INTERACTIVE ANALYTICS OF MARKET SENTIMENT USING LLM-LABELED DATA

Team 176: Michael Pucci, Matthew Cheung, Yiming Feng, Xin Lin

ABSTRACT

We developed an interactive system that tracks financial sentiment across news and Reddit discussions, linking sentiment scores to specific tickers and sectors. By combining LLM-based labeling with neural network classification and cloud-native architecture, our system captures rapid shifts in market mood that conventional dashboards often miss.

MOTIVATION & OBJECTIVES

The Challenge:

•Traditional sentiment tools focus only on structured financial news

•Fast-moving online communities (Reddit's r/stocks, r/wallstreetbets) drive market narratives

•Need for unified system capturing both formal and informal sentiment

Our Goal: Create a practical, dynamic tool that:

•Detects ticker and sector mentions in financial text

•Assigns accurate sentiment scores across diverse sources

•Visualizes aggregate sentiment trends interactively

•Helps analysts identify market mood shifts early

SYSTEM ARCHITECTURE

•**Two-Phase Pipeline:**

•**Phase 1: Model Training**

•Financial news headlines labeled via Google Gemini

•High-quality training dataset generated automatically

•Feed-forward neural network trained on labeled data

•Efficient alternative to computationally expensive transformers

•**Phase 2: Deployment & Visualization**

•Real-time Reddit data ingestion via PRAW API

•Sentiment prediction using trained model

•PostgreSQL database for storage and aggregation

•Interactive Dash/Plotly dashboard for exploration

•**Cloud Infrastructure:**

•End-to-end Google Cloud Platform deployment

•Scalable, near real-time processing

•SQLAlchemy for efficient data queries

Method: Data Sources

•**Training Data**: 200,000+ financial news headlines (Webz.io)

•**Deployment Data**: Reddit posts from r/stocks, r/investing, r/wallstreetbets

	uuid	url	site	site_full	title	author	published	country	language
0	b1478dba6	https://ww	under30ce	www.under	US stock in	Ashley Niel	2024-12-16	US	english
1	2bf229308	https://fakt	fakti.bg	fakti.bg	Over 230,0	СУ Св. Клим	2024-12-17	BG	english
2	121445e9f	https://ww	businessto	www.busin	Crypto taxa	@business_	2024-12-17	IN	english
3	18e3d7a74	https://ww	plymouthh	www.plymo	Lloyds Bank	Kate Lally	2024-12-17	GB	english
4	4873a12db	https://ww	finanznach	www.finanz	Cheetah M	AFX News	2024-12-17	DE	english
5	848a15249	https://the	theeagleon	theeagleon	Nigeria's in	Hassan Mu	2024-12-17	NG	english
6	39284c4de	https://ww	investing.c	www.invest	German bu	Reuters	2024-12-17	CN	english
7	ef76c36c1	https://ww	armstronge	www.armst	Food Inflati	Martin Arm	2024-12-17	US	english
8	10c6b30e8	https://ww	gazettextra	www.gazet	How crimin	David Danz	2024-12-17	US	english
9	162d2e246	https://ww	surinenglis	www.surine	High wage	Lucía Palac	2024-12-17	ES	english
10	bf3724345	https://ww	ktbs.com	www.ktbs.c	Markets mc	AFP	2024-12-17	US	english
11	06f9db913	https://ww	foxbangor.c	www.foxba	Markets mc	AFP	2024-12-17	US	english
12	39c008c3c	https://ww	sharecast.c	www.share	London ope	Michele Ma	2024-12-17	GB	english
13	6efb4ae7a	https://ww	sharecast.c	www.share	Bunzl share	Josh White	2024-12-17	GB	english
14	61d9c531c	https://trib	tribune.net	tribune.net	CoA Report	Edjen Oliqu	2024-12-17	PH	english
15	23f6e79f23	https://ww	business-s	www.busin	Telecom st	Business St	2024-12-17T08:00:00.		english
16	c7e9ee71d	https://the	thedailygua	thedailygua	Why Olaf Sc	@TheDaily(	2024-12-17	IN	english
17	4d020713c	https://ww	rappler.con	www.rapple	FACT CHEC	Ailla Dela C	2024-12-17	PH	english
18	8b1c1e900	https://mw	mwnation.c	mwnation.c	PER capita	Grace Phiri	2024-12-17	AF	english
19	2a9a9cfedl	https://ww	fxstreet.cor	www.fxstre	EUR/USD s	FXStreet	2024-12-17	US	english

Figure 1. Data snapshot

Method - Key Technical Innovations:

•**LLM-Assisted Labeling**
- Automated sentiment labeling via Google Gemini
- Consistent, high-quality labels without manual annotation
- Structured prompt template for classification

•**Finance-Aware Text Processing**
- Preserves ticker symbols ($TSLA, AAPL formats)
- Retains financial indicators (%, $, emojis)
- Pattern matching for ticker extraction

•**Neural Network Classifier**
- Learns financial-specific language patterns
- Fast inference for real-time streams
- Balances accuracy with computational efficiency

•**Multi-Level Aggregation**
- Daily sentiment averages
- 90-day rolling statistics
- Sector-level summaries by industry

INTERACTIVE DASHBOARD FEATURES

User Controls:

•Toggle between Reddit-only vs. merged news+Reddit data

•Select ticker-level or sector-level analysis

•Customize date ranges and sentiment types

Visualization Components:

•**1. Sentiment Time-Series Panel**

•Daily sentiment scores (positive/neutral/negative)

•90-day rolling statistics for trend detection

•Optional stock price overlay for correlation analysis

•Multi-ticker comparison capability

•**2. Industry Heatmap Panel**

•Sector-wide sentiment aggregation

•Diverging color scale for intensity visualization

•Temporal patterns across industries

•Quick identification of market-driving sectors

RESULTS & EVALUATION

Classifier Performance:

•Stable training/validation loss convergence (80/20 split)

•Successful generalization to informal Reddit text

•Strong alignment with human judgment on sampled predictions

•Challenges: Sarcasm and highly context-dependent posts

Pipeline Validation:

•Consistent ticker extraction across format variations

•Accurate sentiment predictions written to database

•Verified aggregations (daily averages, rolling stats, sector summaries)

Dashboard Accuracy:

•Visualization outputs match direct SQL query results

•Clear display of daily sentiment shifts

•Effective highlighting of long-term patterns via rolling statistics

•Successful identification of sector-level market movements

RESULTS & EVALUATION

Overall Accuracy: 84% with balanced performance across classes (Fig. 2)

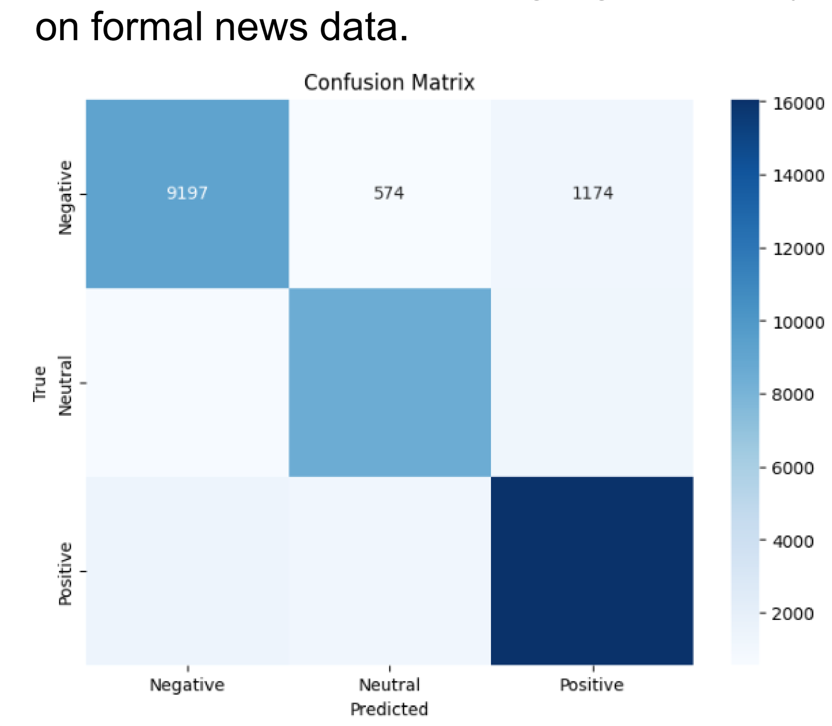
•**Positive**: Precision 0.87, Recall 0.86, F1 0.87

•**Neutral**: Precision 0.83, Recall 0.82, F1 0.83

•**Negative**: Precision 0.82, Recall 0.84, F1 0.83

•**Macro-averaged F1-score: 0.85** across 39,901 test samples

Most misclassifications occur between adjacent sentiment categories, reflecting the nuanced nature of financial language. The model successfully generalizes to informal Reddit text while maintaining high accuracy on formal news data.



Classification Report:

	precision	recall	f1-score	support
Negative	0.82	0.84	0.83	10945
Neutral	0.83	0.82	0.83	10363
Positive	0.87	0.86	0.87	18593
accuracy			0.85	39901
macro avg	0.84	0.84	0.84	39901
weighted avg	0.85	0.85	0.85	39901

Figure 2. Confusion Matrix of Classification Performance

DASHBOARD VISUALIZATION

•**Time-Series Panel**: Track sentiment evolution for multiple tickers with daily or 90-day rolling averages (scale: -100 to +100). Example shows divergent trends across five selected stocks (Figure 3).

•**Industry Heatmap**: Compare daily sentiment across 11 sectors using color intensity (blue = positive, red = negative, white = neutral). Quickly identify market-driving industries and temporal patterns. (Figure 4)

•**User Controls**: Toggle data sources, select ticker/sector granularity, filter date ranges, and choose multiple tickers or industries for comparison.

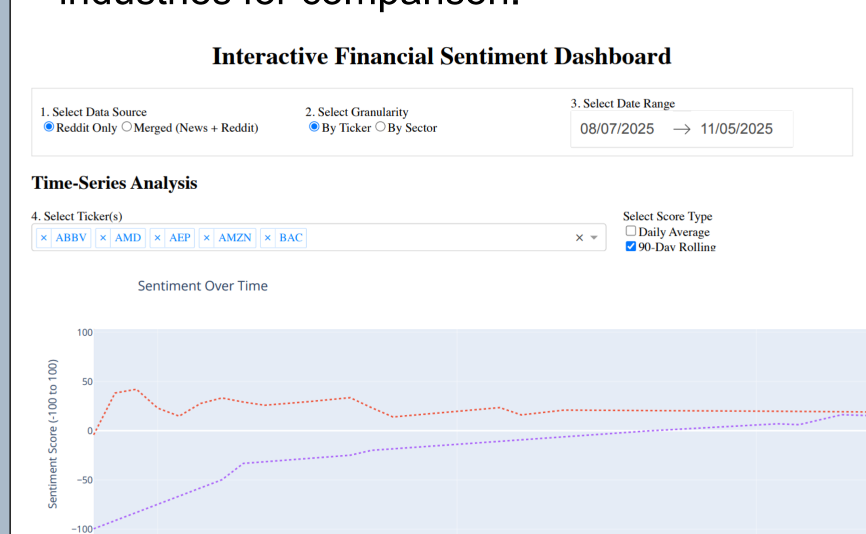


Figure 3. Time Series Sentiment visualization

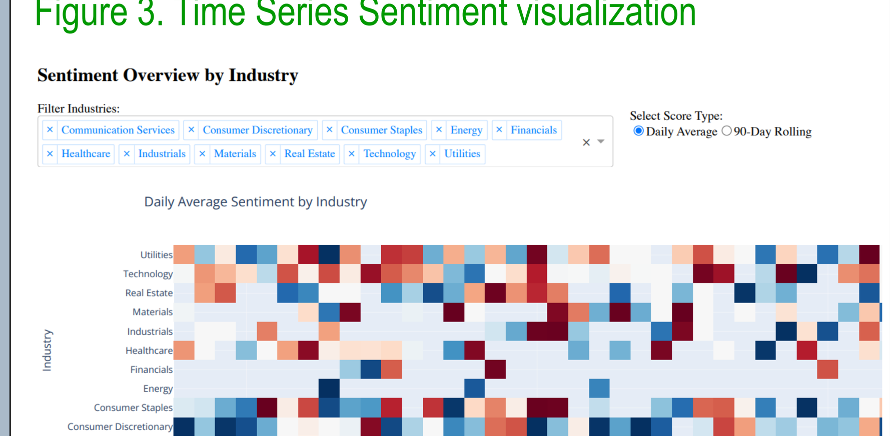


Figure 4. Sentiment Overview by Industry heat map

KEY FINDINGS

✓ **LLM labeling + lightweight neural network = efficient, accurate sentiment classification**

✓ **System generalizes from formal news to informal social media language**

✓ **Multi-level visualization reveals sentiment patterns invisible in static reports**

✓ **Cloud-native architecture enables scalable, near real-time processing**

✓ **Combined news+Reddit sentiment provides richer market narrative**

CONCLUSIONS & FUTURE WORK

Achievements:

•Demonstrated effective hybrid approach combining LLM labeling with neural classification

•Created interpretable system for tracking ticker and sector sentiment

•Validated alignment between sentiment shifts and real-world market narratives

Future Enhancements:

•**Live streaming**: Implement continuous real-time Reddit ingestion

•**Model upgrades**: Integrate transformer-based models for nuance detection

•**Expanded coverage**: Include additional subreddits and data sources

•**Predictive analytics**: Explore sentiment-price relationships with live market feeds

REFERENCES

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063. https://arxiv.org/abs/1908.10063

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1–8.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT.

Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). Deep learning for event-driven stock prediction. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Havre, S., Hetzler, E., & Nowell, L. (2002). ThemeRiver: Visualizing thematic changes in large document collections. IEEE Transactions on Visualization and Computer Graphics, 8(1), 9–20.

Heer, J., & Shneiderman, B. (2012). Interactive dynamics for visual analysis. Communications of the ACM, 55(4), 45–54.

Kirtac, K., & Germano, G. (2025). Large language models in finance: What is financial sentiment? doi: 10.2139/ssrn.5166656.

Liu, T., et al. (2020). Event extraction as machine reading comprehension. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Ong, K., Van Der Heever, W., Satapathy, R., Cambria, E., & Mengaldo, G. (2023). FinXABSA: Explainable finance through aspect-based sentiment analysis. IEEE International Conference on Data Mining Workshops (ICDMW), 773–782. doi: 10.1109/ICDMW60847.2023.00105.

Park, J., Lee, H. J., & Cho, S. (2021). Automatic construction of context-aware sentiment lexicon in the financial domain using direction-dependent words. arXiv preprint arXiv:2106.05723. http://arxiv.org/abs/2106.05723

Shiller, R. J. (2017). Narrative economics. American Economic Review: Papers & Proceedings, 107(4), 967–1004.

Sun, Y., Yuan, H., & Xu, F. (2025). Financial sentiment analysis for pre-trained language models incorporating dictionary knowledge and neutral features. Natural Language Processing Journal, 11, 100148. doi: 10.1016/j.nlp.2025.100148.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. The Review of Financial Studies, 21(2), 1–37.