have limited the robustness and accuracy of its conclusions. Also, the original work focused on market direction rather than volatility, which is a gap that the SentiVol will aim to address.

Relevance

Bollen et al.’s work provides the conceptual and significant foundation for SentiVol, demonstrating that social media sentiment can act as a market indicator making this study extremely relevant. While their study focused on market movements, SentiVol looks into volatility which is of greater use in risk calculations by investors and traders. The techniques used by Bollen et al., such as sentiment aggregation and causality testing, will be adapted and extended on by using more advanced NLP models and a broader data set which is filterable by users on a self service basis.

Citation breadth

The extensive citation of Bollen et al. reinforces its impact across disciplines. It has nearly 4000 citations (as of April 2025) on ScienceDirect alone (Science Direct) and has inspired further research into behavioral finance and machine learning applications in market prediction including Lachanski, M. and Pav, S. (2017) who revisited and extended the Twitter (now X) sentiment model 6 years later with nearly 10,000 downloads as of April 2025.

Subsequent studies have both built on and critiqued its methodologies, making it a good baseline reference point for SentiVol which will acknowledge these contributions while also explicitly addressing its critiques such as the fixed time period bias mentioned above.

Requirements Capture & Research

In order to better understand the requirements and approaches needed for SentiVol, research is required to be carried out as well as data gathering sessions.

Research

The system requirements for SentiVol will be captured through an organised methodology. Initially, a comprehensive literature review will be conducted. Two essential academic works by Bollen (2011) and Lachanski (2017) will be studied, presenting established methods for connecting social media sentiment analysis with market data. The research presents methods for data collection and sentiment analysis which uses both lexicon-based and multi-dimensional mood metrics along with predictive modeling through statistical tests including Granger causality and machine learning models. In addition, the X API and Supabase technical documentations will be reviewed providing information about data retrieval together with system constraints and integration limitations.

Data gathering

The next step requires stakeholder engagement which is a step aside from development (SE7). Structured interviews and surveys with a mix of open and closed questions will be created to gather data from potential users who work as traders and risk managers and data analysts providing a mix of quantitative and qualitative data. The participants will be from a wide range of ages and experience levels to ensure an unbiased sample group and cover all major intended stakeholder groups of the app.

The planned sessions will obtain comprehensive user stories together with particular requirements from them. For example, a trader would need tweet filters that monitor market-moving events whilst a risk manager would want to see daily sentiment patterns which relate to volatility changes. The system development must solve real operational issues and maintain industry standards which this stakeholder engagement process will uncover and highlight.

Rapid prototyping of essential system components within the UI will allow for an early assessment of core functionalities (SE7). The evaluation of prototypes happens through repeated sessions where feedback gets documented until it gets integrated with the main requirements. A requirements traceability matrix will then be implemented to connect every requirement to its original owner from

academic research, stakeholder feedback or technical limitations while ensuring all items remain quantifiable and most importantly measurable.

Data ingestion and data processing, analytical modeling and user interface form the distinct categories into which the requirements will be organised. Prioritisation sessions will be conducted with stakeholders to establish core features for the initial minimum viable product (MVP) whilst also scheduling requirements for future implementation (SE7).

The project will maintain a scheduled review process which also includes updates. The system will utilise scheduled meetings combined with feedback loops to track modifications in market conditions, technological progress and user requirements. The requirements will stay up to date through continuous reviews which enable the system to address new challenges.

Must have requirements

The below are a mix of functional and nonfunctional requirements SentiVol must have in order to be deemed a success.

Functional requirements:

1. Aggregate X sentiment based on filter criteria accurately ensuring no nuances in the natural language processing
2. A system in place to fetch data. Both X tweets and financial markets data from public apis such as yFinance
3. Aggregated sentiment scores stored in a database that supports real time integrations
4. Forecasting model that achieves at least 80% accuracy in predicting VIX swings based off of sentiment trends
5. Error handling ability to ensure the app can restart in the event of failure
6. Intuitive user interface to allow stakeholders to conduct their own research into VIX movements by inputting criterias to filter our social media posts
7. UI dashboard to visualise forecasting an in built charting software to ensure SentiVol is an all-in-one analytics platform for traders making it more seamless for them to use

Non-functional requirements:

1. Performance of the app should remain constant regardless of data volumes and user load ensuring a seamless experience
2. The app should be deployed in such a manner to allow for quick restarts to ensure minimal downtime
3. The UI must be intuitive and seamless to use
4. SentiVol should be maintainable by any developer regardless of their knowledge of the inner workings of the app
5. The app should be coded to allow for easy integration of new APIs in the future when they become available. For example a new and faster market data streaming API.

Implementation

The implementation of SentiVol will be carried out in modular stages, integrating robust data collection, processing, analysis, and a UI to ensure ease of maintainability and allow for isolated improvements, subsequently ensuring the robustness of the system (SE2).

Data collection

There will be 2 main data sources; social media and market data. X data will be accessed via the X API using the Tweepy python library. Although current API access is limited, necessary credentials under an academic research license will need to be requested. VIX data will be sourced from the yFinance APIs providing real time data. This data will be stored daily and a data pipeline job should be scheduled to run at the end of each day to pull in the latest day, after an initial -10 year data population, of data to save on compute.

Supabase

Supabase is chosen for its simplicity, scalability, and real-time capabilities. It's essentially a PostgreSQL database which will allow for storing user preferences, aggregated sentiment and more.

Data preprocessing

Data stored in Supabase will be retrieved for preprocessing. This will be coded in Python due to its vast set of libraries such as Pandas ensuring fast memory processing allowing the models to be run on demand by users. The following steps will be applied:

1. Text Cleaning: Remove URLs, user mentions, special characters, and non-relevant symbols.
2. Normalisation: Convert all text to lowercase and remove stop words using libraries such as NLTK or spaCy.

Sentiment analysis

Two parallel python pipelines will be set up:

1. Lexicon-Based pipeline: VADER will be used to give sentiment scores to each tweet, which will then be combined daily to form aggregated averages.
2. Machine Learning-Based pipeline: A transformer based model will be used such as SOFNN that will be applied to capture deeper sentiment nuances. Individual tweets will be scored and then the daily aggregates will be stored back into Supabase. The precomputed aggregates will be from common user filters that SentiVol has noticed to be commonly used filters which will save on computation time.

Time series analysis and predictive modeling

Before the app can be published, the model has to be created and tuned. The model will be coded in Python using libraries such as pandas, NumPy, statsmodels and or scikit-learn. Some of the steps that will be followed are below:

1. The daily sentiment index will be merged with the VIX data retrieved from Supabase.
2. Exploratory data analysis will be conducted using visualisations using Matplotlib to assess trends and correlations.
3. Statistical tests, including Pearson correlation and Granger causality, will be applied to determine if past sentiment can causally predict future volatility.
4. A forecasting model will be built using regression techniques such as Random Forests and time-series models such as GARCH augmented with exogenous sentiment variables. Models incorporating sentiment features will be compared against baselines that rely solely on historical market data as that is what the industry is currently relying on for risk forecasting.
5. Once a model is curated and tested, this will be modularised so that the SentiVol web application can reference this working forecasting model.

User interface

The UI will be built using React Typescript with VITE and will connect with the Supabase via API and will be deployed using Vercel. This tech stack follows web industry standards and will be developed following best practices (SE3). The UI will:

1. Allow user defined inputs: Users can select a desired time period and input specific keywords or hashtags related to their needs.
2. Trigger analysis on demand: With the click of a button, the system will run the complete analysis using the modularised model completed in the previous implementation step.
3. Display interactive visualisations: The UI will present results through dynamic dashboards that include charts, graphs, and tables. These visualisations will show time-series data for sentiment and volatility, correlation analyses, and forecasting outputs. D3, a JS library, will be used to provide the dynamic metrics and visuals whilst also using AGgrid to house all the static data in tabular form.

Version control

Github will be used to track and store the codebase for this project which is common practice in the industry.

Modularity

The project will be structured into independent modules with the aim for them to integrate seamlessly together. This modular approach allows for project decomposition ensuring each aspect of SentiVol is coded to the highest of quality. This modular design architecture improves maintainability and allows for independent updates or improvements to each component ensuring robust code (SE2).

1. Data collection module: Connects to the X API and financial data sources, storing data in Supabase.
2. Preprocessing module: Retrieves and cleans raw data from Supabase.
3. Sentiment analysis module: Processes the cleaned data to generate sentiment scores, saving outputs back to Supabase.
4. Analysis & modeling module: Merges and analyzes sentiment and market data, performs statistical tests, and develops predictive models.
5. Frontend module: A UI enabling users to trigger analysis, and view interactive results.

Testing and Evaluation

Testing is integral to ensuring that SentiVol functions reliably, robustly and provides meaningful and accurate insights regardless of data volume and user load (SE5).

Data quality and preprocessing testing

1. Functional tests: Scripts will cross-reference data from the X API and financial providers against known benchmarks to ensure accuracy and consistency (SE11).
2. Unit tests: These are tests that will cover the base logic of the preprocessing such as cleaning methods utilising regex (SE11).

Sentiment analysis evaluation

1. Benchmark comparison: Sentiment scores from VADER and transformer-based models will be compared against a manually annotated dataset. Metrics such as accuracy and F1-score will be calculated.
2. Model selection: The performance of both pipelines will be assessed to select the method that most accurately captures shifts in sentiment during key market events.

Time series and forecasting model testing

1. Statistical testing: Pearson correlation and Granger causality tests will be conducted to evaluate if historical sentiment data significantly predicts future volatility.
2. Predictive modeling: Forecasting accuracy will be measured using RMSE, MAE, and classification metrics like precision and recall.
3. Backtesting: The models will be run and tested against old data to validate their predictive ability.

UI testing

1. Usability testing: The web interface will be tested with a group of users representing different technical backgrounds to ensure that parameter selection and data visualisations are intuitive.
2. Integration testing: End-to-end tests using Cypress will verify that when a user selects a time period and key terms, the backend processes the data correctly and updates the dashboard in real time (SE11).

Robustness and sensitivity analysis

1. Parameter sensitivity: Models will be run with varying lag periods and alternative feature sets to assess stability.
2. Event focused evaluation: Special analysis will be conducted during known periods of market turbulence to determine if sentiment signals are particularly strong during these times.

Conclusion

SentiVol uses modern NLP techniques and machine learning to perform sentiment analysis to discover a causal relationship between social media and financial market volatility forecasting. Users can utilise the self service UI to customise their analysis through time period selection and key term choice which automatically triggers data processing and forecasting before presenting results through charts, graphs and tables.

Though this is a greenfield project, it extends the study by Bollen et al.'s (2011) original work by focusing on market volatility instead of direction while allowing users to analyze any time period instead of being limited to 2008. The project will use advanced sentiment analysis methods together with statistical testing including Granger causality tests and forecasting to determine if public mood fluctuations can predict market volatility. The essential goal remains to establish cause-effect relationships instead of accidental correlations because machine learning analysis always faces this possibility.

The identification of a strong predictive relationship would enable the development of useful risk management tools and trading strategies. On the other hand, if a weak relationship is found, the project will provide important theoretical insights about the forecasting capabilities of social media sentiment advancing behavioural finance theories.

References

Arcuri, M.C., Gandolfi, G. and Russo, I. (2023) 'Does fake news impact stock returns? Evidence from US and EU stock markets', Journal of Economics and Business, 125–126, 106130. [online] Available at: https://doi.org/10.1016/j.jeconbus.2023.106130 (Accessed: 2025).

Bollen, J., Mao, H. and Zeng, X. (2011) "Twitter mood predicts the stock market." *Journal of Computational Science*, 2(1), pp. 1–8. [online] Available at: https://doi.org/10.1016/j.jocs.2010.12.007 (Accessed: 2025).

Investopedia (2023) "CBOE Volatility Index (VIX)." [online] Available at: https://www.investopedia.com (Accessed: 2025).

Investopedia (n.d.) "Using Quantitative Investment Strategies." [online] Available at: https://www.investopedia.com/articles/trading/09/quant-strategies.asp (Accessed: 2025).

Lachanski, M. and Pav, S. (2017) "Shy of the Character Limit: 'Twitter Mood Predicts the Stock Market' Revisited." *Econ Journal Watch*, 14(3), pp. 302–345.

Science Direct (n.d.) [online] Available at: https://www.sciencedirect.com (Accessed: 2025).

X, Inc. (n.d.) "X API Documentation, Historical Search." [online] Available at: https://developer.x.com/en (Accessed: 2025).

World Wide Web Consortium (W3C) (n.d.) WCAG 2 Overview | Web Accessibility Initiative (WAI). [online] Available at: https://www.w3.org/WAI/standards-guidelines/wcag (Accessed: 2025).

Specialised KSBs

SE2: Undertake analysis and design to create artefacts, such as use cases to produce robust software

SE3: Produce high quality code with sound syntax in at least one language following best practices and standards.

SE5: Test code to ensure that the functional and non-functional requirements have been met.

SE7: How to operate at all stages of the software development lifecycle.

SE11: How to perform functional and unit testing.