A User-Configurable Deep Learning Framework for Automated Cryptocurrency Market Prediction Using Multilingual LLM-Based Sentiment Fusion

By **Ashley Beebakee**, MSc
Submitted to The University of Nottingham
In **September 2025**

In partial fulfilment of the conditions for the award
of the degree of **Master of Science in Computer Science**
with Artificial Intelligence

20232303

Supervised by Kai Xu

School of Computer Science

University of Nottingham

Declaration of Authorship

I hereby declare that this dissertation is all my own work, except as indicated in the text:

Signature: AB

Date: 12/09/2025

Abstract

This dissertation presents the design and evaluation of a user-configurable deep learning framework for cryptocurrency market prediction via the integration of multilingual LLM-based sentiment analysis with historical price data. Due to volatility of cryptocurrency markets and their dependence on public sentiment, the combination of financial price time-series with sentiment signals may yield higher predictive performance. Existing approaches often rely on price-only models or English-only sentiment signals, not enough frameworks allow configurable, multilingual sentiment fusion with deep learning architectures. The framework supports multiple LLMs (i.e. GPT-5, LLaMA-3.1, Orca2 and BLOOMZ) for sentiment extraction (manual and automatic), two fusion strategies (early, late), and configurable deep learning models (LSTM, CNN, CNN-LSTM). Over 85% of the system was implemented, including sentiment extraction pipelines, fusion modules, and configurable training with evaluation metrics. Experiments conducted across two months of Bitcoin, Ethereum and Dogecoin historical data (individually), demonstrate that sentiment-enhanced models, especially with early fusion, outperform price-only models.

Keywords: *deep learning; sentiment analysis; cryptocurrency; Bitcoin (BTC); Ethereum (ETH; Dogecoin (DOGE); multilingual language models (LLMs); prompt engineering; configurable framework; Streamlit; MLflow.*

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Kai Xu, for his invaluable guidance, constructive feedback, and continuous support throughout the duration of this project.

I am deeply thankful to my girlfriend, for her unwavering encouragement, patience and motivation, which kept me focused during challenging times, in turn allowing me to handle time management better than I would usually do.

Finally, I would like to extend my heartfelt appreciation to my family for their constant love, belief in me, and support that made this journey possible right up to the very end.

Table of Contents

List of Tables

List of Acronyms

ACRONYM	DEFINITION
AI	Artificial Intelligence
API	Application Programming Interface
ARIMA	Autoregressive Integrated Moving Average
BLOOMZ	BigScience Large Open-science Open-access Multilingual Language Model – Zero-shot
BTC	Bitcoin
CNN	Convolutional Neural Network
DA	Directional Accuracy
EMA	Exponential Moving Average
ETH	Ethereum
DOGE	Dogecoin
FX	Foreign Exchange
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
GPT	Generative Pre-trained Transformer
LLAMA	Large Language Model Meta AI
LSTM	Long Short-Term Memory
MAPE	Mean Absolute Percentage Error
MDA	Mean Directional Accuracy
MOM	Minutes of Meeting
NASDAQ	National Association of Securities Dealers Automated Quotations
NOD	Notes of Development
NYSE	New York Stock Exchange
MSE	Mean Squared Error
RMSE	Root Mean Squared Error

1. Introduction

1.1 Context & Background

When observing the financial ecosystems of the twenty-first century, one might be familiar with traditional foreign exchange markets (FX), the New York Stock Exchange (NYSE) or perhaps the National Association of Securities Dealers Automated Quotations (NASDAQ), however, the most dynamic and unpredictable one that has been evolving rapidly throughout the past two decades is the cryptocurrency market. Operating around the clock globally, are cryptocurrencies such as Bitcoin (BTC), Ethereum (ETH) or Dogecoin (DOGE), which on top of being decentralised, are frequently driven by shifts in public and investor sentiment rather than relying solely on fundamental data (Gurgul, Lessmann and Karl Härdle, 2025; Jung, Lee and Kim, 2025). This volatility represented by sudden jumps and periods of sharp reversals present significant challenges for when it comes to predictive modelling. Due to the dynamics of non-linearity and sensitivity to sentiment, conventional econometric models such as autoregressive integrated moving average (ARIMA) and generalized autoregressive conditional heteroskedasticity (GARCH) are increasingly insufficient. As a response to this, modern research took a turn towards the usage of advanced deep learning architectures: long short-term memory networks (LSTM), convolutional neural networks (CNNs), or bidirectional LSTM (Bi-LSTM) which allow temporal dependencies to be captured better in markets with high volatility (Gurgul, Lessmann and Karl Härdle, 2025).

Even so, sophisticated price-based models often fall short due to the social and psychological aspects of cryptocurrency trading not being taken into consideration. Trading behaviour and price movements have been shown to be significantly influenced by market sentiment communicated via platforms like Reddit and various news sources (Roumeliotis, Tselikas and Nasiopoulos, 2024; Long *et al.*, 2025). In particular, sentiment signals extracted from textual data can act as an early warning indicator for short-term market trends that traditional time-series models fail to capture (Jung, Lee and Kim, 2025). Despite this, the scope of existing literature remains narrow, due to either focusing on single-cryptocurrency models, limiting sentiment to the English language, or lacking a structured comparison of different fusion strategies between sentiment and price data (Gurgul, Lessmann and Härdle, 2025).

The development of Large Language Models (LLMs), which include models like generative pre-trained model (GPT), large language model meta AI (LLaMA), or BigScience Large Open-science Open-access Multilingual Language Model – Zero-shot (BLOOMZ), has increased the accuracy and subtlety of interpreting the feelings behind words when performing sentiment analysis, beyond lexicon-based methods which implies just looking up words in a list (Roumeliotis, Tselikas and Nasiopoulos, 2024). Even from noisy financial texts, the performance in sentiment extraction exhibited by these LLMs appears to be robust. Some recent models, i.e. FinBERT, have been fine-tuned on domain specific corpora resulting in better parsing of finance-oriented language (Dashtaki *et al.*, 2025). Nevertheless, the application of LLMs for the extraction of multilingual sentiment related to cryptocurrency remains rare, despite the global market revolving around its linguistic nature and diversity.

Currently, a vast number of predictive systems are designed as static, one-off pipelines that lack modularity. This creates difficulties in terms of adaption, configuration, or extension without being deeply involved in its technical aspects, thereby raising concerns about reproducibility and experimentation (Gurgul, Lessmann and Härdle, 2025; Jung, Lee and

Kim, 2025). To contrast this, a user-configurable framework could enable researchers and traders (rookies to experts) to explore different data sources, LLM models, sentiment models, and time intervals. In turn, this enhances flexibility, transparency, and reproducibility, a contribution that the literature currently lacks.

1.2 Aims and Project Objectives

In order for these gaps to be addressed, this study presents the design, implementation, and evaluation of a user-configurable deep learning framework adjusted for cryptocurrency market prediction via the integration of multilingual sentiment analysis performed by LLMs, along with its fusion with historical time series data. The framework is structured to handle multiple cryptocurrencies: BTC, ETH and DOGE whilst operating across flexible temporal resolutions (from five-minute intervals to daily aggregates), depending on what the user selects as part of their desired configuration. Moreover, the sentiment data is sourced from Reddit (legacy) and NewsAPI, preprocessed, and aligned with OHLCV (Open-High-Low-Close-Volume).

For the generation of sentiment scores, a batch of carefully selected quantised multilingual open-source LLMs will be used: LLaMA 3.1 Instruct (4-bit and 2-bit), Orca 2 7B (6-bit) and BLOOMZ 7B (4-bit). On other hand, close-source LLMs will only be used for testing purposes on individual sentiment score extraction. Although fusion strategies such as early fusion, late fusion, and attention-based fusion are planned but not yet implemented, they are foundational to the framework design. Lastly, evaluation will be the analysis of performance metrics such as Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Mean Directional Accuracy (MDA), through the comparison of models across different configurations. The entire system will be accessible through a Streamlit-based GUI, with a planned-out design that promotes flexibility, transparency and reproducibility.

To adhere with this aim, this study will follow these project objectives (POs):

- **PO1:** to acquire and preprocess sentiment-related textual data from Reddit (legacy) and NewsAPI for alignment with cryptocurrency time-series data at configurable temporal resolutions.
- **PO2:** to apply multilingual sentiment extraction using quantised LLMs (LLaMA 3.1 8B Instruct, Orca 2 7B, BLOOMZ 7B, GPT-5) to generate sentiment scores.
- **PO3:** to architect and implement the framework's modular pipeline capable of fusing sentiment and price data, configurable by end users via a GUI.
- **PO4:** to design and compare fusion strategies, specifically early fusion, late fusion, and attention-based fusion, within the deep learning pipeline.
- **PO5:** to evaluate model performance using statistical and financial metrics (RMSE, MAPE, directional accuracy), comparing sentiment-augmented models to price-only models.
- **PO6:** to deliver a reproducible and flexible framework that includes an interactive GUI, enabling experimentation without in-depth programming knowledge.

1.3 Research Questions

From these project objectives, the study poses the following Research Questions (RQs):

- **RQ1:** does incorporating sentiment extracted from multilingual LLMs improve predictive performance over price-only models?

- **RQ2:** which fusion strategy (early fusion, late fusion, or attention-based fusion) most effectively integrates sentiment and price data?
- **RQ3:** are multilingual LLM-generated sentiment features advantageous compared to sentiment extracted solely in English?
- **RQ4:** can a modular, use-configurable framework systematically support experimentation and hypothesis testing in cryptocurrency prediction?

1.4 Contributions

This dissertation contributes to the following fields:

1. **Engineering Contribution:** development of an open-source, modular, user-configurable framework that enables the integration of multilingual sentiment with time-series data for cryptocurrency forecasting. The inclusion of a GUI extends accessibility to non-expert users.
2. **Research Contribution:** empirical evaluation of fusion strategies and the comparative performance of multilingual LLMs in the context of cryptocurrency prediction, thus filling a significant gap in empirical literature.
3. **Practical Contribution:** a system enabling experimentation by individual researchers, analysts, or hobbyists, who can explore different cryptocurrencies, resolutions, sentiment sources, and configurations without coding from scratch.

1.5 Dissertation Structure

This dissertation has been structured in an organised manner to concisely convey the way that every phase unfolded and how they are interlinked with each other.

- **Chapter 3:** a comprehensive Literature Review exploring prior work in cryptocurrency forecasting sentiment analysis, fusion methods, multilingual NLP, and configurable predictive frameworks.
- **Chapter 4:** Methodology, detailing data acquisition, preprocessing steps, sentiment scoring pipelines, and the proposed fusion strategies.
- **Chapter 5:** Implementation, covering system architecture, modular components, GUI design, and details of LLM integration and quantisation.
- **Chapter 6:** Testing, presenting comparative performance across model configurations via RMSE, MAPE and directional accuracy, including baseline comparisons and analysis.
- **Chapter 7:** Discussion, reflecting on results, drawing connections with literature, acknowledging limitations, and exploring theoretical and practical implications.
- **Chapter 8:** Conclusion and Future Work, summarising key findings, addressing research questions explicitly, and suggesting directions for extending the framework.

By blending advanced natural language processing (NLP) techniques with a configurable architecture for the system, this study aims to expand both academic understanding and support practical application in predicting the cryptocurrency market. A strong emphasis on reproducibility, flexibility and transparency can be observed through how the project was thought out, these qualities are essential when it comes to computational finance research, especially in a highly volatile and sentiment-driven market environment.

2. Literature Review

2.1 Cryptocurrency Forecasting and Time-Series Methods

When it comes to cryptocurrency markets, Bitcoin, Ethereum and Dogecoin exhibit similar behaviour to high-volatility assets (sharp jumps, volatility clustering and changing regimes), yet the intensity and frequency in which it occurs, often exceeds those observed in traditional markets. Early empirical work modelled crypto volatility using GARCH-family processes, showing that heavy-tailed, leverage-sensitive specifications outperform simpler baselines for Bitcoin returns (Katsiampa, 2017). These studies established that volatility is both persistent and asymmetric, and that model choice influences risk estimation and forecast stability. This is an observation that motivates more expressive, non-linear approaches when it comes to moving from volatility modelling to price or return forecasting (Katsiampa, 2017).

Conventional econometric models (i.e. ARIMA, VAR or GARCH variants) remain important as transparent baselines, but they struggle when confronted with non-stationarity, structural breaks, and strong non-linearities, which are all quite common in crypto time-series data. Recent studies report that while classical methods can perform competitively on simpler series or short horizons (Bouteska *et al.*, 2024), ensemble learning and deep learning methods tend to dominate as complexity grows. This occurrence especially takes place for multi-asset settings (i.e. BTC, ETH, XRP, LTC) and multi-step horizons. Notably, modern evaluations highlight pitfalls such as in-sample overfitting and the need for robust out-of-sample validation, walk-forward testing, and feature engineering when assessing models for crypto forecasting (Bouteska *et al.*, 2024).

The shift from linear models to deep learning (DL) in financial forecasting grew exponentially as a result of the ability of neural architectures being able learn non-linear, long-range dependencies. LSTM networks and their bidirectional or gated variants (i.e. GRU) became popular due to their gating mechanisms, allowing the mitigation of vanishing gradients and enabling the modelling of longer context windows. Seminal work on equities showed that LSTMs can outperform traditional methods in directional prediction (Fischer and Krauss, 2018), and subsequent crypto-focused studies adopted LSTM or GRU to capture rapidly evolving temporal patterns. This resulted in improvements being often reported for RMSE and MAPE or directional accuracy (DA) relative to classical baselines (Fischer and Krauss, 2018). Nevertheless, results are sensitive to the forecasting horizon, data preprocessing (scaling, windowing), and the stability of underlying regimes, which are factors that can erode gains if not carefully managed (Fischer and Krauss, 2018; John, Binnewies and Stantic, 2024).

Beyond recurrent models, attention mechanisms and transformer-based approaches have gained traction because of how they can learn long-range and cross-feature interactions more directly. The Temporal Fusion Transformer (TFT) is a notable example designed for multi-horizon forecasting. It integrates recurrent layers to identify local patterns with interpretable self-attention for long-term dependencies and introduces gating to focus on salient covariates (Lim *et al.*, 2020). TFT-style models are particularly relevant for crypto because of how they can incorporate these (i.e. trading volume, macro proxies) alongside price histories. They are also able to provide variable-importance insights that help practitioners observe about model behaviour in volatile regimes (Lim *et al.*, 2020). Emerging crypto studies increasingly benchmark transformers against LSTM and CNN hybrids, often finding competitive or superior performance when sufficient data and regularisation is available (John, Binnewies and Stantic, 2024; Zhang, Cai and Wen, 2024).

Recent surveys consolidate these trends. Comprehensive reviews from the years of 2024 and 2025 report that deep learning methods (LSTM, GRU, CNN-LSTM hybrids, Transformers) generally outperform classical baselines for cryptocurrency price forecasting, while warning that a set of experimental protocols can inflate reported gains. They further emphasise that model comparisons are often fragmented by asset, horizon, and metric choice, complicating cross-study synthesis (John, Binnewies and Stantic, 2024; Wu *et al.*, 2024; Zhang, Cai and Wen, 2024). These surveys also note that while predictive accuracy can improve, interpretability and robustness remain open concerns, encouraging the use of attention and variable-importance, along with stress tests across market regimes (John, Binnewies and Stantic, 2024; Wu *et al.*, 2024; Zhang, Cai and Wen, 2024).

Two additional methodological points are noticeable for this dissertation. First, horizon design matters: crypto forecasting tasks range from very short-term (i.e. 5-minute intervals) to daily or weekly horizons. Different algorithms may dominate at different horizons. LSTM and GRUs often excel at short to medium horizons when trained with appropriate sliding windows, whereas Transformer variants can handle longer horizons and multiple inputs effectively (Fischer and Krauss, 2018; Lim *et al.*, 2020). Second, feature sets influence outcomes: univariate close-price models are simpler but frequently underperform multivariate models that incorporate OHLCV, technical indicators, and regime proxies. High-capacity models also require careful regularisation (dropout, early stopping), learning-rate scheduling, and hyperparameter search to avoid overfitting, issues that will be addressed explicitly in this project's methodology and evaluation design.

In summary, the literature establishes that crypto assets present challenging statistical properties that strain classical forecasting methods, deep architectures, especially LSTM, GRU and Transformer families, provide measurable benefits when evaluated consistently practice demands strict out-of-sample protocols and interpretability analysis (including multilingual LLMs) and then fusion strategies that combine sentiment with price-based predictors, two dimensions central to this dissertation's hypothesis and system design.

3. Methodology

3.1 Research Design & Overview

This study follows a straightforward idea: take what we know works in time-series modelling, add sentiment to it since markets are not able to do that, then make the whole thing configurable so different users can test their own hypotheses without rewriting any code. The framework targets BTC, ETH, and DOGE, because they represent large-cap, smart-contract and "meme" behaviour respectively. This is useful when it comes to diversity, for testing whether sentiment actually adds any value beyond prices.

Methodologically, the work blends a design-science approach (building a working system that people can configure and run) with an experimental evaluation (compare price-only vs sentiment-augmented models across settings). This aligns with best practice for forecasting research: stating a clear protocol, using transparent baselines, and avoiding easy mistakes. In practice, that implies clearing data pipelines, chronological splits, and metrics that make sense for both error and direction.

Price time-series data is sourced via Yahoo Finance through the yfinance library. Within the current build, hourly (lh) bars are more reliable since daily (1d) bars are not sufficient to undergo fusion due to a lack of new sources data. This is a reasonable starting point given the stability of hourly aggregates and the libraries used support for hourly frequency, with the usual reminder that Yahoo's data is intended for personal use and therefore should respect their terms and conditions.

The textual side intentionally mixes community chatter (Reddit) and news (NewsAPl) to convey different information flows. Reddit captures the retail pulse and narratives that often move crypto quickly, we call this “hype”, whereas the news offers faster, higher-signal events (policies, exchange incidents or large corporate actions). The Reddit pipeline focuses on domain-relevant subreddits: r/Cryptocurrency, r/Bitcoin, r/Ethereum, r/Dogecoin and r/CryptoMarkets, along with a simple but effective URL-based de-duplication to remove repeated scraped posts. This choice is supported by recent work showing that subreddit activity and mood correlate with Bitcoin market dynamics and can explain short-term variation. On the news side, the framework connects to the NewsAPl endpoint and exposes language filters (English, German, French, Spanish, Italian, or “All”) with a cap of up to one hundred articles per call since that is the documented maximum page size. The starting date defaults to 1 July 2025 (modifiable by the user) so it's easy to run contemporary experiments without having to scrape far back. The NewsAPl docs clearly describe the page size limits and paging behaviour, which the framework respects.

The framework computes a single sentiment score in the range of -1 to 1 for each text item, where -1 is negative, 0 is neutral, and 1 is positive. Scores are produced using multiple LLMs that are practical to run locally or via API in lightweight form: LLaMA 3.1 8B Instruct (2-bit & 4-bit), Orca 2 7B (6-bit), BLOOMZ 7B (4-bit), and GPT-5, with the latter of the closed-source LLMs not having been implemented due to time constraints. The point is not to claim one "best" model, but to let users compare them on the same pipeline. Finance evidence recently supports this direction. Transformer models and instruction-tuned LLMs extract sentiment signals that outperform older lexicon-based approaches, as a result of this, these signals can be useful for return forecasting when integrated carefully (Kelly and Xiu, 2023).

4. Experiments & Results

The purpose of this section was to turn our study into evidence by reporting what was tested, how it was carried out, and what conclusions it yielded in relation to the Research Questions (RQs) we set out in Section 1.3. The design of these tests followed standard forecasting disciplines: transparent baselines, strictly out-of-sample evaluation, and statistical tests to support any claims of improvements (Diebold and Mariano, 1995; Hyndman and Koehler, 2006).

4.1 Experimentation Setup

All experiments carried out used BTC, ETH and DOGE with hourly (1h) bars, following the procedures described in Sections 3 and 4. News sources data come from legacy Reddit (specific subreddits) and NewsAPI (configurable languages). Sentiment is scored per row between -1 (negative) and 1 (positive) with LLMs and aggregated to hourly features (mean, median, dispersion, counts), as specified in Section 3.3 and Section 3.4.

Each of the three assets are split chronologically into 70% train, 15% validation and 15% test whereas the transforms (scalers, rolling stats) are fitted on train only and applied forward. The full windows and split boundaries are summarised in **Error! Reference source not found.**. Seeds are fixed, and every run is logged, with the inclusion of all parameters, metrics, artifacts. This allows results to be reproduced and compared, as recommended in the forecasting literature and common ML practice (Hyndman and Koehler, 2006).

For conciseness, the experiments to answer the research questions were organised as follows, as shown in Table 1:

- Experiment 1 (Exp1): price-only baseline.
- Experiment 2 (Exp2): early fusion (Reddit & News, EN language filter).
- Experiment 3 (Exp3): early fusion (ALL languages) to test multilingual value.
- Experiment 4 (Exp4): Sentiment-model choice (i.e. LLaMA 3.1 4-bit & 2-bit vs Orca 2).

Note bene: experiment 5 does not include 'BLOOMZ 7B (4-bit)' due to time constraints not allowing for an optimised prompt template for it to extract sentiment scores from the news sources appropriately. As a result, the column 'Sentiment_Bloomz7b1' is incomplete, thus cannot be utilised for evaluation.

Table 1. Experiment matrix (Exp1 to Exp4)

ID	Asset(s)	Language	Fusion	Architecture	Sentiment LLM(s)
Exp1	BTC, ETH & DOGE	N/A	None	LSTM, CNN & CNN-LSTM	N/A
Exp2	BTC, ETH & DOGE	EN	Early	LSTM, CNN & CNN-LSTM	LLaMA 3.1 (4-bit)
Exp3	BTC, ETH & DOGE	ALL	Early	LSTM, CNN & CNN-LSTM	LLaMA 3.1 (4-bit)
Exp4	BTC, ETH & DOGE	EN (or ALL)	Early	LSTM, CNN & CNN-LSTM	LLaMA 3.1 (4-bit) & Orca 2

4.2 Model Training

A standard loop with early stopping on validation loss is used for the model training with fixed hyperparameters per model type (LSTM, CNN, CNN-LSTM) and reused across Exp1 to Exp4 to prevent 'moving target' comparisons. The metrics for the experiments are always computed on the test slice.

4.3 Results

Refer to Tables A.1, A.2 and A.3 in Appendix A.

5. Reflection

It has been quite an adventurous three months, with this project forcing me to balance ambition (LLM sentiment, multilingual news, fusion) with practicality (hourly data only, early fusion, strict out-of-sample testing). A share of lessons stood out. It is clear that I spent too much time developing the framework instead of writing the report, hence its incompletion.

5.1 Positives (What went well?)

Steadily building a clean, configurable pipeline paid off. Once all components (scraping, sentiment scoring, fusion, etc.) became modular, I was able to swap models and settings without having to rewrite any code, this meant that I could find the results more reliable due to their consistency.

With the use of a configuration file (.yaml) and the integration of MLflow, reproducibility was achieved to a certain extent, meaning that I spent less time guessing what was changing and more time interpreting outcomes.

From the results, BTC showed clear gains when sentiment was fused, whilst DOGE showed small but consistent gains with English-only sentiment. These validate the idea that text helps when retail attention is high.

5.2 Negatives (What went wrong?)

Multilingual news sources were not beneficial as I thought it would be. For ETH and DOGE, it introduced noise and negatively affected RMSE and MDA. Naively aggregating languages without refining the source was optimistic, but not the best idea.

MAPE for cryptocurrency was not very informative at an hourly horizon (especially for DOGE). I decided to keep it for completeness, but RMSE and DA conveyed the truth of what was going on.

I ran out of time to implement late and attention-based fusion, including the evaluation of additional metrics with debugging. Dropping them kept the dissertation honest, but it also limited what I could say about fusion choice and performance between LLM sentiment and deep learning architectures.

5.3 Project Management

- The creation of a Minutes of Meeting (MoM) and a Notes of Development (NoD) log helped me a lot in terms of keeping track of what had been discussed with the supervisor and what actions need to be taken, also allowing me to spot dead-ends quickly. However, the complexity of this increased as the project size increased.
- The main component I would change during the course of a future project would be to prepare an evaluation plan much earlier with a higher number of distinct experiments to avoid spending hours and hours on features that never made the final cut (though I hope it will be of some use to future students).

5.4 Ethics

- All data used within this project came from public sources
- API usage adhered to provider terms and rate limits
- The developed pipeline avoids collecting personal data and keeps experiment artifacts minimal (config, metrics & plots), to allow the work to be reproducible.

6. Conclusion & Future Work

By linking our evidence back to our research questions in Section 1.3, we can produce conclusive answers:

- **RQ1:** does incorporating sentiment extracted from multilingual LLMs improve predictive performance over price-only models?
    - Yes, BTC benefitted the most (12-21% RMSE reduction) whilst DOGE saw consistent gains with English-only sentiment and ETH gained slightly in direction but not in magnitude.
- **RQ2:** which fusion strategy (early fusion, late fusion, or attention-based fusion) most effectively integrates sentiment and price data?
    - In this project, only early fusion was evaluated, which in turn worked well for BTC, however without late and attention-based fusion implemented, we cannot produce a definitive answer to this research question.
- **RQ3:** are multilingual LLM-generated sentiment features advantageous compared to sentiment extracted solely in English?

  - Multilingual value seems to be asset dependent and seemed to be beneficial to BTC when paired with the right LLM but negatively affected ETH and DOGE, this implies that stricter source selection or per-language weighting is required.
- **RQ4:** can a modular, use-configurable framework systematically support experimentation and hypothesis testing in cryptocurrency prediction?
  - Yes, chronological splits, fixed seeds, and logged experiment runs delivered transparency and reproducibility

6.1 Limitations

I acknowledge that the prototype that was developed contains some bugs and glitches, and is limited to a certain extent:

- Only daily (1d) & hourly (1h) price time-series data can be downloaded from Yahoo Finance (yfinance) using the GUI.
- Late fusion was partly implemented whilst attention-based fusion was not included at all in this prototype.
- Lack of news sources data (Reddit & NewsAPI), meaning that long date ranges cannot be tested for a more robust model to be trained on.
- Multilingual aspect of the news sources data lacked quality controls beyond language codes.

6.2 Future Work

Lots of time was taken to build upon this project itself in a modular way, allowing future students who will undertake this project 'LLM for Automated Trading' to have access to everything that was thought, written, mistaken and subsequently developed.

All the source code of the project (including backups, diagrams, samples), along with clearly written guides (i.e. for installation) are readily available in the 'Vis4Sense' project folder under 'Ashley Beebakee'. Furthermore, GitHub issues were created to denote the Work Packages (WPs) with their respective Milestones, those include well-written details about specific features implemented into the project.

There are countless additional features one would have loved to implement, however, due to the time constraint of only three months, maybe such features can be developed by a future student:

- *To include additional open-source LLMs into the framework (legacy to modern).*
- *To integrate all closed-source LLM APIs with error handling and limits to control credit usage (i.e. Gemini & Claude).*
- *To benchmark and format the prompt templates for each LLM (open-source & closed-source) for them to give the best sentiment score they possibly can.*
- *To fully implement late and attention-based fusion to carry out further experiments for comparison between price-only and sentiment-augmentation.*
- *To enhance the integration of MLflow tracking when the model is training (i.e. define an experiment naming convention based on the configuration of the framework.*
- *To allow the model to train on different price time-series intervals, i.e. 1m, 5m, 15m, etc. which requires debugging of the yfinance package or potentially the integration of the Alpaca API.*

- *To implement additional little features for practice, i.e. custom prompt for chosen LLM to see if you can make it output a sentiment score manually.*
- *To restructure the GUI based on user testing, especially in terms of navigability.*
- *To include more data from the news sources, i.e. descriptions for the Reddit posts that are being scraped and more sources to get data from (not just Reddit & NewsAPI).*

This list is endless, thereby we should halt it here with the more significant points having been highlighted.

To conclude, this study delivers a working, reproducible framework that shows when, and how sentiment can be beneficial for crypto forecasting. The gains are obtained where narrative risk is high (BTC, DOGE) and more modest where the baseline is already strong (ETH). The immediate next steps would be the implementation of the italic points above, they should be fairly straightforward, except requiring careful restructuring of existing scripts, these improvements should make the framework have better consistency, interpretability and reproducibility.

Bibliography

Bouteska, A. *et al.* (2024) ‘Cryptocurrency price forecasting – A comparative analysis of ensemble learning and deep learning methods’, *International Review of Financial Analysis*, 92, p. 103055. doi: 10.1016/j.irfa.2023.103055.

Dashtaki, S.M. *et al.* (2025) ‘A Multisource Fusion Framework for Cryptocurrency Price Movement Prediction’. arXiv. doi: 10.48550/arXiv.2409.18895.

Diebold, F.X. and Mariano, R.S. (1995) ‘Comparing Predictive Accuracy’, *Journal of Business & Economic Statistics*, 13(3), pp. 253–263. doi: 10.1080/07350015.1995.10524599.

Fischer, T. and Krauss, C. (2018) ‘Deep learning with long short-term memory networks for financial market predictions’, *European Journal of Operational Research*, 270(2), pp. 654–669. doi: 10.1016/j.ejor.2017.11.054.

Gurgul, V., Lessmann, S. and Härdle, W.K. (2025) ‘Deep learning and NLP in cryptocurrency forecasting: Integrating financial, blockchain, and social media data’, *International Journal of Forecasting*, 41(4), pp. 1666–1695. doi: 10.1016/j.ijforecast.2025.02.007.

Gurgul, V., Lessmann, S. and Karl Härdle, W. (2025) ‘Deep learning and NLP in cryptocurrency forecasting: Integrating financial, blockchain, and social media data’, *International Journal of Forecasting* [Preprint]. doi: 10.1016/j.ijforecast.2025.02.007.

John, D.L., Binnewies, S. and Stantic, B. (2024) ‘Cryptocurrency Price Prediction Algorithms: A Survey and Future Directions’, *Forecasting*, 6(3), pp. 637–671. doi: 10.3390/forecast6030034.

Jung, H.S., Lee, H. and Kim, J.H. (2025) ‘Detecting Bitcoin Sentiment: Leveraging Language Model Applications in Sentiment Analysis for Bitcoin Price Prediction’, *Neural Processing Letters*, 57(77). doi: 10.1007/s11063-025-11787-1.

Katsiampa, P. (2017) ‘Volatility estimation for Bitcoin: A comparison of GARCH models’, *Economics Letters*, 158, pp. 3–6. doi: 10.1016/j.econlet.2017.06.023.

Kelly, B. and Xiu, D. (2023) ‘Financial Machine Learning’, *Foundations and Trends® in Finance*, 13(3–4), pp. 205–363. doi: 10.1561/0500000064.

Lim, B. *et al.* (2020) ‘Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting’. arXiv. doi: 10.48550/arXiv.1912.09363..

Long, S. *et al.* (2025) ‘From whales to waves: Social media sentiment, volatility, and whales in cryptocurrency markets’, *The British Accounting Review*, p. 101682. doi: 10.1016/j.bar.2025.101682.

Roumeliotis, K.I., Tselikas, N.D. and Nasiopoulos, D.K. (2024) ‘LLMs and NLP Models in Cryptocurrency Sentiment Analysis: A Comparative Classification Study’, *Big Data and Cognitive Computing*, 8(6), p. 63. doi: 10.3390/bdcc8060063.

Wu, J. *et al.* (2024) ‘Review of deep learning models for crypto price prediction: implementation and evaluation’. arXiv. doi: 10.48550/arXiv.2405.11431.

Zhang, J., Cai, K. and Wen, J. (2024) ‘A survey of deep learning applications in cryptocurrency’, *iScience*, 27(1), p. 108509. doi: 10.1016/j.isci.2023.108509.

Appendix A – Experiment Tables

Table A.1 BTC full experiment breakdown (Exp1 to Exp4)

ID	Fusion	Language	LLM	Architecture	RMSE	MAPE (%)	MDA (%)	R²
Exp1	None	None	None	LSTM	704.77	10.86	0.45	0.90
	None	None	None	CNN	1461.51	20.98	0.47	0.56
	None	None	None	CNN-LSTM	780.28	11.58	0.44	0.87
Exp2	Early	EN	LLaMA 3.1 (4-bit)	LSTM	964.83	15.23	0.45	0.81
	Early	EN	LLaMA 3.1 (4-bit)	CNN	943.34	13.45	0.49	0.82
	Early	EN	LLaMA 3.1 (4-bit)	CNN-LSTM	617.09	9.27	0.46	0.92
Exp3	Early	ALL	LLaMA 3.1 (4-bit)	LSTM	854.15	12.95	0.47	0.85
	Early	ALL	LLaMA 3.1 (4-bit)	CNN	1746.02	30.84	0.47	0.37
	Early	ALL	LLaMA 3.1 (4-bit)	CNN-LSTM	621.64	10.24	0.46	0.92
Exp4	Early	ALL	LLaMA 3.1 (2-bit)	CNN-LSTM	560.32	8.78	0.47	0.94
	Early	ALL	LLaMA 3.1 (4-bit)	CNN-LSTM	636.58	9.76	0.45	0.92
	Early	ALL	Orca 2 (4-bit)	CNN-LSTM	556.37	8.72	0.48	0.94

Table A.2 ETH full experiment breakdown (Exp1 to Exp4)

ID	Fusion	Language	LLM	Architecture	RMSE	MAPE (%)	MDA (%)	R²
Exp1	None	None	None	LSTM	59.27	4.53	0.49	0.84
	None	None	None	CNN	152.99	13.41	0.54	-0.069
	None	None	None	CNN-LSTM	57.14	4.39	0.47	0.85
Exp2	Early	EN	LLaMA 3.1 (4-bit)	LSTM	60.02	4.65	0.47	0.84
	Early	EN	LLaMA 3.1 (4-bit)	CNN	105.90	9.21	0.51	0.49
	Early	EN	LLaMA 3.1 (4-bit)	CNN-LSTM	57.83	4.44	0.49	0.85
Exp3	Early	ALL	LLaMA 3.1 (4-bit)	LSTM	75.35	5.78	0.50	0.74
	Early	ALL	LLaMA 3.1 (4-bit)	CNN	173.29	15.95	0.51	-0.37
	Early	ALL	LLaMA 3.1 (4-bit)	CNN-LSTM	92.25	7.69	0.49	0.61
Exp4	Early	ALL	LLaMA 3.1 (2-bit)	CNN-LSTM	83.54	7.02	0.43	0.68
	Early	ALL	LLaMA 3.1 (4-bit)	CNN-LSTM	87.67	6.72	0.47	065
	Early	ALL	Orca 2 (4-bit)	CNN-LSTM	85.48	6.40	0.48	0.67

Table A.3 DOGE full experiment breakdown (Exp1 to Exp4)

ID	Fusion	Language	LLM	Architecture	RMSE	MAPE (%)	MDA (%)	R²
Exp1	None	None	None	LSTM	0.002672	118.50	0.49	0.88
	None	None	None	CNN	0.003008	103.25	0.44	0.85
	None	None	None	CNN-LSTM	0.002660	110.82	0.47	0.88
Exp2	Early	EN	LLaMA 3.1 (4-bit)	LSTM	0.002736	120.95	0.49	0.87
	Early	EN	LLaMA 3.1 (4-bit)	CNN	0.009063	859.47	0.48	-0.39
	Early	EN	LLaMA 3.1 (4-bit)	CNN-LSTM	0.002549	121.05	0.48	0.89
Exp3	Early	ALL	LLaMA 3.1 (4-bit)	LSTM	0.008840	469.16	0.43	-0.32
	Early	ALL	LLaMA 3.1 (4-bit)	CNN	0.007460	424.54	0.45	0.06
	Early	ALL	LLaMA 3.1 (4-bit)	CNN-LSTM	0.007828	389.76	0.45	-0.04
Exp4	Early	ALL	LLaMA 3.1 (2-bit)	CNN-LSTM	0.005955	304.36	0.44	0.40
	Early	ALL	LLaMA 3.1 (4-bit)	CNN-LSTM	0.005309	281.82	0.43	0.52
	Early	ALL	Orca 2 (4-bit)	CNN-LSTM	0.005843	277.36	0.43	0.42

Appendix B – Project Code

The project consisted of a vast amount of code, therefore, head to the project folder of the 'Vis4Sense' repository below to view and replicate the system:

https://github.com/Vis4Sense/student-projects