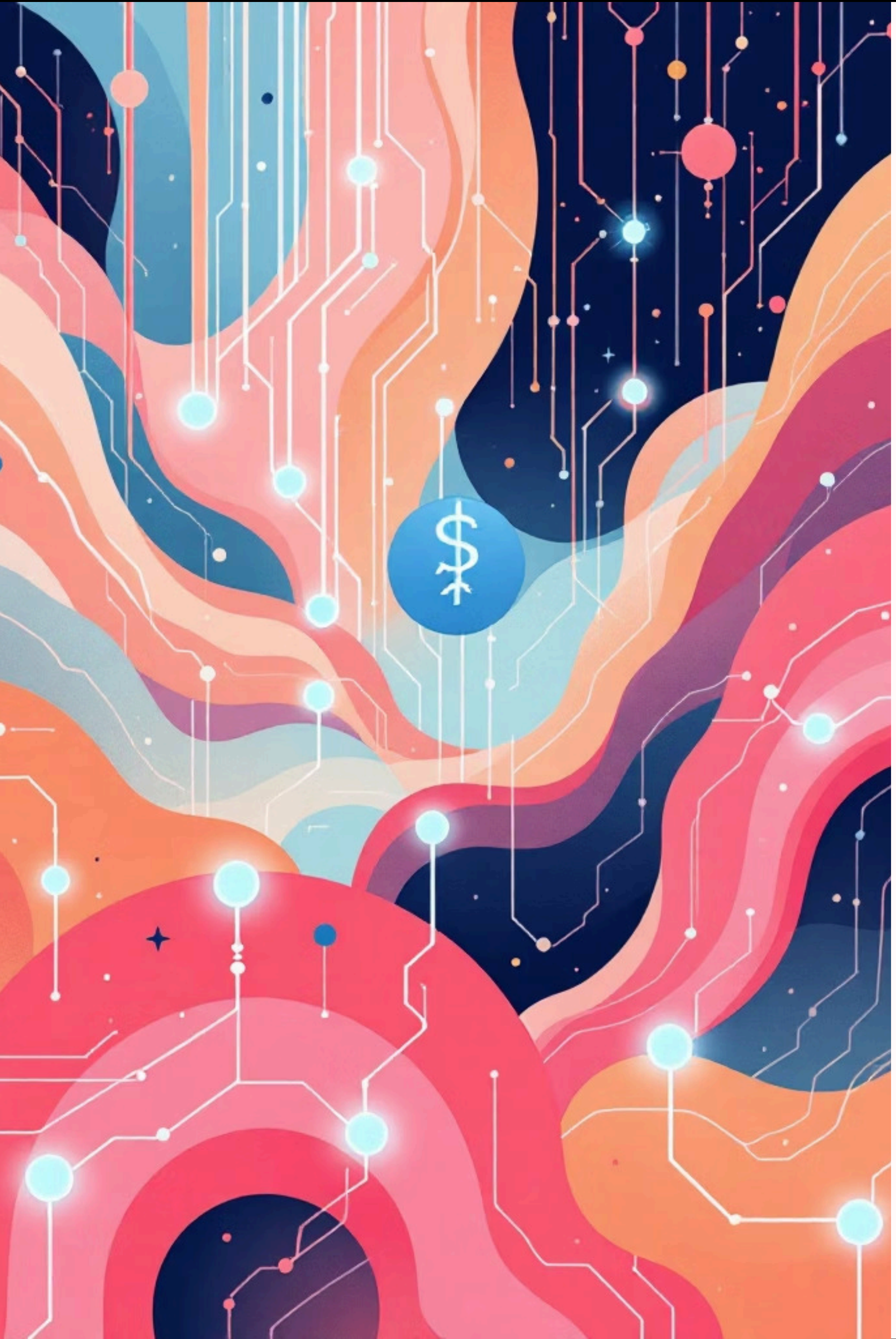


# Generative AI Foundations

## Stock Market News Sentiment Analysis

February 28, 2026

# Contents

EXECUTIVE SUMMARY

# Sentence Transformer: The Recommended Architecture

## The Problem

Manual NASDAQ monitoring is slow due to the growing volume of stock related news and inherently biased. Our automated sentiment analysis delivers objective, real-time signals — optimizing investment strategies and client outcomes.

## Top Model Results

| 77.8% | +9% |
|---|---|
| Accuracy | F1 vs. Word2Vec |

## Next Steps & Recommendations

### Volatility Alert System

Flag negative headlines in real-time to enable proactive hedging before price drops

### Sentiment-Weighted Sizing

Integrate sentiment probability scores into trading algorithms to auto-adjust positions

### F1 as Primary KPI

Prioritize F1-score over accuracy to avoid the Majority Class Trap in bullish cycles

EXECUTIVE SUMMARY

# From Classification Tool to Decision-Support System

Implementing these recommendations elevates the model into a dynamic, real-time engine for capital allocation.

### Proactive Risk Mitigation

Flag negative headlines before sentiment shifts are fully priced in — execute hedges and exit strategies ahead of the market

### Navigate Liquidity Events

High-intensity sentiment is a statistical leading indicator for volume spikes, providing a crucial early-warning advantage

### Eliminate Cognitive Bias

Automate position adjustments based on sentiment probability scores — removing subjectivity from capital allocation entirely

### Optimized Capital Efficiency

Scale aggressively during high-conviction positive windows; auto-reduce exposure during Neutral or Negative risk periods


BUSINESS PROBLEM

# Manual Monitoring Can't Keep Pace With the Market

The daily influx of financial news materially moves stock prices — yet manual analysis remains slow, subjective, and inconsistent at scale.

### Information Overload

Analysts cannot systematically process the volume, velocity, and diversity of incoming financial data

### Subjectivity & Inconsistency

Manual sentiment interpretation is inherently biased and unreliable — two analysts reading the same headline may reach opposite conclusions

### Missed Market Opportunities

Failure to quantify qualitative data leads to delayed reactions, missed returns, and elevated risk exposure

**The scale of the problem demands an immediate, automated, and unbiased solution.**

SOLUTION APPROACH

# An AI-Driven Sentiment Engine for Real-Time Market Intelligence


## The Solution

A robust, transformer-based sentiment analysis system that automatically processes financial news to generate objective, real-time market signals.

## Expected Results

- Quantifiable reduction in risk exposure
- Optimized investment strategy decisions
- Improved client outcomes at scale

→ Analyze breaking news **as it is reported** — no lag

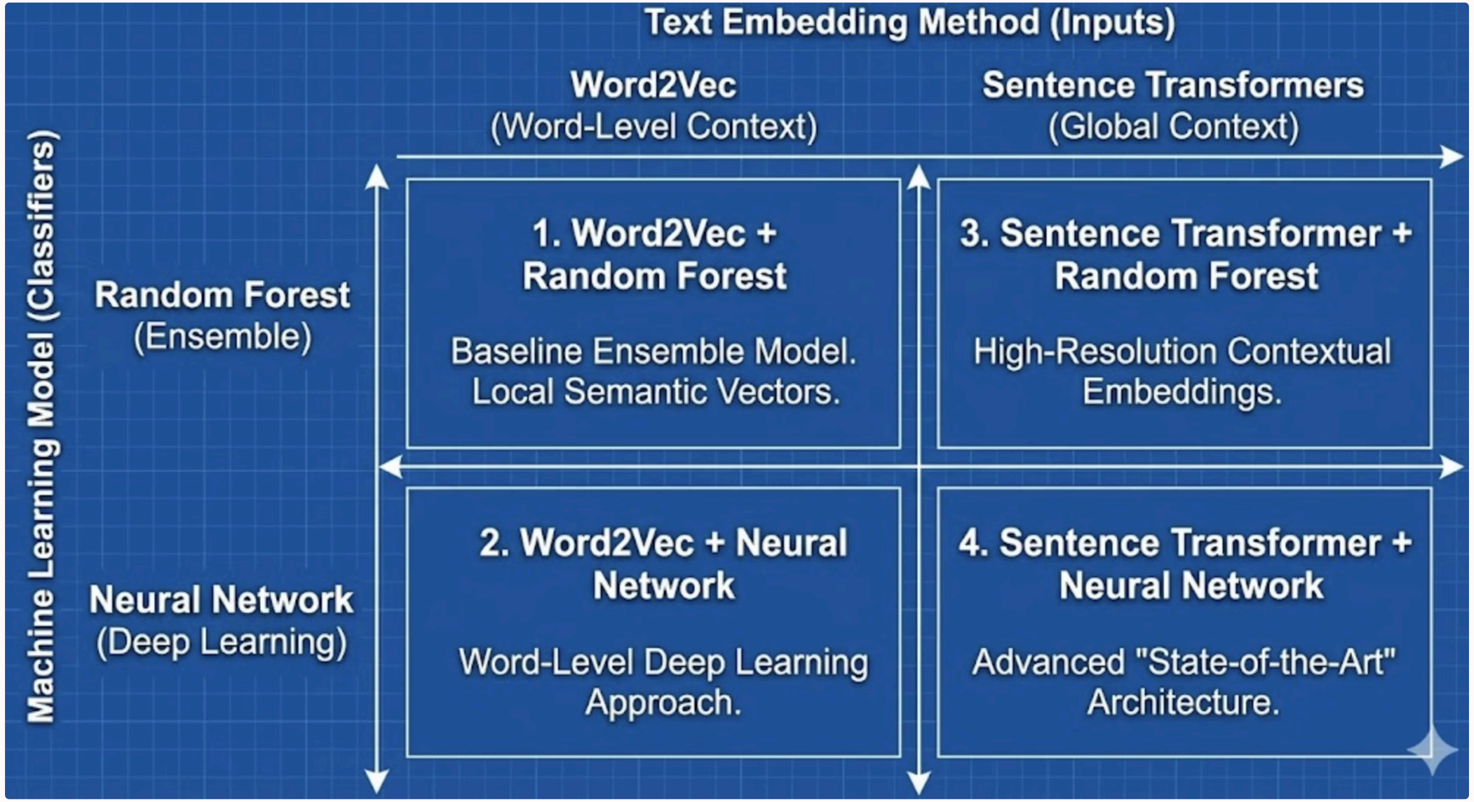
→ System-generated signals driven by **sentiment scores**

→ Reduce time-to-act from **hours to real-time**

SOLUTION METHODOLOGY

# Four-Model Comparative Framework

To identify the most robust solution, we benchmarked traditional word-level embeddings against modern transformer-based contextual architectures.



SOLUTION METHODOLOGY

# Why a Four-Model Framework?

This methodology guaranteed the final system achieved optimal **accuracy** and **stability** — not just on average, but against the most challenging edge cases in financial data.

### 1

#### Feature Representation vs. Architecture

Isolate whether performance gains come from *how* words are represented or *how* those representations are processed

### 2

#### Establishing Performance Baselines

Word2Vec + Random Forest sets the "floor." Every subsequent model must demonstrably prove its value-add over this foundation

### 3

#### Addressing the "Small Data" Challenge

With only 418 records, we needed the approach most resistant to overfitting while remaining sensitive to market-moving news events

### 4

#### Class Imbalance Verification

Confirmed the model identifies genuine minority "Negative" signals — not merely tracking the 64.6% positive market trend


# EDA Results – Univariate Analysis

Understanding the balance of the data and key observations on individual variables — the foundation for every modeling decision that follows.

EXPLORATORY DATA ANALYSIS

EDA RESULTS

# Dataset Overview & Feature Engineering

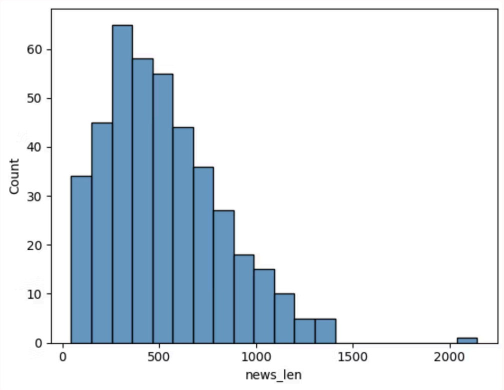
**Duration:** ~4 months (Jan 2 – Apr 29, 2019) | **Volume:** 418 records | **Source:** NASDAQ-listed company daily news, prices, and trading volumes

| Variable | Role in Analysis | Description |
|---|---|---|
| **News** (Headlines) | Primary Predictor (*X*) | Unstructured text input used to generate semantic embeddings for all four models |
| **Label** (Sentiment) | Target Variable (*y*) | Categorical classification: −1 = Negative, 0 = Neutral, 1 = Positive |
| **Close Price & Volume** | Key Contextual Metrics | Used in EDA to validate that sentiment labels correlate with actual market value and liquidity |
| **Date, Open, High, Low** | Supporting Financial Data | Provides a complete OHLC view for granular intraday market context |

EDA RESULTS

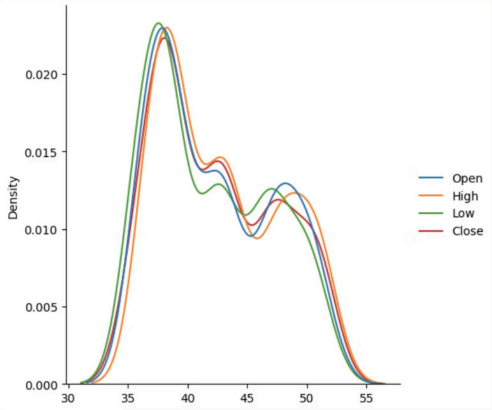
# Univariate Analysis: Key Observations

Exploratory analysis of individual variables confirms both the technical necessity of transformer models and the business validity of the sentiment approach.




## News Headline Length

Headlines range from **44 to 500+ words**. This variability makes context-preservation critical — Word2Vec loses semantic meaning in longer articles, confirming the need for Sentence Transformers.



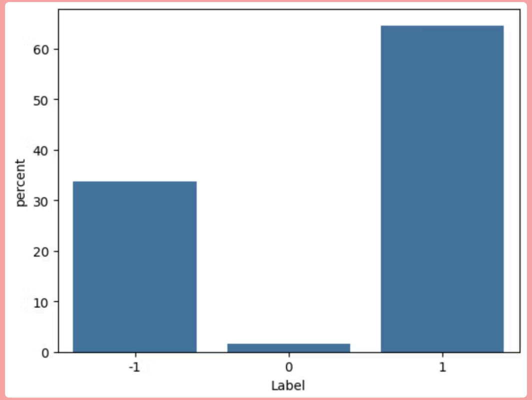

## OHLC Price Spread

Daily High–Low spread shows a **consistent 2–4% intraday fluctuation**. This confirms sufficient liquidity and volatility for sentiment signals to carry predictive value — validating the entire business case.

# Univariate Analysis: Sentiment & Volume

Two variables reveal the core statistical properties of the dataset — and the risks they introduce for naïve modeling approaches.

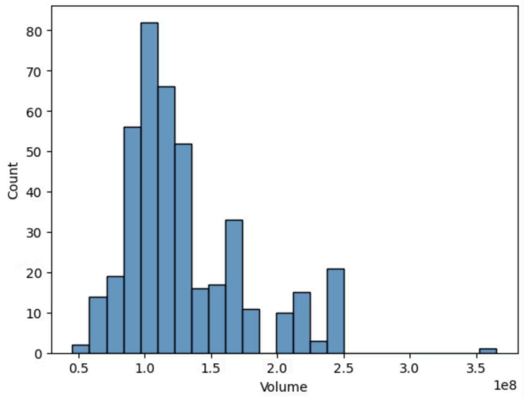
## Variable: Sentiment




The dataset is **64.6% positive**. Positive (1) outnumbers Negative (-1) and Neutral (0) by nearly 2:1 — a direct reflection of a sustained bull market in the observation window.

**F1-Score implication:** A model could achieve high accuracy simply by predicting 'Positive' for every headline. We addressed this by prioritizing **F1-score over accuracy** throughout evaluation.

## Variable: Volume



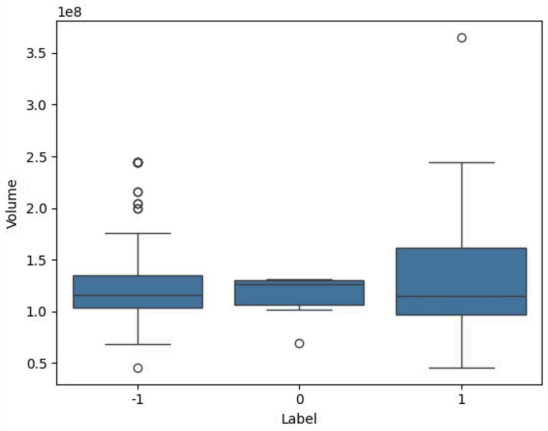

Volume is **not normally distributed** — it exhibits a long right tail. Most trading days cluster around standard volume levels, but extreme outliers reach **3× the average**.

**Insight:** These outliers are not noise. They represent identifiable **market shock events** where news-driven investor behavior produced dramatic liquidity surges.

# EDA Results – Bivariate Analysis

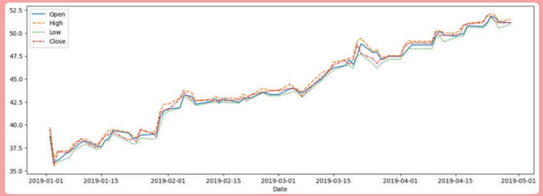
Examining the relationship between sentiment labels, trading volume, and OHLC price data to understand how news intensity maps onto market behavior.

## Variables: Label & Volume




Negative (-1) and Positive (1) sentiment days are consistently associated with **higher trading volumes** compared to Neutral days. High-sentiment headlines function as a **volume catalyst** — with negative news correlating with the sharpest volume spikes, indicating high-velocity trading and elevated liquidity risk.

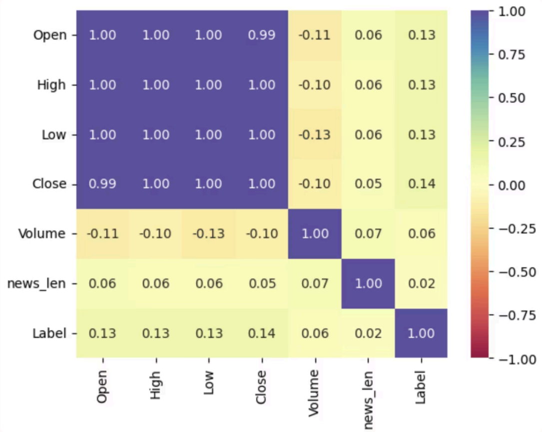
## Variables: Date, Open, High, Low, Close




The OHLC time series shows a **sustained upward climb** from ~$36 in January to over **$50 in April** — a confirmed bull run. This directly explains the high density of positive headlines (64%) and reinforces that our dataset is regime-dependent.

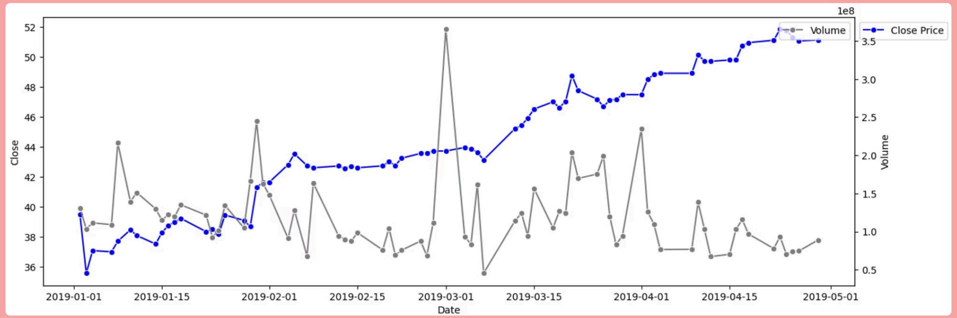
# Bivariate Analysis: Price Correlation & Volume Spikes

## Variables: Open, High, Low, Close




The correlation heatmap reveals **near-perfect collinearity** between Open, High, Low, and Close prices. This confirms significant **feature redundancy** among price metrics, justifying the decision to use **unstructured news text** as the primary predictive signal rather than price alone.

## Variables: Date, Volume, Close Price

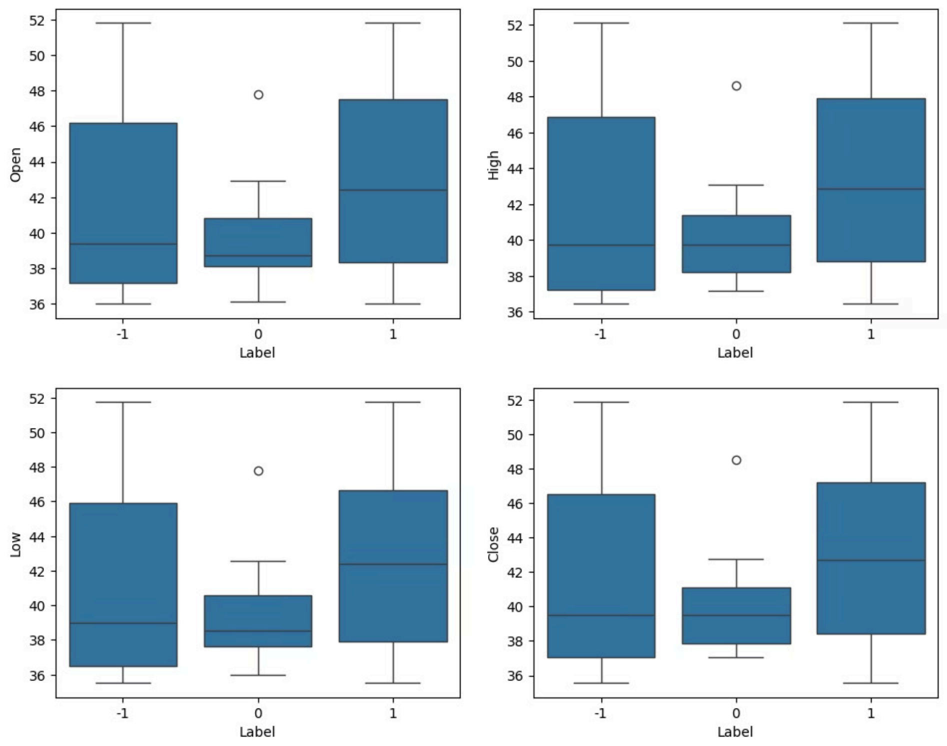



While the price trend is steady and gradual, trading volume is characterized by **massive, isolated spikes** that occur largely independent of the long-term price trajectory. This strongly suggests that **external news events** — not price momentum — are triggering mass liquidity shifts and investor repositioning.

# Bivariate Analysis: Sentiment vs. Price Spread

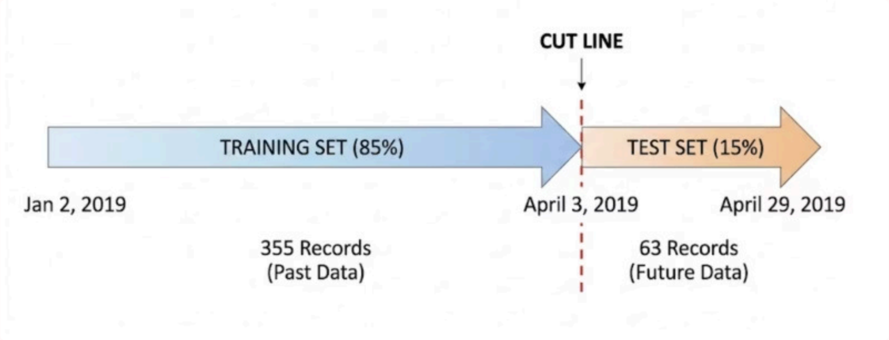
## Variables: Open, High, Low, Close, Label

Both Negative (-1) and Positive (1) sentiment days exhibit significantly **wider interquartile ranges (IQR)** and broader overall price spreads compared to Neutral (0) days. Neutral days show the most **compressed price ranges** and lower median prices — a clear signal that the **absence of impactful news directly correlates with market stability**. This finding has direct implications for volatility-sensitive strategies and options pricing models.



# Data Preprocessing

Preparing the dataset for embedding generation and model training — with special care to preserve the temporal integrity of financial time-series data.




We applied a **fixed chronological 85/15% split** to preserve the time-series structure of the data — training on the first 355 records (Jan 2 – Apr 3, 2019) and testing on the final 63 records (post-Apr 3, 2019).

### Why 85% Training?

With only 418 records, maximizing training volume was critical. It gave the Neural Network sufficient **linguistic variety** to learn complex financial semantics without sacrificing test-set significance.

### Why No Random Shuffle?

Random shuffling enables **data leakage** — the model would see future events to predict the past. The chronological split ensures the model is tested on a truly held-out future window, faithfully mimicking real-time deployment.

# Word Embeddings

Converting unstructured news headlines into **numerical vectors** that capture financial sentiment and semantic meaning — a foundational step before any classifier can operate.

To make news headlines machine-readable, we tested two distinct embedding strategies — each with a fundamentally different philosophy for how language should be represented.

## Word2Vec (Baseline)

A foundational NLP technique that encodes individual words into vectors. To represent a full sentence, we compute a **Mean Pool** (averaging all word vectors). In financial text, this is a critical limitation — word order and syntax encode the "who" and "what" of a headline, and simple averaging **destroys those relationships**.

## Sentence Transformers

Transformers treat the **entire headline as a single semantic unit**, using self-attention to weigh each word relative to the others. Context is fully preserved. This architectural advantage is what delivered the **9% boost in F1-score** — allowing the model to correctly distinguish nuanced market-moving signals.

# Sentiment Analysis – Model Matrix

We tested two distinct embedding technologies against two classifier types, yielding four experiments with increasing representational power.

| Embedding Strategy | Classifier: Random Forest (ML) | Classifier: Neural Network (DL) | Design Rationale |
|---|---|---|---|
| **Word2Vec (Baseline)** | **Experiment 1:** Tests traditional ensemble logic on Bag-of-Words averages | **Experiment 2:** Tests how a Multi-Layer Perceptron (MLP) handles averaged word vectors | Establishes the performance floor with classical NLP representations |
| **Sentence Transformers** | **Experiment 3:** Tests if advanced contextual embeddings improve a traditional classifier | **Experiment 4 — Winning Model.** Combines contextual deep learning with a Neural Network for maximum precision | Determines if representational richness unlocks gains in minority-class detection |

The 2×2 matrix design isolates the contribution of each component — embedding quality vs. classifier architecture — enabling a clean attribution of performance gains.

# Sentiment Analysis – Evaluation Framework

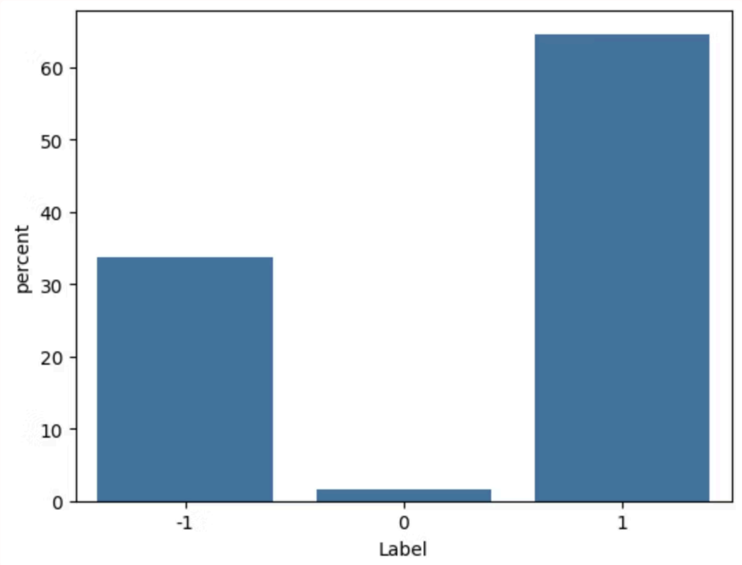
To ensure the model delivers **actionable intelligence** — not just trend-following — we designed an evaluation framework that balances general reliability with specialized risk detection.

| Metric | Significance | Target Outcome | Priority |
|---|---|---|---|
| **Accuracy** | System Reliability — overall correctness across all classes | High overall correctness (>70%) | Secondary |
| **Precision** | Operational Confidence — reliability in flagging Negative signals | High reliability, minimizing false alarms | High |
| **Recall** | Risk Exposure Coverage — comprehensive capture of market-moving headlines | Comprehensive detection of critical events | High |
| **F1-Score** | Performance Equilibrium — harmonic mean of precision and recall | Balanced detection across all sentiment classes | **Primary** |

# Evaluation Strategy: The Case for F1-Score

## Why Not Accuracy?

Accuracy is susceptible to the **"Majority Class Trap."** With 64.6% of our dataset labeled Positive, a model that blindly predicts 'Positive' for every headline would achieve over 64% accuracy — without learning anything meaningful.




## Why F1-Score?

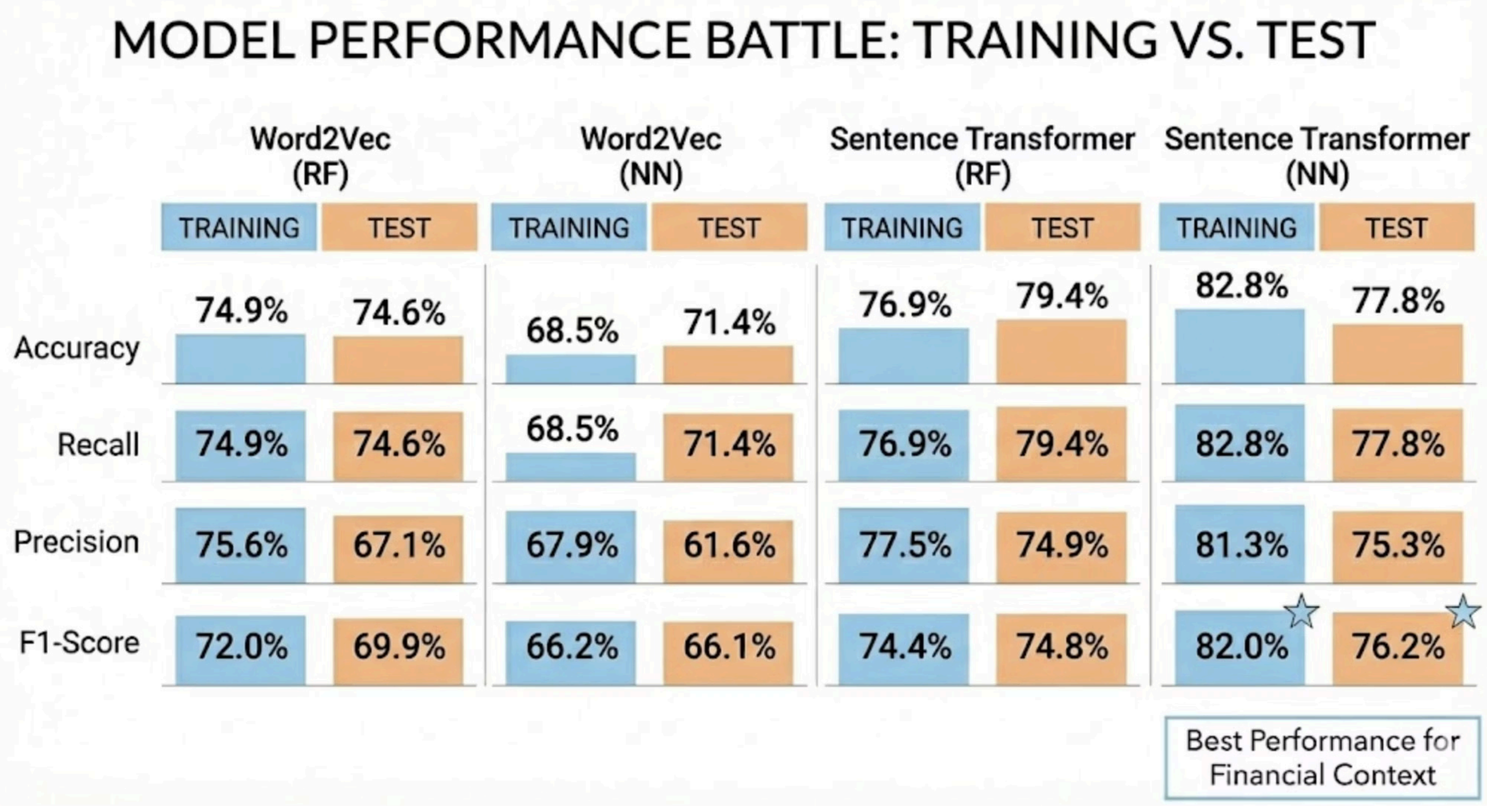
F1-Score is the **harmonic mean of Precision and Recall**, forcing the model to perform well on both dimensions simultaneously. For a risk management application, missing a Negative signal is far more costly than a false positive — F1-Score ensures we never sacrifice minority-class detection for headline accuracy numbers.

By prioritizing F1-Score, we ensure the model is genuinely effective at detecting the rare but critical **Negative and Neutral signals** that drive hedging and risk management decisions.

# Test Results: Winning Model

Highest F1-Score: 76.2%

The **Sentence Transformer + Neural Network** achieved the highest F1-Score at **76.2%** — selected as the final model because it most effectively balances precision and recall for the critical Negative and Neutral classes that raw accuracy consistently misrepresents.



# Sentiment Analysis – Full Results Comparison

All four experiments, evaluated across Accuracy, F1-Score, Precision, and Recall — with key observations on each model's strengths and failure modes.

| Embedding | Classifier | Accuracy | F1-Score | Precision | Recall | Key Observation |
|---|---|---|---|---|---|---|
| Word2Vec | Random Forest | 74.6% | 69.9% | 67.1% | 74.6% | Strong recall but lower precision suggests difficulty distinguishing true negative signals from general market noise |
| Word2Vec | Neural Network | 71.4% | 66.1% | 61.6% | 71.4% | NN underperformed the RF with Word2Vec — simple averaging provides insufficient complexity for deeper architectures |
| Sentence Transformer | Random Forest | 79.4% | 74.8% | 74.9% | 79.4% | Major performance jump across all metrics — contextual embeddings significantly improve traditional classifiers on financial syntax |
| **Sentence Transformer** | **Neural Network** | **77.8%** | **76.2%** | **75.3%** | **77.8%** | **Highest F1-Score and Precision — the superior choice for identifying critical risk outliers despite marginally lower accuracy than RF** |

# Deep Learning Architecture

**Model Chosen: Sentence Transformer + Neural Network.** Each design decision was driven by the constraints of a small, financially specialized dataset.

### Leaky ReLU Activation

Unlike standard ReLU, Leaky ReLU ensures neurons don't "die" when processing negative sentiment vectors — maintaining gradient flow for critical bearish market signals.

### Dropout (0.2)

Prevents the model from memorizing specific training headlines (overfitting), forcing it to generalize linguistic patterns across the broader financial vocabulary.

### Optimizer & Loss

Adam optimizer paired with Categorical Cross-Entropy for precise multi-class alignment across all three sentiment classes.

### Architectural Depth

**32-neuron input layer → 16-neuron hidden layer → 3-neuron output layer** (12,899 total parameters).

The decision to limit the hidden layer to **32 neurons** was the most consequential architectural choice. It creates a compression bottleneck that forces the high-dimensional Sentence Transformer vectors to encode **universal financial concepts** — rather than memorizing the exact vocabulary of the training data. This single constraint is what drives generalization on an out-of-sample month.

# Model Hyperparameters – Random Forest

Systematic tuning focused on structural parameters to prevent overfitting on our compact 418-record dataset. Maintaining shallow architectures delivered the most stable generalization — with a narrow **~5% performance delta** between training and test results.

| Model Type | Variable | Optimal Setting | Observations & Impact of Alteration |
|---|---|---|---|
| **Random Forest** | n_estimators | **11** | Increasing beyond 11 showed diminishing returns across F1-score, accuracy, precision, and recall — additional trees added complexity without improving signal discrimination |
| **Random Forest** | max_depth | **3** | Increasing depth (e.g., to 10) caused overfitting where the model captured both noise and signal indiscriminately — shallow trees enforced the necessary regularization for a small dataset |

Both hyperparameters converge on the same principle: **constrain complexity** to prevent memorization. With 418 records, the model's generalization budget is limited — shallow and sparse wins.

# Model Hyperparameters — Neural Network

| Variable | Optimal Setting | Alternatives Tested | Observations & Impact of Alteration |
|---|---|---|---|
| Hidden Layers | **1 (32 neurons)** | 2+ layers, 64+ neurons | Adding a second hidden layer or scaling to 64+ neurons caused immediate **overfitting** on our 418-record dataset — the bottleneck architecture was essential for generalization |
| Activation | **Leaky ReLU** | ReLU, Sigmoid, Tanh | Leaky ReLU consistently outperformed all alternatives by preserving gradient flow for **negative sentiment vectors** — standard ReLU caused neuron death on bearish signals |
| Optimizer | **Adam** | SGD | Switching from SGD to Adam provided a **2–3% boost** in all evaluation metrics through adaptive learning rates per parameter |
| Batch Size | **16** | 64+ | Larger batches (64+) degraded performance — small batches introduced beneficial noise that improved generalization given the limited data volume |
| Epochs | **15** | 30+ | Higher epoch counts led to memorization of training headlines rather than learning generalizable patterns — early stopping at 15 preserved out-of-sample performance |

REFERENCE MATERIAL

# Appendix

Supporting data, background context, and technical reference tables for the models and datasets described in this analysis.

# Data Background and Contents

## Dataset Overview

**Duration:** ~4 Months (January 2, 2019 – April 29, 2019)

**Volume:** 418 records

**Content:** Curated daily NASDAQ news headlines paired with corresponding market activity metrics and a human-labeled sentiment score per trading day.

## Variable Reference

| Variable | Function in Analysis |
| --- | --- |
| **News (Headlines)** | Primary Predictor (*X*): Unstructured text input used to generate semantic embeddings |
| **Label (Sentiment)** | Target Variable (*y*): Categorical classification — -1 = Negative, 0 = Neutral, 1 = Positive |
| **Close Price & Volume** | Key Contextual Metrics: Used in EDA to validate that sentiment labels correlate with actual market behavior |
| **Date, Open, High, Low** | Supporting Financial Data: Provides complete OHLC context for granular market analysis |

# Data Background – Sample Records

An illustrative snapshot of the raw dataset structure, showing the alignment of daily news headlines with OHLC price data, trading volume, and sentiment labels as used in model training and evaluation.

Each record pairs a single trading day's headline with its full OHLC + Volume market snapshot and a categorical sentiment label – creating a clean supervised learning dataset aligned to real market conditions.

| Date | News | Open | High | Low | Close | Volume | Label |
|---|---|---|---|---|---|---|---|
| 01-02-2019 | The dollar minutes ago tumbled to 106 67 from above 109 a t | 38.72 | 39.71 | 38.56 | 39.48 | 130672400 | 1 |
| 01-02-2019 | By Wayne Cole and Swati Pandey SYDNEY Reuters The Ja | 38.72 | 39.71 | 38.56 | 39.48 | 130672400 | -1 |
| 01-02-2019 | By Stephen Culp NEW YORK Reuters Wall Street edged hi | 38.72 | 39.71 | 38.56 | 39.48 | 130672400 | 0 |
| 01-02-2019 | By Wayne Cole SYDNEY Reuters The Australian dollar was | 38.72 | 39.71 | 38.56 | 39.48 | 130672400 | -1 |
| 01-02-2019 | Investing com Asian equities fell in morning trade on Thursda | 38.72 | 39.71 | 38.56 | 39.48 | 130672400 | 1 |
| 01-02-2019 | Bloomberg Apple Inc s Asian suppliers tumbled following | 38.72 | 39.71 | 38.56 | 39.48 | 130672400 | -1 |
| 01-02-2019 | Investing com The Japanese yen jumped on Thursday in Asi | 38.72 | 39.71 | 38.56 | 39.48 | 130672400 | 1 |
| 01-02-2019 | RBC Capital reiterates its Outperform rating and 220 price tar | 38.72 | 39.71 | 38.56 | 39.48 | 130672400 | 0 |
| 01-02-2019 | Reuters Roku Inc O ROKU said on Wednesday it will beg | 38.72 | 39.71 | 38.56 | 39.48 | 130672400 | 1 |

# Word Embeddings — Technical Comparison

A direct head-to-head of both embedding strategies across the dimensions that matter most for financial NLP — from syntactic preservation to measurable model performance.

| Feature | Word2Vec (Baseline) | Sentence Transformer (Winner) |
|---|---|---|
| **Embedding Level** | Word-level (requires post-hoc averaging) | Sentence-level (context-aware from the ground up) |
| **Output Type** | Static average of individual word parts | Holistic semantic unit representing the full headline |
| **Syntactic Logic** | Lost — Bag-of-Words ignores word order and relationships | Preserved — Self-Attention weights each word relative to all others |
| **Financial Relevance** | Destroys the "who" and "what" critical to headline interpretation | Captures subject-action-object relationships in market news |
| **Market Result** | Higher rate of missed Negative and Neutral signals | **9% F1-Score improvement** — superior minority-class detection |