SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Sentiment-Informed Price Prediction in Algorithmic Trading: An Emoji-Based Approach

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Paolo Pietro Bucci

Advisor: Prof. Simone Formentin

Co-advisor: Prof. Daniele Mazzarina, Dr. Raffaele Giuseppe Cestari

Academic year: 2024-25

---

1. Introduction

To fully capture financial market dynamics has always been a complex challenge. Recently, this challenge has intensified due to the growing influence of digital communication on markets. Online platforms such as Twitter, Stocktwits, and Reddit now provide spaces where retail and institutional investors openly share opinions, expectations, and emotions that can strongly influence market behavior. Traditional forecasting approaches, which rely primarily on historical quantitative indicators such as returns, volumes, and volatility, remain essential but fail to capture these behavioral dimensions that often shape short-term price fluctuations. Recent advances in Natural Language Processing (NLP) have made it possible to systematically extract sentiment from textual contents, opening new opportunities for market prediction. The objective of this thesis is to investigate the contribution of sentiment information to financial time-series forecasting, with a particular focus on extracting sentiment from unstructured, colloquial and emoji-rich online content. To this end, we develop and evaluate sentiment classification models specifically adapted to financial texts enriched with emojis, which are increasingly common in online investor discussions. The extracted sentiment signals are then integrated into a price prediction model based on a Long Short-Term Memory (LSTM) architecture, a deep learning method well suited for time series forecasting. By comparing models trained with and without sentiment features, and by analyzing the impact of different levels of sentiment classification accuracy, the study aims to quantify the added value that sentiment analysis provides in predicting market movements.

2. Objectives

The thesis pursues two main objectives:
First we want to **classify market sentiment** from online unstructured content (e.g. news headlines, investing forums, Twitter posts) for a target financial asset. This leads to the necessity to design a sentiment classifier capable of extracting investor sentiment from social media content, with a particular **focus on short, emoji-rich and informal text**. Unlike traditional financial news, online investor discussions often rely on slang, irony, and emojis, which pose challenges for conventional NLP tools. To address this challenge, we will extend BERT-based models with enriched vocabularies and

Attention layers and train them to classify posts into three categories: **Bullish, Bearish, and Neutral**.

The second objective is to **use sentiment to predict financial returns**. By aggregating post's classification results across time windows, we compute a Sentiment Score, defined as the number of Bullish minus Bearish posts. This score acts as a synthetic indicator of overall market sentiment. Therefore, the Sentiment Score is integrated with classical financial indicators such as past prices and traded volumes for price forecasting. A LSTM network is employed to capture temporal dependencies in the data and predict future returns. The LSTM outputs serve as the basis for a trading pipeline, where upward predictions trigger long positions and downward predictions trigger short positions. This framework allows us to assess whether sentiment signals derived from social media improve predictive accuracy and can be translated into profitable trading decisions.

Together, these objectives establish an end-to-end pipeline: from extracting sentiment in emoji-rich online environments to incorporating it as a predictive feature in algorithmic trading strategies.

3. Literature Review

3.1. Natural Language Processing

Natural Language Processing techniques for sentiment analysis begin by transforming text into numerical representations, or vectors, that models can process. This first step is called tokenization: the foundational text pre-processing step, which has evolved from simple white space word-level split of text to more sophisticated sub-word methods such as Byte Pair Encoding and WordPiece to improve robustness, reduce vocabulary size, and handle rare or multilingual terms more effectively [11, 15]. The second step to turn text into vectorial representations is the encoding phase. Traditional encoding like Bag-of-Words and TF-IDF, often paired with models such as Support Vector Machines, represent documents as unordered collections of words, which fails to capture context, word order, and subtleties like negation or sarcasm. Word2Vec algorithm advanced the field by mapping words into dense embeddings that preserve semantic relationships, though these remain context-independent, giving the same representation to polysemous words like bank [9]. A breakthrough came with the attention mechanism, which enabled models to focus dynamically on relevant tokens and paved the way for the Transformer architecture [14]. Transformers rely on embeddings, positional encodings, multi-head self-attention, residual connections, and feed-forward networks to build contextualized embeddings that encode meaning, order, and inter-dependencies among words and can heavily exploit the advantages offered by parallel computing. Building on this foundation, BERT introduced a deep encoder-only model that generates rich contextual representations for tasks such as natural language inference, question answering, named entity recognition, and text classification [3]. BERT uses special tokens like [CLS] for sequence-level classification and [SEP] for separating paired inputs. Its training paradigm involves two phases: pre-training with masked language modeling and next sentence prediction to learn general-purpose language representations, followed by fine-tuning, where the model parameters are adjusted for specific downstream tasks such as sentiment classification. This shift to contextualized embeddings through Transformers has significantly advanced sentiment analysis beyond the limits of static or bag-of-words approaches.

3.2. Sentiment Analysis and Price Forecasting

Over the last decade, emojis have became a crucial component of online communication, particularly on platforms like Twitter and Instagram, where they enrich short texts and often determine sentiment. Their interpretation is essential, as they can reinforce or even invert the meaning of a message. Early approaches to emoji sentiment analysis relied on dictionaries [7] or embeddings such as Emoji2Vec, which learned representations from textual descriptions [5]. More recent methods leveraging deep learning models and in particular Transformers, which achieve superior performance in NLP tasks are limited by the absence of emojis in most pre-trained vocabularies [8]. In financial markets, sentiment analysis complements traditional approaches: fundamental analysis,

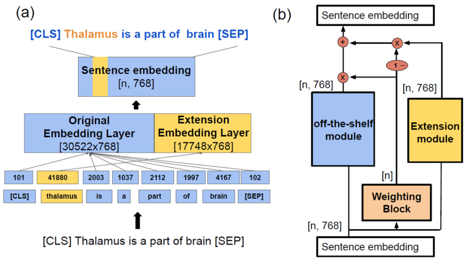


Figure 1: exBert model architecture [12]

which evaluates stocks using financial and economic data [4], and technical analysis, which focuses on statistical price and volume patterns [10]. A growing body of work shows that sentiment from news and social media influences returns and volatility. For example, Tetlock [13] demonstrated that pessimistic media tone depresses market prices, while Twitter-based indices such as TFSI provide predictive signals for daily stock returns [1]. Classical statistical models like ARIMA [2] are effective for stationary linear processes, but nonlinearity in financial data and temporal complexity demand more advanced tools. While machine learning methods such as SVR and Random Forests can capture nonlinearities, they still require handcrafted features. Deep neural networks, especially LSTMs [6], overcome these limitations by learning sequential dependencies directly, making them state-of-the-art models for forecasting in finance.

4. Model Design and Methods

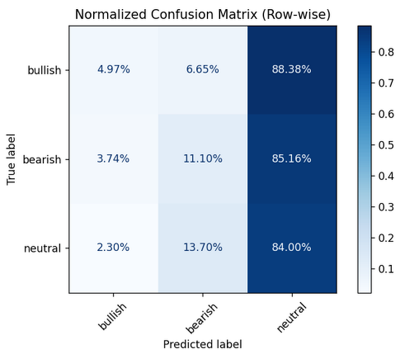
This thesis introduces **exBERT**, an extension methodology for adapting pre-trained BERT-based models to the domain of financial social media posts enriched with emojis. The approach enlarges the original model vocabulary with emojis which are used with financial domain-specific meaning, adds an additional embedding for new tokens and an extension module to each transformer layer. Outputs from the base and extension modules are combined via a learnable weighted sum, enabling the model to reuse pre-trained weights while learning specialized embeddings for new tokens. This allows efficient adaptation under constrained resources while retaining the knowledge of the original model.

To conduct experiments, two datasets were used. The **SPY StockTweets Dataset** contains 1.4M labeled posts about the SPY ETF spanning 498 business days (from 2020 to 2022), with more than 210k posts containing emojis. The **Stocktwits-Emoji Dataset**, covering BTC, ETH, and SHIB, provides 92k posts with balanced labels across bullish, bearish, and neutral sentiment. The latter also includes a held-out test set of 11,984 posts. Together, these datasets supply both large-scale, finance-specific data and emoji-rich balanced samples.

As off-the-shelf models two BERT-based models were selected: **FinBERT**, pretrained on financial text, and **Twitter-RoBERTa**, pretrained on 124M tweets. Both were extended with the exBERT framework, producing FinBERT-based exBERT and RoBERTa-based exBERT (ROexBERT). Pretraining was performed using a custom Masked Language Modeling task with emphasis on emoji tokens, while finetuning used under sampled balanced subsets to mitigate class imbalance. Progressive unfreezing was applied: first training only the new embedding layer, then progressively unfreezing classifier, pooler, and full network.

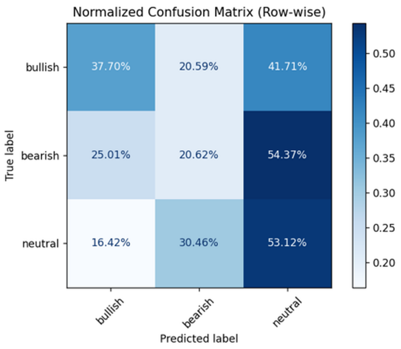
4.1. Evaluation and Results

Model evaluation relied on accuracy, precision, recall, and F1-score, supported by confusion matrices.



Metric	Value
Accuracy	0.2435
Precision	0.3845
Recall	0.3336
F1-score	0.2017

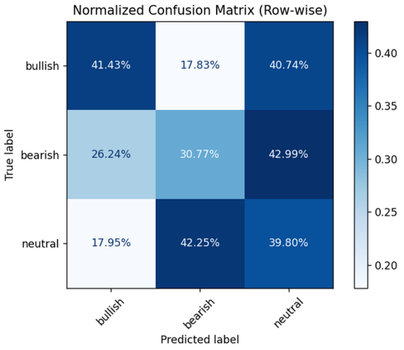
Figure 2: Confusion matrix and macro-averaged metrics for Finbert on the StockTweetsEmoji test set.



Metric	Value
Accuracy	0.3508
Precision	0.3748
Recall	0.3715
F1-score	0.3441

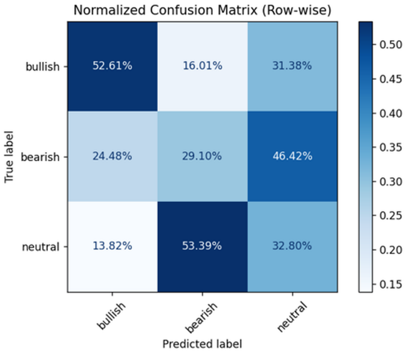
Figure 3: Confusion matrix and macro-averaged metrics for exBert FinBERT-based on the StockTweetsEmoji test set.

The first confusion matrix, corresponding to FinBERT, reveals the models strong bias toward predicting the Neutral class. This behavior results in low overall accuracy of about 24%, with the vast majority of predictions concentrated in a single class.The second confusion matrix, corresponding to exBERT a significant improvement of about 10 percentage points on accuracy and a more evenly distributed labeling.



Metric	Value
Accuracy	0.3734
Precision	0.3921
Recall	0.3733
F1-score	0.3670

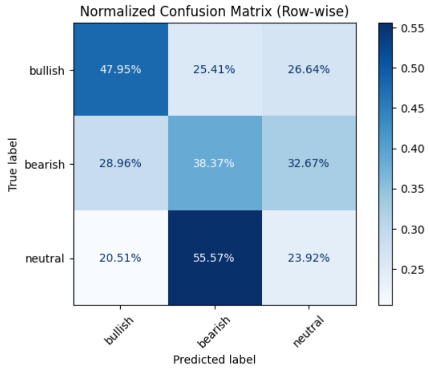
Figure 4: Confusion matrix and macro-averaged metrics for ROBertA on the Stocktweet-Emoji dataset.



Metric	Value
Accuracy	0.4505
Precision	0.4034
Recall	0.3817
F1-score	0.3831

Figure 5: Confusion matrix and macro-averaged metrics for exBERT ROBertA-Based on the Stocktweet-Emoji dataset.

RoBERTa as a baseline model proves to be originally stronger than FinBERT with an accuracy of 37% and its exteded version ROexBERT further improved results to 45% on the Stocktwits-Emoji test set.



Metric	Value
Accuracy	0.3935
Precision	0.3798
Recall	0.3675
F1-score	0.3706

Figure 6: Confusion matrix and macro-averaged metrics for finetuned Finbert (no extention module) on the Stocktweets-Emoji test set.

FinBERT finetuned without extension also improved around 39%, showing that while emoji-aware extensions are most useful in emoji-dense contexts, fine-tuning alone can also yield gains. Human annotators from Politecnico di Milano achieved 42% for the single vote classification and 49% accuracy on the majority voting classification, placing ROexBERT at near-human performance.

Summing up the results, the exBERT approach demonstrates that extending base models with domain-specific vocabulary substantially improves sentiment classification on financial social media, especially for emoji-rich posts. While fine-tuning helps, the extension module provides additional robustness where emojis are central to sentiment expression. Moreover, automated models reach performance comparable to human annotators, justifying their use in large-scale sentiment extraction pipelines.

5. Price Prediction Model

The objective of this chapter is to evaluate whether incorporating market sentiment into time-series forecasting can improve financial price prediction. Sentiment is extracted from social media posts and aggregated into a score defined as:

$$\text{Sentiment Score} = \#\text{Bullish} - \#\text{Bearish}.$$

which provides a simple and interpretable proxy for market sentiment. This novel feature is combined with two classical predictors:**returns and volumes**. Returns are used instead of raw prices because of their stationarity, while volumes measure trading intensity. Together, these three features constitute the input for the predictive model.

The forecasting framework is based on a deep recurrent architecture with three stacked bidirectional LSTM layers (256 - 256 - 384 hidden units), followed by dropout (0.1) and a dense ReLU output layer. The network takes as input sequences of length 15 and is trained with the Adam optimizer and MSE loss, while both MSE and MAE are tracked as evaluation metrics. Hyper-parameters are optimized with Bayesian methods in TensorFlow Keras. The model has more than 14 million parameters, reflecting its capacity to capture nonlinear dependencies, but also requiring careful regularization to avoid overfitting. To build the dataset used in the experiments, 3.5 million tweets about the S&P 500 index ($SPY ticker) were collected between April 2020 and March 2022. Sentiment labels were assigned using the models developed in the previous chapter, and aggregated over two different time frequencies: daily and hourly. Historical financial data (returns and volumes) were aligned with these aggregation windows to form synchronized time series with three features.

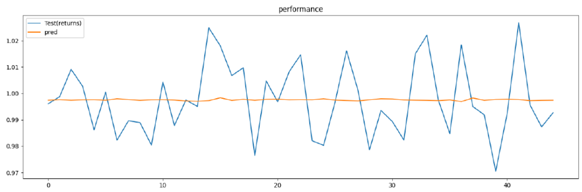


Figure 7: Daily returns predictions without sentiment score feature.

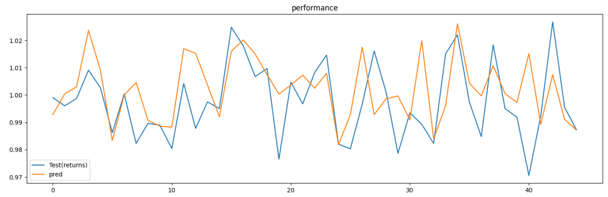


Figure 8: Daily returns predictions with sentiment score feature.

The experiments begin with daily frequency. As shown in Figure 7, the baseline LSTM trained only on returns and volumes collapses to predicting flat averages around 1.00, minimizing error but failing to capture short-term movements. By contrast, when the sentiment score is added, predictions better follow the actual returns and capture local dynamics as we can observe in Figure 8. The improvement is also evident in trading performance. A naive rule-based strategy, based only on past returns, was compared against an informed strategy using model predictions. For the selected trading window the sentiment-informed approach yields higher cumulative gains.

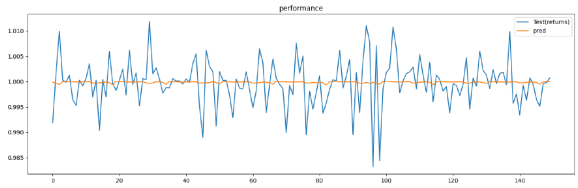


Figure 9: Hourly returns predictions without sentiment score feature.

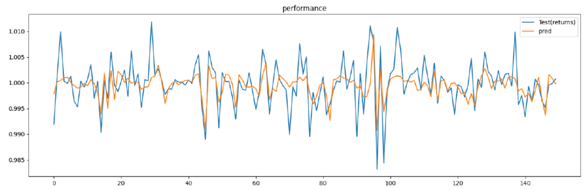


Figure 10: Hourly returns predictions with sentiment score feature.

When the same experiment is repeated at hourly frequency, the impact of sentiment becomes even stronger. Predictions with sentiment align more closely with intraday fluctuations, while daily aggregation tends to smooth out short-term variations. This confirms that sentiment contributes more significantly at higher frequencies, where investor mood shifts rapidly and directly influences market behavior. Such settings also highlight the advantage of automation, as algorithmic trading systems can exploit hourly signals more effectively than humans.

Table 1: Comparison of Models using MSE and MAE

Features	**MSE**	**MAE**
FinBERT	1.75533e-05	0.002964
RoBERTa-twitter	1.675195e-05	0.002830
FinBERT FT	1.481982e-05	0.002781
exBert	1.506633e-05	0.002764
ROexBERT	1.424594-05	0.002564

Finally, the quality of the sentiment classification models directly affects forecasting performance. Table 1 reports error metrics for models trained with sentiment scores derived from different classifiers. FinBERT performs worst, while RoBERTa-twitter and fine-tuned FinBERT improve results. The best-performing model over the classification task, ROexBERT, achieves the lowest error values when used to calculate the sentiment score feature (MSE 1.42×10^{-5}, MAE 0.0026), confirming that improved handling of emojis and finance-specific text leads to more accurate forecasts.

In summary, incorporating sentiment into the forecasting pipeline improves predictive accuracy and trading profitability. The results show that sentiment is especially valuable in high-frequency (hourly) settings and that the accuracy of upstream sentiment classificationparticularly with emoji-aware modelssubstantially enhances downstream price prediction performance.

6. Conclusions and Future Work

This thesis investigated the use of domain-adapted language models for sentiment analysis on financial market through social media post and its integration into price forecasting. The first contribution was the development of exBERT, an approach to extend pre-trained models with emoji-aware vocabulary. Results showed that this method significantly improves sentiment classification, with ROexBERT achieving the best performance and confirming the efficiency of vocabulary and model extension. The second contribution was the integration of sentiment scores into an LSTM forecasting model alongside classical financial features such as returns and volumes. Experiments demonstrated that sentiment improves predictive accuracy, particularly at the hourly level where market movements are strongly driven by short-term investor mood. Moreover, better classification models translated directly into more accurate forecasts, while trading simulations confirmed the practical value of sentiment-informed strategies, which outperformed naive baselines in cumulative gains.

Future work includes building a fully automated trading pipeline that processes posts, classifies sentiment, and generates signals in real time, while accounting for latency and transaction costs. Another direction is hybrid classification, where emoji-heavy posts are routed to extended models like ROexBERT, and text-heavy posts to lighter models, improving overall accuracy and forecasting quality.

In sum, this work validates vocabulary extension as a method to enhance sentiment analysis by leveraging emojis and non-verbal cues in finance and demonstrates the concrete benefits of incorporating sentiment into forecasting pipelines.

References

[1] Travis Adams, Andrea Ajello, Diego Silva, and Francisco Vazquez-Grande. More than words: Twitter chatter and financial market sentiment. *Finance and Economics Discussion Series*, (2022-009), 2022.

[2] George EP Box and Gwilym M Jenkins. *Time Series Analysis: Forecasting and Control.* Holden-Day, 1970.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[4] David Dodd and Benjamin Graham. *Security Analysis.* Whittlesey House, New York, 1934.

[5] Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. emoji2vec: Learning emoji representations from their description. *arXiv preprint arXiv:1609.08359*, 2016.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[7] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of emojis. *PLOS ONE*, 10(12):e0144296, 2015.

[8] Yao Lou, Yating Zhang, Fan Li, Ting Qian, and Donghong Ji. Emoji-based sentiment analysis using attention networks. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(5):1–13, 2020.

[9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*, 2013.

[10] John J. Murphy. *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications.* Penguin, 1999.

[11] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, 2016.

[12] Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. exbert: Extending pre-trained models with domain-specific vocabulary under constrained training resources. *arXiv preprint arXiv:2004.00018*, 2020.

[13] Paul C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[15] Yonghui Wu, Mike Schuster, Zhifeng Chen, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.