

AI-Driven Sentiment: Smarter Insights for Smarter Investments

Stock Market News Analysis

November 2, 2025

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- Data Background and Contents
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Actionable Insights and Next Steps
- Appendix

Executive Summary

- Project evaluates how financial news sentiment influences stock performance.
- Raw news articles converted into numeric embeddings for machine learning.
- Tested two NLP approaches: Word2Vec vs. Sentence Transformer embeddings.
- Random Forest models trained and evaluated using standard classification metrics.
- Sentence Transformer achieved significantly higher performance, demonstrating stronger ability to capture financial context.

Business Problem Overview and Solution Approach

Business Context

- Markets move fast, driven by performance, innovation, and buzz.
- News and media can swing investor sentiment in minutes.
- Analysts face a flood of information that's tough to track and interpret.
- The opportunity is to turn chaotic data into clarity with AI.

Problem Definition

- Build an AI powered sentiment engine to decode market mood from news.
- Use historical NASDAQ data including news, prices, and volumes to train insights.
- Deliver weekly sentiment summaries that sharpen stock predictions and guide smarter investments.

Key Insights and Recommendations

Key Insights

- News sentiment is a meaningful indicator of stock movement.
- Sentence Transformer embeddings yield better prediction precision than Word2Vec.
- Traditional ML models can leverage advanced language embeddings effectively.
- Broader datasets likely to further improve precision and robustness.

Recommendations

- Standardize on Sentence Transformer embeddings for sentiment modeling.
- Expand dataset across additional companies, sectors, and longer time periods.
- Evaluate other ML models (e.g., XGBoost, fine-tuned transformer models).
- Implement rolling time-series validation to mimic real-world trading conditions.
- Incorporate results into analyst dashboards to support data-driven investment decisions.

Data Background and Contents

- Dataset includes daily financial news articles for a NASDAQ-listed company.
- Each record contains the corresponding daily stock market information.
- Data spans 73 trading days, providing aligned news and market movement context.
- Text data (news articles) is included in both raw and cleaned formats.
- Sentiment labels indicate market reaction categories (three-class sentiment setup).
- Price-related fields (Open, High, Low, Close) and trading Volume included for market trend analysis.
- Final dataset contains 418 total news articles matched to market data.

Data Overview

- **No missing values** across any column.
- Dataset includes **9 columns**: Date, News, Open, High, Low, Close, Volume, Label, and Cleaned_News.

Uniqueness Summary

- **Date**: 73 unique values
- **News**: 418 unique entries
- **Open**: 69 unique values
- **High**: 67 unique values
- **Low**: 66 unique values
- **Close**: 68 unique values
- **Volume**: 73 unique values
- **Label**: 3 unique sentiment classes
- **Cleaned_News**: 417 unique processed news entries

Insights

- Data is **complete and clean** with no nulls.
- **News and Cleaned_News** columns have the highest diversity, ideal for text analysis.
- **Label** column indicates a **multi-class sentiment problem** (3 categories).
- **Price and volume features** show strong variability suitable for correlation analysis with sentiment.

Data Preprocessing

- Word2Vec embeddings were generated for the news articles, creating a (418, 100) array where each article is represented by a 100-dimensional vector.
- Sentence Transformer embeddings were generated using the 'all-MiniLM-L6-v2' model, creating a (418, 384) array where each article is represented by a 384-dimensional vector.
- The dataset was split into training (75%) and testing (25%) sets using stratified sampling to maintain label balance.

Model Evaluation Criteria

- **Accuracy:** Measures the overall percentage of correctly predicted labels out of all predictions.
- **Precision:** Evaluates how many of the positive predictions made by the model were actually correct (focuses on quality of positive predictions).
- **Recall:** Measures the model's ability to identify all relevant positive instances (focuses on completeness).
- **F1-Score:** Harmonic mean of precision and recall, providing a balanced metric when both false positives and false negatives are important.
- **Confusion Matrix:** Summarizes true positives, false positives, true negatives, and false negatives to visualize classification performance.

Sentiment Analysis - Model Building & Evaluation

Model	Embedding Type	Train Precision*	Test Precision	Reasoning
Random Forest	Word2Vec (100-dim)	~0.80*	~0.64	Lower precision → more false-positive trade signals; less reliable for execution decisions
Random Forest	Sentence Transformer (384-dim)	~0.88*	~0.72	Higher precision → fewer false positives; better confidence in trade signals

Sentiment Analysis - Model Building & Evaluation

Model Overview & Observations

Word2Vec + Random Forest

- Converts text into 100-dim word vectors averaged per article.
- Efficient and simple to implement.
- *Observation:* Performed reasonably well, but lacked deeper context and nuance in financial language.

Sentence Transformer + Random Forest

- Generates 384-dim contextual embeddings at sentence level.
- Captures meaning, tone, and semantic relationships more effectively.
- *Observation:* Outperformed Word2Vec across all metrics, showing strong ability to interpret financial news sentiment.

Key Takeaway

- Sentence Transformer provided a more meaningful representation of news text, leading to higher precision.
- *Comment:* Modern transformer-based embeddings are better suited for sentiment-driven market prediction tasks than traditional word embedding approaches.

Sentiment Analysis - Model Building & Evaluation

Model Analysis (Precision Focus)

Word2Vec + Random Forest

- Moderate performance
- Lower precision → more false-positive sentiment signals
- Higher risk of unnecessary or incorrect trades

Sentence Transformer + Random Forest

- Stronger performance across metrics
- **Higher precision** → more reliable signal quality
- Better fit for decision-making in financial markets

Conclusion

- **Sentence Transformer model chosen**
- Delivers higher-confidence sentiment predictions and reduces costly false signals

Actionable Insights and Next Steps

Insights

- Precision is critical — **false signals = bad trades + costs**
- Sentence Transformers capture financial language and tone more accurately
- Traditional embeddings (Word2Vec) struggle with context → weaker signal reliability

Next Steps

- Adopt Sentence Transformer embeddings going forward
- Optimize model for **precision** (threshold tuning, PR curve evaluation)
- Expand dataset (longer timeline, more tickers, broader news sources)
- Test advanced models (XGBoost, fine-tuned transformers, ensembles)
- Validate using rolling time windows and live backtesting

Data & Features

- 418 news articles aligned to daily market data (73 trading days)
- Fields: News text, cleaned text, OHLC prices, volume, sentiment labels
- Three-class sentiment classification (positive / neutral / negative)

Embedding Methods

- **Word2Vec:** 100-dim averaged word vectors
- **Sentence Transformer (MiniLM-L6-v2):** 384-dim contextual sentence embeddings

APPENDIX

Modeling

- Classifier: Random Forest
- Train/Test Split: 75% / 25% stratified
- Evaluation focus: **Precision** to avoid costly false-positive signals in trading

Tools

- Python, Pandas, Scikit-Learn, Gensim, HuggingFace Transformers

Assumptions & Limitations

- Single-company dataset; short time window
- No intraday pricing or real-time news feed yet
- Future work: expand universe, apply time-series validation, backtest results

APPENDIX



Happy Learning !

