**The Pulse of the Market: Visualizing the Lag-Correlation between Social Sentiment and Asset Volatility**

**Introduction & Goals**

Traditional financial models often assume market rationality. However, the historic GameStop (GME) Short Squeeze of early 2021 demonstrated that asset prices can be driven by the "irrational exuberance" of retail investors. Existing visual analytics tools typically simplify sentiment into a binary dimension (Positive/Negative). This oversimplification fails to capture complex market psychology—for instance, while "Panic" and "Anger" are both negative, their temporal impacts on the GME price volatility might differ significantly.

This project aims to build a visual analytics poster that leverages the 12-class fine-grained emotion taxonomy proposed by *StockEmotions* (Lee et al., 2023). By revisiting the GME event with this novel lens, we seek to identify specific emotional signals that served as "leading indicators" for the 2021 crash and map the evolutionary trajectory of crowd psychology during this historic market anomaly.

**Research Questions**

RQ1: Among the 12 specific emotions, which ones act as the strongest leading indicators for price volatility?

RQ2: To what extent is market volatility driven by the arousal of the discussion rather than the valence?

RQ3: How does the composition of community sentiment evolve over time during periods of significant market movement?

**Dataset**

We will combine textual data with financial time-series data, specifically filtering for the GameStop ($GME) timeline (Jan 2021 - Feb 2021):

1. **Training ground truth:** <https://github.com/adlnlp/StockEmotions>

**Size**: 10,000 expert-annotated comments

**Variables**: text & emotion\_labels

We will fine-tune a pre-trained DistilBERT model using the StockEmotions dataset. This enables the classification of 12 fine-grained emotions(e.g.,Panic, belief), surpassing the limitations of generic sentiment analysis.

1. **Text data:** [Reddit WallStreetBets Posts (Kaggle)](https://www.google.com/search?q=https://www.kaggle.com/datasets/gpreda/reddit-wallstreetbets-posts)

**Size**: after filtered ~500000 comments

**Variables**: Timestamp, text, SentimentWeight, Derived Attributes

We will utilize the WallStreetBets dataset from Kaggle as our primary set The fine-tuned model is applied to the data to generate specific emotion labels and confidence scores for every comment.

1. **Market Data:** Daily stock price of GME and volume data via Yahoo Finance

**Source:** Yahoo Finance API (yfinance).

**Variables (Attributes): Date, ClosePrice, Volume, High/Low Volatility**

We merge historical price data via Yahoo Finance and merge the dataset with the Kaggle one using temporal resampling(daily aggregation) to create a unified time-series dataset where sentiment metrics align with stock price movement

**Method & Tools**

Pipeline: python for data processing and metric calculation

Visualization: using matplotlib or the tools we learnt from labs to generate the visualization plot.

User task: identify leading signals; correlate intensity with volatility; trace narrative evolution.

**Visualization**

Vis 1: The Time-Lagged Correlation Analysis (Type: Cross-Correlation Function (CCF) Plot)

* Significance: To validate RQ1, this chart statistically identifies whether specific emotions (e.g., Panic) exhibit a peak correlation with price volatility at negative time lags, confirming their role as statistical leading indicators.

Vis 2: The Arousal-Anomaly Scatter Plot (Type: Scatter Plot)

* Significance: Addressing RQ2, this visualization examines whether abnormal trading volumes are correlated more strongly with the intensity of emotional arousal (y-axis) rather than the direction of sentiment (x-axis).

Vis 3: The Sentiment Evolution Stream (Type: 100% Stacked Area Chart)

* Significance: Answering RQ3, this chart visualizes the structural shift in community sentiment over time, illustrating how negative emotions (like Fear) progressively displace positive ones (like Optimism) during market stress events.

Vis 4: The Market State Heatmap (Type: Categorical Heatmap)

* Significance: This matrix reveals the structural relationship between discrete market states (e.g., "High Volatility," "Crash") and the dominant emotion of the day, uncovering distinct emotional signatures for different market phases.

Vis 5: The Predictive Alignment View(Type: Dual-Axis Line Chart )

* Significance: Providing intuitive visual proof for RQ1, this chart superimposes the "Panic" metric (shifted forward by the calculated optimal lag) onto the "Volatility" metric to demonstrate their temporal alignment and predictive power.

Sketches can be seen in Appendix.

**Deliverables for the Interim Presentation**

* Datasets: Both raw inputs (Reddit/Yahoo) and the processed master CSV containing aligned timestamps and predicted 12-emotion labels.
* Static Visualizations: High-fidelity static prototypes for Vis 1, Vis 3, and Vis 5, demonstrating core data trends without interactivity.
* Interaction Plan: A concise guide outlining planned user controls (e.g., time-shift sliders, tooltips, brushing) for each visualization.
* Codebase: Python notebooks for the NLP pipeline (DistilBERT fine-tuning), data cleaning, and static plot generation.

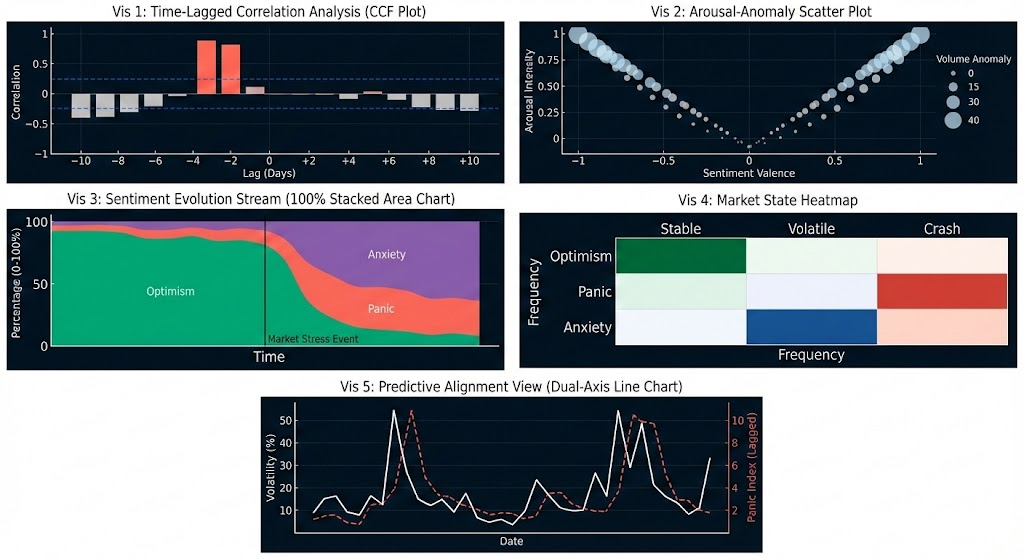
**Roles**

* Yinhan Gao (Data/NLP): Builds the backend pipeline; fine-tunes the DistilBERT model for emotion classification and merges multi-source datasets.
* Ke Ning (Visualization): Develops the 5 visualization idioms using Python; ensures design consistency and implements interactive features.
* Yanpei Yu (Analysis): Conducts statistical tests (e.g., Lag-Correlation); synthesizes insights to address RQs and compiles the final academic report.

**Weekly Timeline**

* Week 4 (Data Prep): Download GME data and perform EDA.
* Week 5 (Modeling): Fine-tune model on StockEmotions and classify GME comments.
* Week 6 (Visualization): Develop and refine Vis 1 through Vis 5.
* Week 7 (Delivery): Integrate charts into the final poster and record presentation.

**Appendix for the reference of our visualizations**



Source: This comprehensive visualization design framework was authored and optimized by Gemini

**Reference**

Lee, J., Youn, H. L., Poon, J., & Han, S. C. (2023). Stockemotions: Discover investor emotions for financial sentiment analysis and multivariate time series. *arXiv preprint arXiv:2301.09279*.