Refining Emoji Semantics for Financial Sentiment Analysis

by

Kaylee Xu

An honors thesis submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Science

Business and Economics Honors Program

NYU Shanghai

May 2024

Professor Marti G. Subrahmanyam — Professor Xi Chen

Professor Christina Wang

Professor Wendy Jin

Faculty Advisers Professor — Thesis Adviser

Table of Contents

Abstract

This study investigates the role of emoji-enhanced representations in financial sentiment analysis (FSA) through the application of classical machine learning models and large language models (LLMs). Drawing upon the StockEmotions dataset and an additional self-collected corpus of emoji-rich Twitter posts related to financial entities, this project explores how different representations of emojis—original, demoji-based, and refined semantic mappings—affect sentiment classification performance. This project proposes a refined version of emoji representation that contextualizes emoji meanings within the financial domain, improving interpretability for downstream tasks.

Three versions of each text instance are constructed: (1) the original version, containing preprocessed text with emoji; (2) the demoji version, where emojis are converted into literal textual descriptions according to the mapping list from Unicode; and (3) the refined version, where selected high-frequency emojis are replaced with domain-specific phrases informed by Emojipedia and manual annotation. These versions are evaluated using logistic regression, support vector machines (SVM), and FinBERT models, including a fine-tuned version.

The results show that incorporating refined emoji semantics leads to consistent improvements in classification performance, with the fine-tuned FinBERT model achieving the highest accuracy and F1 scores across all settings. Additionally, the study demonstrates the robustness of the refined representation when applied to an unseen Twitter dataset. These findings highlight the value of semantically meaningful emoji integration and pave the way for more expressive sentiment modeling in finance-focused NLP applications.

Key words: Large Language Models, Financial Sentiment Analysis, Emoji, NLP

Preface

As a double major in Data Science and Business & Finance at NYU Shanghai, I have long been fascinated by the intersection of machine learning and financial analysis. Thanks to the generous academic support at my beloved undergraduate institution, I have continuously sought to bridge the gap between technical modeling and practical applications in financial markets.

The rise of large language models (LLMs), particularly following the release of ChatGPT in 2022, inspired me to explore how these models can be leveraged in financial sentiment analysis (FSA). With the invaluable support of the Business and Economics Honors Program and the guidance of Professor Xi Chen at NYU Stern, I have had the opportunity to dedicate a year-long research effort to investigating how LLMs can enhance emoji understanding within financial texts on social media platforms.

This thesis paper, though not as professional as a mature researcher, represents my efforts of that journey—a reflection of both academic curiosity and a deep desire to bridge knowledge with real-world relevance.

I. Introduction

With the exponential growth of online financial discussions on platforms such as StockTwits, Reddit, and X (formerly Twitter), social media has become a rich and dynamic source of investor sentiment. Retail traders and institutional investors alike increasingly rely on crowd-sourced information to gauge market sentiment, which in turn affects market behavior in real time (Bollen et al., 2011; Chen et al., 2014). As a result, financial sentiment analysis (FSA)—the task of inferring investor attitudes such as bullishness, bearishness, or neutrality from text—has gained significant attention in the computational finance and natural language processing (NLP) communities.

A critical challenge in FSA is the accurate interpretation of non-standard textual elements such as emojis, which are frequently used in financial discussions to convey emotional nuance, conviction, or sarcasm. Prior studies often preprocessed emojis by simply removing them or replacing them with generic aliases using emoji libraries, resulting in a loss of potentially valuable sentiment signals (Felbo et al., 2017). Recent efforts have demonstrated that emojis, particularly in domain-specific contexts like finance, carry distinct semantic implications that can influence sentiment classification when adequately represented (Lee et al., 2020).

Meanwhile, the emergence of large language models (LLMs) such as BERT and its financial variant FinBERT (Araci, 2019) has dramatically improved the performance of NLP tasks by enabling better contextual understanding. However, even these models face limitations when emojis are reduced to superficial descriptors. To address this, this study proposes a refined emoji representation pipeline, which maps high-frequency emojis to semantically meaningful financial phrases through manual curation and domain-specific interpretation.

In this thesis, I examine how different emoji representations—the original, demoji, and refined version—impact the performance of various classification models, including traditional machine learning algorithms (e.g., logistic regression and SVM) and deep learning models such as FinBERT. By comparing performances across these representations on the StockEmotions dataset (Lee et al., 2020), it aims to shed light on the role of emoji semantics in improving FSA and suggest directions for more interpretable and context-aware sentiment modeling in financial applications.

II. Literature Review

The task of financial sentiment analysis (FSA) has attracted increasing attention in recent years, particularly due to the rise of retail trading communities and the proliferation of investor-generated content on social media. Prior studies have demonstrated that investor sentiment extracted from platforms such as Twitter and StockTwits can significantly influence market volatility and stock returns (Bollen et al., 2011; Chen et al., 2014). These findings have encouraged the development of machine learning systems capable of mining and interpreting user-generated texts to capture public market mood. Traditional approaches to FSA primarily rely on lexicon-based sentiment dictionaries or shallow machine learning models such as logistic regression or support vector machines trained on bag-of-words or TF-IDF features (Li et al., 2014). While such models can capture basic polarity, they often fall short when faced with informal or symbolic content like emojis, which are prevalent in online discourse. Emojis, though initially designed for casual communication, have been shown to carry sentiment-laden

information in domain-specific settings. For instance, the rocket emoji (🚀) is often used in financial communities to express confidence in upward price movement (Barbieri et al., 2018).

However, most sentiment analysis pipelines ignore emojis or treat them as noise due to the lack of semantically grounded representations. Recent work has attempted to bridge this gap by using emoji embeddings trained on massive corpora (Eisner et al., 2016), or by mapping emojis to descriptive phrases using resources like Emojipedia. Yet, such generic mappings frequently fall short of capturing the domain-specific semantics of emojis when used in finance. The StockEmotions dataset (Lee et al., 2020) is one of the first to systematically annotate emojis with fine-grained emotional and sentiment labels in a financial context, laying the groundwork for deeper exploration of emoji-enhanced sentiment models.

Concurrently, the advent of transformer-based language models such as BERT (Devlin et al., 2019) and its financial variant FinBERT (Araci, 2019) has revolutionized NLP by providing deep contextualized representations. FinBERT, pre-trained on large-scale financial texts, has achieved state-of-the-art performance in various sentiment classification tasks. Yet, despite their power, these models are often not equipped to interpret emojis semantically, unless explicitly trained or fine-tuned with emoji-aware data and representations.

In response, some recent studies propose refined emoji modeling techniques, in which high-frequency emojis are manually mapped to sentiment-rich, domain-relevant phrases before being fed into models (Han et al., 2023). This design can serve as a lightweight yet effective method to enhance model interpretability and performance, especially when fine-tuning large language models on relatively small but sentiment-dense datasets like StockEmotions.

III. Datasets and Preprocessing

This part will introduce the selected datasets and recent real data collected from social media. It will also go through the data preprocessing work including data cleaning, input representation design, and the descriptive statistics of the cleaned data used for model training.

3.1 Data Sources

This study utilizes two distinct yet complementary sources of financial text data.

StockEmotions Dataset (Lee et al., 2020) consists of 10,000 English comments collected from StockTwits, a financial social media platform. This dataset has two sentiment classes, namely bullish (positive) and bearish (negative), and 12 emotion classes, including amusement, anger, anxiety, etc., corresponding to 12 emojis. Among the 10000 data records, 80% are used as train data, 10% are used for validation, and 10% are for the test set. Within the train set, 4382 are labeled bullish and 3618 are labeled bearish, yielding a class ratio of approximately 1.2 to 1. This represents a relatively balanced distribution (See Figure 1). Thus, no additional class rebalancing was performed during training. The distribution of emotion labels is also fairly balanced (See Figure 2). Among the twelve predefined emotional categories, optimism (18.2%), anxiety (13.7%), and excitement (13.6%) are the most frequently occurring, collectively accounting for over 45% of the data. Emotions such as belief (9.1%), disgust (11.0%), and ambiguous (8.7%) occupy a secondary tier, while categories like panic, depression, and surprise are relatively rare, each representing less than 5% of the dataset. The presence of the ambiguous label highlights the complexity of interpreting emotional signals in financial texts, particularly when emojis carry multiple or context-specific meanings. Overall, the distribution is moderately balanced, which

helps mitigate issues of class imbalance during model training and enables more robust sentiment generalization across emotional states.

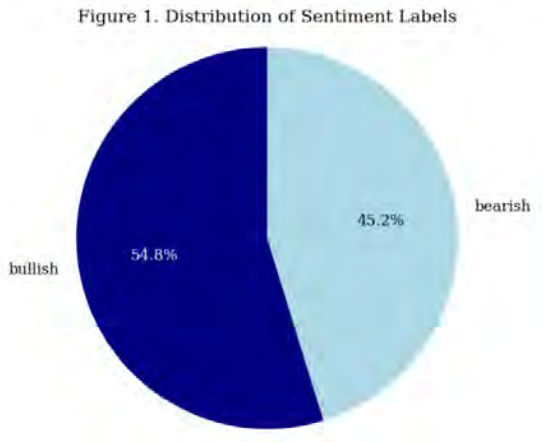


Figure 1

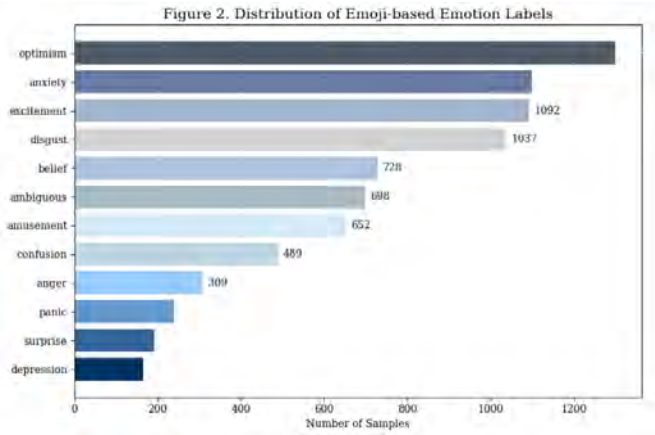


Figure 2

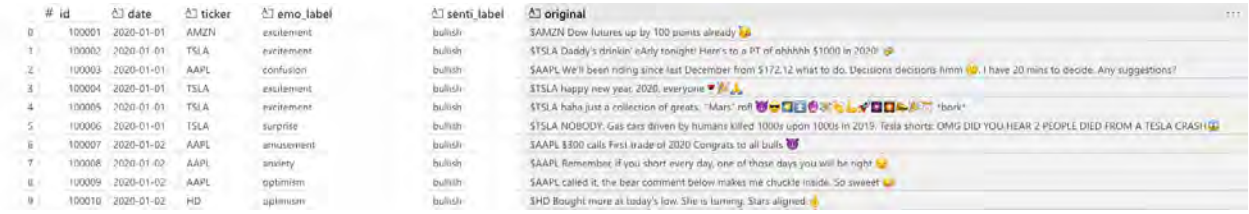
Besides the StockEmotions Dataset, the X2025 dataset provides most recent posts collected from X (formerly Twitter) since January to April 2025. This dataset is used for testifying the generalizability and robustness of the models trained by the StockEmotions Dataset.

3.2 Data Cleaning

In order to ensure textual consistency and remove noise from the user-generated content, especially within the Twitter dataset, I performed a series of standard normalization steps. While the StockEmotions dataset has undergone prior preprocessing by its original curators, the tweets collected for this study required additional cleaning due to their informal and unstructured nature. Specifically, I removed all hyperlinks and URLs, as they rarely carry sentiment-relevant information. Punctuation marks and non-alphanumeric characters (excluding emojis, which may convey affective cues) were filtered out or normalized. Moreover, superfluous line breaks introduced for stylistic purposes on social media platforms were eliminated to improve text

uniformity. These steps facilitate more effective downstream tokenization and model training. Unlike traditional NLP pipelines that discard emojis, I retained all emojis as essential elements for financial sentiment interpretation.

Here presents some records from the cleaned StockEmotions (Figure 3) and X2025 (Figure 4).



	id	date	ticker	emo_label	senti_label	original
0	100001	2020-01-01	AMZN	excitement	bullish	$AMZN Dow futures up by 100 points already
1	100002	2020-01-01	TSLA	excitement	bullish	$TSLA Daddy's drinkin' eArly tonight! Here's to a PT of ahhhhh $1000 in 2020!
2	100003	2020-01-01	AAPL	confusion	bullish	$AAPL We'll been riding since last December from $172.12 what to do. Decisions decisions hmm . I have 20 mins to decide. Any suggestions?
3	100004	2020-01-01	TSLA	excitement	bullish	$TSLA happy new year, 2020, everyone
4	100005	2020-01-01	TSLA	excitement	bullish	$TSLA haha just a collection of greats. "Mars" refi
5	100006	2020-01-01	TSLA	surprise	bullish	$TSLA NOBODY: Gas cars driven by humans killed 1000s upon 1000s in 2019. Tesla shorts: OMG DID YOU HEAR 2 PEOPLE DIED FROM A TESLA CRASH
6	100007	2020-01-02	AAPL	amusement	bullish	$AAPL $300 calls First trade of 2020 Congrats to all bulls
7	100008	2020-01-02	AAPL	anxiety	bullish	$AAPL Remember, if you short every day, one of those days you will be right
8	100009	2020-01-02	AAPL	optimism	bullish	$AAPL called it, the bear comment below makes me chuckle inside. So sweeet
9	100010	2020-01-02	HD	optimism	bullish	$HD Bought more at today's low. She is turning. Stars aligned.

Figure 3: Example Records Of StockEmotion Dataset

	emoji	text	post address	publish time	like_amt	comment_amt	retweet_amt
0		Sweet. Another device (that no one uses) for Apple Intelligence to suck on \n\n$AAPL	https://twitter.com/PeerA	2025-03-31 23:51:48+00:00	4	3	1
1		Ticker TOPP is a recent IPO with a nano float and no dilution.\n\n$NMAX went 800% today - fresh IPO and this is a direct	https://twitter.com/PeerA	2025-03-31 23:06:19+00:00	0	0	0
2		Ticker TOPP is a recent IPO with a nano float and no dilution.\n\n$NMAX went 800% today - fresh IPO and this is a direct	https://twitter.com/PeerA	2025-03-31 23:03:29+00:00	2	0	0
3		@patientinvestl That's nearly half a trillion in cash across 7 companies \n\nBuybacks, AI chips, M&Jump,A, dividends—la	https://twitter.com/PeerA	2025-03-31 22:58:33+00:00	2	2	0
4		After-hours Most Active Stocks by TradingView\n\n$DMN \| Damon \n$BIAF | bioAffinity Technologies \n$NVDA | NVI	https://twitter.com/PeerA	2025-03-31 22:35:35+00:00	4	0	1
5		MAG 7 Market Roundup: Closing Prices \n March 31, 2025\n\n- Meta $META: $576.36 (-0.07%)\n- Apple $AAPL: $	https://twitter.com/PeerA	2025-03-31 22:30:27+00:00	0	0	0
6		Warning:Whale alert\n\nTariffs may lift AI server costs 20%! Cloud giants (MSFT/GOOGL) capex risks. #DataCenters #Clc	https://twitter.com/PeerA	2025-03-31 22:30:00+00:00	4	0	0
7		$INTC bags $3.5B, defense chips on the scene \n $BA hit with strikes, 2008 vibes re-lean \n $BH demand hot, sal	https://twitter.com/PeerA	2025-03-31 22:05:35+00:00	3	0	0
8		So What for Retail Investors?\n Tariff-Exposed Stocks\n\nAvoid: Auto ($F, $GM), China-reliant tech ($NVDA, $AAP	https://twitter.com/PeerA	2025-03-31 22:04:34+00:00	0	1	0
9		Update on the top 5 stocks by market cap \n\n$AAPL Apple Inc. ($222.13, +1.90%) \n$MSFT Microsoft Corp. ($3	https://twitter.com/PeerA	2025-03-31 22:00:39+00:00	0	0	0

Figure 4: Example Records Of X2025 Dataset

3.3 Data Insights and Description

Based on the emoji-level sentiment distributions across the StockEmotions dataset and the manually selected emojis grouped by visual and symbolic themes (e.g., animals, red and green signs, finance-related icons), I present several insights that help contextualize the role of emojis in financial sentiment expression.

Across the dataset, I found that emoji usage is not sentiment-neutral: some emojis demonstrate strong polarity tendencies. For instance, the 🚀 emoji, with a bull ratio of 0.95, is heavily associated with bullish sentiment, consistent with its metaphorical usage to indicate soaring stock prices. Similarly, 💰 and 🤑 also skew bullish, reflecting investor optimism and enthusiasm

around profit and wealth. In contrast, 😭 and 💩 trend strongly bearish, indicating negative emotions such as frustration, regret, or disgust towards market performance (See Table 1). These findings confirm that financial users exploit emojis as compact sentiment signals, often bypassing explicit textual emotion.

Emoji	Bull Count	Bear Count	Bull Ratio	Bear Ratio	Total	Description
😂	929	1429	0.394	0.606	2358	face with tears of joy
🚀	1882	100	0.9495	0.0505	1982	rocket
🤣	416	804	0.341	0.659	1220	rolling on the floor laughing
💰	582	145	0.8006	0.1994	727	money bag
🤑	448	185	0.7077	0.2923	633	money-mouth face
🔥	298	176	0.6287	0.3713	474	fire
🐻	158	196	0.4463	0.5537	354	bear
🤔	144	148	0.4932	0.5068	292	thinking face
😎	181	74	0.7098	0.2902	255	smiling face with sunglasses
🤡	68	156	0.3036	0.6964	224	clown face
📈	210	4	0.9813	0.0187	214	chart increasing
👀	108	70	0.6067	0.3933	178	eyes
👍	108	55	0.6626	0.3374	163	thumbs up
😁	96	65	0.5963	0.4037	161	beaming face with smiling eyes
📉	2	156	0.0127	0.9873	158	chart decreasing
🙏	117	35	0.7697	0.2303	152	folded hands
😆	68	81	0.4564	0.5436	149	grinning squinting face
💩	24	124	0.1622	0.8378	148	pile of poo
💵	131	16	0.8912	0.1088	147	dollar banknote
😭	62	84	0.4247	0.5753	146	loudly crying

Table 1: Top 20 Emojis in StockEmotions

Animal emojis in financial texts reflect deeper symbolic connotations. For instance, 🐻 (bear) appears most frequently among animal emojis and aligns well with its metaphorical bearish meaning (Bull ratio ≈ 0.45). Meanwhile, 🐂, 🐷, and 🐃 (ox, cow variants) appear predominantly in bullish comments, reinforcing their symbolic role as bullish icons in financial vernacular. 🐖 and 🐑 are less frequent but tend to lean bearish, aligning with the derogatory connotation of "pig" or "sheep" in financial slang (See Table 2). Overall, the animal group of emojis is not merely decorative, but semantically aligned with common market metaphors used in trader communities.

Emoji	Bull Count	Bear Count	Bull Ratio	Bear Ratio	Total	Description
🐻	158	196	0.4463	0.5537	354	bear
🐖	18	25	0.4186	0.5814	43	pig
🐑	3	19	0.1364	0.8636	22	sheep
🐂	14	5	0.7368	0.2632	19	ox
🐃	9	0	1.0000	0	9	water buffalo
🐐	5	2	0.7143	0.2857	7	goat
🐷	4	1	0.8000	0.2000	5	cow face
🦍	4	0	1.0000	0	4	gorilla
🐨	0	3	0	1.0000	3	koala
🐱	1	1	0.5000	0.5000	2	cat face
🐵	2	0	1.0000	0	2	monkey face
🦊	2	0	1.0000	0	2	fox
🐄	0	1	0	1.0000	1	cow
🐕	0	1	0	1.0000	1	dog
🐩	1	0	1	0	1	poodle
🐈	1	0	1	0	1	cat

Table 2: Animal Emojis in StockEmotions

Color-coded emojis offer surprisingly sharp sentiment signals: Green emojis such as 💚 (green heart), 🟢 (green circle), and 🟩 (green square) show a 100% bull ratio, suggesting their role in expressing positive market performance or trader optimism. ❤️ (red heart) also show a high chance of bullish attitude of 0.71. Conversely, red-colored emojis such as 🔻 (downward red triangle) show a strong indication of bearish, reflecting their symbolic link with losses, danger, or negative movement in stock prices. Interestingly, ❗ (red exclamation mark) and ❓ (question mark) are more sentiment-neutral or even lean bearish, possibly indicating uncertainty or alertness in market predictions. This shows that visual semantics of emoji—like color, shape, and intensity—map onto financial sentiments in a way that's surprisingly structured.

Emoji	Bull Count	Bear Count	Bull Ratio	Bear Ratio	Total Count	Description
❤️	47	19	0.7121	0.2879	66	red heart
🔻	0	28	0	1.0000	28	red triangle pointed down
💚	19	0	1.0000	0	19	green heart
❓	9	0	1.0000	0	9	red question mark
🟢	6	1	0.8571	0.1429	7	green circle
❗	2	5	0.2857	0.7143	7	red exclamation mark
🔴	0	4	0	1.0000	4	red circle
💔	0	2	0	1.0000	2	broken red heart
🟩	1	0	1.0000	0	1	green square

Table 3: Red and Green Signs in StockEmotions

These findings validate the decision to preserve and refine emoji semantics in financial sentiment models. Rather than treating emojis as noise or trivial artifacts, they should be treated as important sentiment carriers with domain-specific polarity. Additionally, refined phrase

representations, grounded in contextual financial meaning (e.g., “stocks rising quickly in price” for 🚀), may further improve the interpretability and performance of sentiment models.

IV. Methodology

This part will describe the methodology in detail from the input representation engineering to model selection process.

4.1 Representation Engineering

To better encode the sentiment expressed by emojis, I performed a form of representation transformation, yielding three parallel textual variants: 1) original version, 2) demoji version, 3) designed version. Descriptions of the three versions are in Table 4.

No.	Version Name	Description
1	Original Version	The original unmodified text, with emojis retained in their Unicode form.
2	Demoji Version	Each emoji is mapped to its default Unicode alias using the `emoji.demojize()` function (e.g., 💰 → “money bag”). This representation attempts to expose emoji semantics to language models that may not natively interpret Unicode emojis.
3	Refined Version	Building upon the demoji version, a manually curated subset of emojis (selected from the top 30 most frequent ones) were mapped to financial domain-specific phrases based on contextual interpretation (e.g., 🚀 → "stocks rising quickly in price") with reference to Emojipedia (Figure 5).

Table 4: Descriptions of Three Text Variants

Figure 5: Example from Emojipedia

Table 5 shows some examples of how original emoji descriptions are refined to be closer to the financial contexts. Emojis not included in the refined dictionary fall back to their standard alias.

Emoji	Demoji Version	Refined Version
🚀	Rocket	stocks rising quickly in price
📈	Chart Increasing	increasing stock prices
🙏	Folded hands	praying

Table 5: Examples of Different Representations

These three versions provide an opportunity to test whether fine-grained semantic control over emoji representation improves the alignment between model input and human-perceived sentiment in financial contexts.

4.2 Modeling Approaches

To evaluate the effectiveness of emoji-informed textual representations in financial sentiment analysis (FSA), I experimented with both traditional machine learning models and state-of-the-art transformer-based language models. Specifically:

- **Logistic Regression** and **Support Vector Machine (SVM)** were employed as classical baselines. These models were trained using TF-IDF feature representations derived from various forms of the processed text.
- **FinBERT (base)**, a domain-specific BERT model pre-trained on financial texts by ProsusAI, was utilized for zero-shot inference without additional task-specific fine-tuning.
- A **fine-tuned variant of FinBERT** was further trained on the StockEmotions dataset to adapt the model to the informal, emoji-rich linguistic patterns prevalent in social media-based financial discourse.

This combination allows us to not only benchmark performance across architectures, but also to evaluate how well each model generalizes to different textual representations of the same financial message.

4.3 Fine-Tuning Setup

For the fine-tuned FinBERT experiments, I initialized from the pre-trained `ProsusAI/finbert` checkpoint and performed supervised training on the binary sentiment classification task (bullish vs. bearish). The neutral class present in the original dataset was merged into the bearish class to create a binary classification setting consistent with prior FinBERT implementations. During training, I used Huggingface's `Trainer` API with dynamic padding and gradient clipping to ensure stability and reproducibility.

V. Experiments and Results

5.1 Transformed Data Sample

To illustrate the effectiveness of our emoji transformation strategies, Figure 6 presents selected examples of the same text rendered under three different representation schemes: the Original Version, the Demoji Version, and the Refined Version.

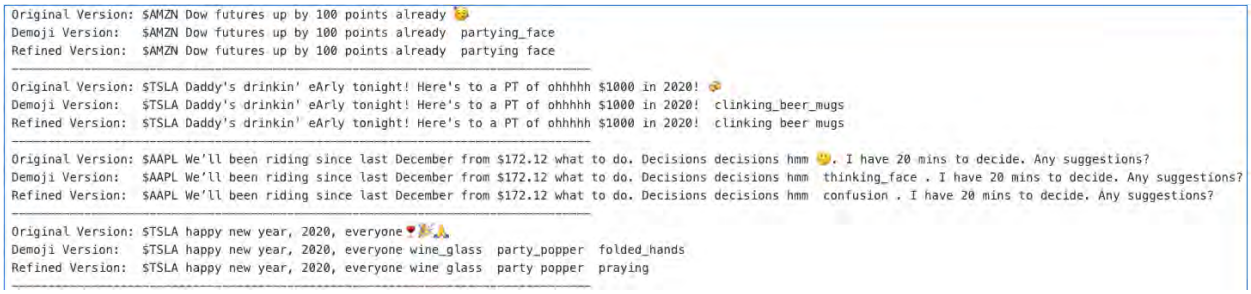


Figure 6: Sample Representations

- **The Original Version** preserves the raw text collected from social media platforms, including emojis and informal user expressions.
- **The Demoji Version** converts emojis into general descriptive phrases based on their Unicode alias using the `emoji` Python package. For instance, emojis such as 🥳 are translated into “partying face”.
- **The Refined Version** further enhances semantic clarity by replacing frequent emojis with financial-context-aware descriptions**,** derived from Emojipedia and refined manually. For example, 🤔 is mapped to “confusion” instead of “thinking_face”, a more contextually relevant phrase for financial sentiment analysis.

As shown, these transformations enable progressively richer semantic representations. While the Demoji Version provides a literal interpretation of emojis, the Refined Version bridges the gap

between surface-level form and sentiment-carrying content—crucial for downstream modeling. This layered transformation pipeline thus serves as the foundation for evaluating the influence of emoji representation strategies on model performance.

5.2 Model Performance

Based on the results from the logistic regression, SVM, and FinBERT models across three text input variants—Original, Demoji, and Refined—several performance trends emerge that provide valuable insights into the role of emoji representation in financial sentiment analysis.

1) Logistic Regression and SVM Models

Both traditional machine learning models, logistic regression (LOGREG) and support vector machine (SVM), demonstrated substantial sensitivity to different textual representations. LOGREG achieved the highest F1-score of 0.776 on the Demoji version, indicating that replacing emojis with their textual descriptions helped the model better capture the sentiment signals embedded in social media texts. Interestingly, the Refined version—which applies domain-specific interpretations to selected emojis—did not lead to further improvements, with F1 at 0.768, slightly lower than the Demoji version but still outperforming the Original version (F1 = 0.719). This suggests that while general description phrases already provide a significant boost, further fine-tuning of these representations yields diminishing returns under classical models.

A similar pattern is observed with SVM: the F1-score improves from 0.697 (Original) to 0.751 (Demoji), and remains stable at 0.764 under the Refined version. These results confirm the benefit of semantically enriching the input by converting emojis into meaningful text, especially for models that rely on explicit text features like TF-IDF.

Model	Version	Accuracy	F1_weighted	Precision	Recall
LOGREG	Original	0.719	0.71930904	0.71979215	0.719
LOGREG	Demoji	**0.775**	**0.77563957**	**0.77888164**	**0.775**
LOGREG	Refined	0.767	0.7676743	0.77118132	0.767
SVM	Original	0.696	0.69676343	0.69894922	0.696
SVM	Demoji	0.751	0.75161363	0.75366917	0.751
SVM	Refined	**0.764**	**0.76446046**	**0.76566058**	**0.764**

Table 6: Experiment Results of The FinBERT Base Model

2) FinBERT Base Model

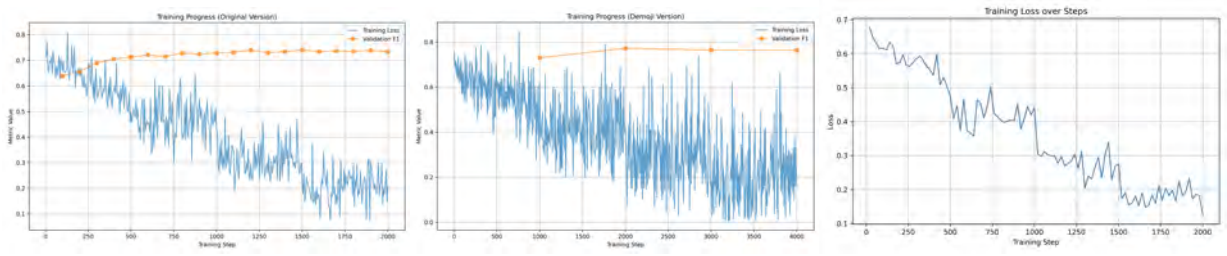
Unlike the classical models, FinBERT—though powerful—struggled more with the noisy and informal nature of financial social media texts. When evaluated without fine-tuning, its performance lagged behind LOGREG and SVM. For instance, the F1-score on the Original version was 0.378. Even though the metric increases to 0.465 on the Refined version, it still appears much poorer than the traditional model performance. The large gap in performance suggests that FinBERT's pretrained knowledge from formal financial texts may not directly transfer to emoji-rich, user-generated content. However, the improvement from Original to Refined reflects the utility of enhanced emoji phrasing even in deep learning contexts. With more targeted fine-tuning, further gains are likely achievable.

Model	Version	Accuracy	F1_weighted	Precision	Recall
FinBERT	Original	0.481	0.37836397	0.60389354	0.481
FinBERT	Demoji	0.461	0.33080711	0.57459826	0.461
FinBERT	Refined	**0.53**	**0.46482441**	**0.66081168**	**0.53**

Table 7: Experiment Results of The FinBERT Base Model

3) Fine-tuned FinBERT Model.

The fine-tuned FinBERT model demonstrates the strongest performance across all evaluated model architectures and input representations. As illustrated in the training and validation loss curves below, the model converges steadily over the training steps, with both training loss and evaluation loss decreasing consistently. This indicates effective learning and stable optimization during fine-tuning.



The performance metrics for FinBERT on the test set further validate its superiority. Among the three input versions, the Refined version achieves the highest overall accuracy of 0.797. It also delivers the best class-wise balance between precision and recall. Specifically, for the bullish class, the model attains a precision of 0.830 and an F1 score of 0.814, while for the bearish class,

it maintains strong performance with a precision of 0.760 and an F1 score of 0.777. These results underscore the added value of semantically enhanced emoji representations in financial sentiment understanding.

Model	Version	Accuracy	Class	Precision	Recall	F1_weighted
Fine-tuned FinBERT	Original	0.772	bearish	0.741648	0.748315	0.744966
			bullish	0.796733	0.790991	0.793852
Fine-tuned FinBERT	Demoji	0.795	bearish	0.770270	0.768539	0.769404
			bullish	0.814748	0.816216	0.815482
Fine-tuned FinBERT	Refined	**0.797**	bearish	**0.759657**	**0.795506**	**0.777168**
			bullish	**0.829588**	**0.798198**	**0.813590**

Table 8: Experiment Results of The FinBERT Finetuned Model

In sum, these results confirm that both fine-tuning on domain-specific data and enhancing emoji semantics with refined descriptions can significantly boost the performance of pre-trained language models in financial sentiment analysis.

5.3 Cross Dataset Validation

When evaluating the fine-tuned FinBERT models on the GPT-labeled Twitter dataset, I observe that the Refined Version, which previously outperformed all others on the original StockEmotions test set, achieved slightly lower performance compared to the Demoji Version.

While the overall results still demonstrate the efficacy of emoji-enhanced representations, this drop highlights important caveats in cross-dataset generalizability.

First, it is important to note that the GPT-generated labels in the Twitter dataset, though carefully crafted, have not undergone extensive human validation. The absence of annotation consensus or inter-annotator agreement introduces a potential source of noise that may affect model evaluation. Second, the fine-tuned models were optimized using hyperparameters specific to the StockEmotions dataset; as such, they may not be ideally tuned for the linguistic and distributional characteristics of Twitter texts. This further complicates direct performance comparisons across domains.

Lastly, the Refined Version is currently limited to a manually constructed set of 30 high-frequency emojis, which might not fully capture the nuanced and diverse emoji usage prevalent on Twitter. As a result, certain emoji-driven sentiment signals may be underrepresented or misinterpreted. Future work could address these challenges by expanding the refined emoji mapping through LLM-based contextual interpretation, validating labels with human annotation, and performing hyperparameter re-tuning for better domain adaptation.

Model	**Version**	**Accuracy**	**F1_weighted**	**Precision**	**Recall**
Fine-tuned FinBERT	Original	0.495	0.375	0.586	0.495
Fine-tuned FinBERT	Demoji	**0.543**	0.467	**0.659**	**0.543**
Fine-tuned FinBERT	Refined	0.539	**0.473**	0.624	0.539

Table 9: Experiment Results of The FinBERT Finetuned Model

VI. Discussion

The experimental results reveal several important findings regarding the effectiveness of emoji representation strategies and modeling choices for financial sentiment analysis. First and foremost, I observe that transforming raw textual inputs through emoji-aware representations leads to consistent improvements across all evaluated models. Both the traditional machine learning classifiers (Logistic Regression and SVM) and transformer-based models (pretrained and fine-tuned FinBERT) benefit from the inclusion of semantically meaningful emoji replacements. Compared to the original (clean) version, the Demoji version—where each emoji is substituted by its Unicode-based textual phrase—offers notable performance gains, suggesting that providing interpretable cues to models helps disambiguate the sentiment conveyed by emojis.

Further improvements are observed with the Refined version, in which only a curated set of high-frequency emojis are replaced with domain-specific financial phrases. For instance, transforming 🚀 into “stocks rising quickly in price” or 🐻 into “not promising” offers sentiment-aware signals grounded in financial discourse. This representation yields measurable performance gains: the Logistic Regression model’s F1-weighted score improves from 0.719 in the original version to 0.768 with the refined representation, while FinBERT’s F1 score rises from 0.378 to 0.465. These gains validate the hypothesis that incorporating domain-adapted emoji semantics significantly enhances model understanding of financial sentiment.

The best performance across all settings is achieved by the Fine-tuned FinBERT model, which attains an overall accuracy of 0.797 and balanced class-wise F1 scores when using the Refined version. In particular, the bullish class achieves a precision of 0.830 and F1 score of 0.814, while the bearish class achieves 0.760 and 0.777, respectively. These results demonstrate the synergy

between carefully crafted textual representations and powerful pre-trained language models when fine-tuned on task-specific data. Notably, the fine-tuned FinBERT not only outperforms all baseline models but also demonstrates robustness across all three input representations.

In summary, these results emphasize the importance of emoji representation as a critical component in financial sentiment pipelines. While traditional text preprocessing often discards emojis as noise, our findings suggest that preserving and refining emoji semantics—especially in a domain-specific manner—can unlock deeper sentiment insights. When coupled with strong modeling architectures like Fine-tuned FinBERT, representation engineering becomes a powerful tool for advancing performance in financial NLP tasks.

VII. Conclusion and Future Work

This study investigates the role of emojis in financial sentiment analysis (FSA), with a specific focus on refining emoji representation to improve model understanding within the financial context. By constructing three parallel textual representations—Original, Unicode Demojized, and Refined (Context-Aware Descriptions)—this research evaluates the impact of emoji semantics on model performance using both traditional machine learning models and a domain-specific large language model, FinBERT.

Experimental results demonstrate that converting emojis into descriptive phrases, particularly those that reflect their financial sentiment usage, can substantially improve classification accuracy and robustness. Among all tested models, the fine-tuned FinBERT achieves the best performance across all metrics, with the Refined representation yielding the highest F1-score and accuracy. This indicates that enhancing semantic alignment between emoji usage and financial interpretation significantly boosts model interpretability and effectiveness. These findings not

only reinforce the value of incorporating emojis in sentiment modeling but also highlight the importance of context-sensitive transformation over generic Unicode mappings.

However, the current refined emoji map is limited to 30 frequently occurring emojis. This restricted vocabulary may constrain the generalizability of the refined approach, especially when applied to more diverse or emerging financial discourse platforms such as Reddit or Discord. Additionally, the refined mapping was constructed manually, and thus may be sensitive to subjectivity or domain shifts.

Future work can address these limitations in several directions. First, expanding the refined emoji dictionary to include a broader and more representative set of emojis—potentially aided by frequency analysis and unsupervised clustering—will improve coverage and robustness. Second, leveraging resources such as Emojipedia sentiment tags, emoji embeddings, or context-driven alignment techniques can further enhance representation quality. Third, exploring the integration of multi-modal information (e.g., emoji image vectors, user behavior metadata) may offer deeper insights into investor emotion. Lastly, evaluating model generalization on multilingual datasets or across financial domains (e.g., cryptocurrency vs. equities) will broaden the applicability and relevance of this work.

Overall, this thesis highlights the importance of emoji-aware representation engineering in financial sentiment analysis, and paves the way for future research that bridges language models with emotional nuance in social finance.

References

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1–8.

Barbieri, F., Espinosa Anke, L., & Camacho-Collados, J. (2018). Multimodal Emoji Prediction. EMNLP.

Chen, H., De, P., Hu, Y., & Hwang, B. H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. Review of Financial Studies, 27(5), 1367–1403.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL.

Eisner, B., Rocktäschel, T., Augenstein, I., Bosnjak, M., & Riedel, S. (2016). emoji2vec: Learning emoji representations from their description. arXiv preprint arXiv:1609.08359.

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. EMNLP.

Han, S. C., Lee, J., Youn, H. L., & Poon, J. (2023). Understanding Emojis for Financial Sentiment Analysis. Proceedings of the 44th International Conference on Information Systems.

Lee, J., Youn, H. L., Poon, J., & Han, S. C. (2020). StockEmotions: Discover investor emotions for financial sentiment analysis and multivariate time series. arXiv preprint arXiv:2005.14595.

Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. Knowledge-Based Systems, 69, 14–23.

Acknowledgements

I would like to express my heartfelt gratitude to Professor Xi Chen, my thesis advisor, for his guidance and support throughout this year-long research journey. His insights and encouragement have been truly invaluable in shaping this work. I am also deeply thankful to Professor Christina Wang, Professor Ye Jin, and Professor Marti G. Subrahmanyam, our faculty advisors, for organizing numerous one-on-one meetings and offering thoughtful feedback along the way. My sincere appreciation goes to the invited seminar speakers, whose weekly talks have enriched my understanding of finance, economics, and marketing. Special thanks to our teaching assistant, Xinyi Yang, for her continuous support and dedication throughout the program.

I am grateful to my fellow Honors Program classmates for sharing this enriching journey, and to all the professors and friends I have met during my undergraduate years for their constant encouragement. Finally, I would like to express my deepest gratitude to my parents—for their unconditional love, patience, and unwavering support throughout both my academic and personal life.