IE 684 Web Mining Project Outline
TweetMiner: NVIDIA Market Sentiment Analysis

Nan Chen[2033010], YuWei Liu[2035212], Yunuo Wu[2033410], Zhiqi Yang[2110635], Hanshi Zhang[2110222]
Team 1

April 6, 2025

1 Problem Statement

Financial markets require timely indicators to capture rapid sentiment shifts around volatile assets like NVIDIA stock. Traditional metrics (e.g., VIX, price volatility) lack granularity in tracking crowd-driven sentiment dynamics on social platforms. With the rise of retail investors on X/Twitter, we propose a targeted framework to address:

- **Prediction Enhancement**: How can social media data improve short-term stock forecasts?
- **Content Analysis**:
  - Identify key financial topics and terminology associated with NVIDIA in Twitter discussions.
  - Detect structural patterns linking keywords to sentiment shifts.
  - Integrate analysis of news articles with social media discussions to enhance sentiment understanding.
- **User Behavior**:
  - Cluster users by discussion topics and engagement patterns.
  - Quantify the correlation between sentiment volatility and posting frequency.
- **Temporal Dynamics**:
  - Map sentiment diffusion pathways during critical events.
  - Analyze synchronization between news cycles and social media discussions.
  - Explore news cycle timing and its influence on social media sentiment.

2 Dataset Statement

2.1 Data Constraints

Historical financial news datasets often require costly proprietary subscriptions. To ensure cost-effectiveness, we utilize:

- **Social Media Data**: 100,000+ NVIDIA-related tweets in 2022 and 2023 from Kaggle https://www.kaggle.com/datasets/soheiltehranipour/100k-nvidia-tweets.
- **Market Data**: Historical price and volume data for NVIDIA, retrieved using the `yfinance` Python library.
- **Macro Sentiment Proxy**: VIX index, serving as a general measure of investor uncertainty.
- **News Data Web-Scraped**: Sources from major financial news websites such as CNBC, Yahoo Finance, and Bloomberg.

2.2 Dataset Characteristics

– **Temporal Coverage**: 2022-2023.

– **Metadata**:

  – Tweet Data: tweet ID, Datetime, text, username.
  – News Data: newsid, Datetime, source, title, summary, content.

3 Methodology

3.1 Preprocessing

3.1.1 Social Media and News (Text Data)

To ensure consistency and prepare the data for sentiment analysis:

– Convert all text to lowercase, strip whitespace, and clean formatting.

– Remove URLs, mentions (@user), hashtags (e.g., #nvidia → nvidia), and emojis.

– Remove stopwords and punctuation.

– Apply stemming or lemmatization.

– Tokenize the sentences into words.

3.1.2 Sentiment Analysis

For both Twitter and news data:

– Use FinBERT (or another financial NLP model) to assign sentiment scores to each entry (positive, neutral, or negative).

– Assign polarity scores to news and tweet data using FinBERT and VADER.

– Calculate sentiment scores weighted by the credibility of the user (for tweets) and the news source (for news).

3.1.3 Stock Price Data

Align stock price data with daily sentiment scores:

– Calculate daily returns and moving averages for baseline comparisons.

3.2 Sentiment Index Construction

3.2.1 Input Components

– **Social Media Sentiment**: Use FinBERT and VADER for polarity detection on tweets, weighted by user credibility.

– **News Sentiment**: Use FinBERT and VADER for polarity detection on news content, weighted by source credibility and article engagement.

– **Market Dynamics**: Utilize normalized volatility metrics from `yfinance` and a smoothed VIX index.

3.2.2 Optimization Framework

Bayesian hyperparameter tuning for component weighting and equal-weight baseline for robustness checks.

3.3 Recommendation System

The recommendation system provides trading signals and suggestions based on sentiment levels:

- **Daily Signals**: Generate daily 'Nvidia Signal' (buy/hold/alert) based on sentiment levels derived from Twitter and news data.
- **Recommendation of Influential Tweets/News**: Recommend high-impact tweets or news articles based on sentiment analysis to enhance investment decisions.
- **Cross-Stock Recommendations**: Optionally, recommend other stocks with similar emotional trends, such as AMD or Intel, based on sentiment comparisons.

3.4 Analytical Framework

Module	Technique	Objective
Content Mining	LDA topic modeling Sentiment-topic correlation	Identify dominant discussion themes Map keyword-sentiment associations
Structure Mining	NetworkX/Gephi propagation graphs Community detection	Visualize sentiment diffusion paths Identify influencer clusters
Usage Mining	Granger causality tests Poisson regression	Quantify temporal relationships Model sentiment-behavior dynamics

3.5 Validation Protocol

Use walk-forward backtesting to validate the model's predictions.

4 Evaluation Metrics

4.1 Predictive Power

- Directional accuracy: percent vs actual price movements
- Sharpe ratio of sentiment-driven trading signals.

4.2 Statistical Significance

- Pearson/Spearman correlations with lagged returns.
- Diebold-Mariano test against benchmark models.

4.3 Network Effects

- Sentiment cascade velocity (nodes/hour).
- Influence hierarchy via PageRank centrality.

5 Expected Outcomes

We expect to show that:

- Public sentiment—especially when strongly positive or negative—precedes short-term price movement in NVDA stock
- Tweets and news exhibit identifiable emotional diffusion patterns, traceable over time and user clusters
- A simple rule-based sentiment threshold can produce tradeable signals
- This approach can be generalized to other individual stocks or sectors for real-time investor sentiment tracking