King Fahd University of Petroleum & Minerals

Mathematics Department

Master of Science in Data Science & Analytics

MATH-619

A CNN-LSTM Hybrid Model for Multivariate Stock Price Forecasting Using Technical and Sentiment Indicators

Dataset Summary

Thursday, 18th September 2025

Student Name & ID	**Advisor**
ABDULLAH ALHARBI g202415480	Dr. Mousa Ahmad Al-Bashrawi **Co-advisor** Mr. Mohammed Agbawi

Securing Data – Summary of Data and Resources

This project integrates high-quality financial and sentiment data to build a multivariate dataset suitable for deep learning-based stock movement prediction. The secured datasets include long-term technical indicators, market volatility measures, and short-term sentiment signals.

2.1 Stock Price Data and Technical Indicators (2015–2025)

Source:

TradingView

Companies:

Apple (AAPL), Google (GOOG), Amazon (AMZN), Microsoft (MSFT), Tesla (TSLA)

Period:

2 January 2015 – 30 October 2025

Extracted Fields:

The technical indicators used in this study are limited to the following:

- EMA (Exponential Moving Average)
- RSI (Relative Strength Index)
- MACD (Moving Average Convergence Divergence)
- MACD Histogram
- Bollinger Bands (Upper, Middle, Lower)
- Volume

These indicators were selected because they capture key aspects of market behavior, including trend direction, momentum, volatility, and buying or selling pressure.

2.2 Market Volatility Data (VIX Index)

Source:

TradingView / Yahoo Finance

Period:

2015–2025

The CBOE Volatility Index (VIX) is included as a macro-level measure of market sentiment and uncertainty. VIX reflects overall market fear and risk perception, providing important context beyond stock-specific technical indicators.

2.3 Twitter Sentiment Data (Pre-Scored)

Source:

Kaggle — *"Stock Tweets for Sentiment Analysis and Prediction"* by Equinxx (2023)

Period Covered:

30 September 2021 – 29 September 2022

Key Points:

- The dataset includes tweets related to the selected companies.
- Sentiment scores are already computed (range −1 to +1). → No FinBERT, VADER, or additional NLP processing was performed in this project.
- The dataset includes engagement metrics such as likes, comments, and retweets.
- Daily average sentiment and tweet volume were aggregated for each company.

This dataset captures public mood and behavioral signals that complement traditional technical indicators.

2.4 Company–Tweet Mapping File

Source:

Kaggle (part of the same dataset)

This file maps each tweet ID to its associated ticker symbol, ensuring that sentiment is accurately assigned to the correct company (AAPL, GOOG, AMZN, MSFT, or TSLA).

2.5 Final Dataset Structure

Dataset A – Technical Only (2015–2025)

- OHLC prices
- EMA
- RSI
- MACD and MACD Histogram
- Bollinger Bands
- Volume
- VIX

Dataset B – Technical + Sentiment (2021–2022)

- All technical indicators listed above
- VIX
- Daily Twitter sentiment score (pre-scored)
- Daily tweet volume

The dual-dataset approach enables a direct and fair comparison between technical-only forecasting and sentiment-enhanced forecasting.

2.6 Rationale for Selecting Technology Companies

The study focuses on technology-sector companies (AAPL, GOOG, AMZN, MSFT, and TSLA) because:

- They are fundamentally strong and highly liquid.
- They exhibit higher volatility than companies in many other sectors, making them suitable for deep learning-based time-series modeling.
- They attract significant social-media and news attention, resulting in rich sentiment data.

This ensures consistent, high-quality data for both technical and sentiment analysis.