Rising like a Phoenix: Emerging from the Pandemic and Reshaping Human Endeavors with Digital Technologies ICIS 2023

Social Media and Digital Collaboration

Dec 11th, 12:00 AM

# Understanding Emojis for Financial Sentiment Analysis

Siyi Chen
*National University of Singapore*, iris921426135@gmail.com

Frank Xing
*National University of Singapore*, xing@nus.edu.sg

# Understanding Emojis for Financial Sentiment Analysis

*Completed Research Paper*

**Siyi Chen**
National University of Singapore
Singapore, 117418
E0376986@u.nus.edu

**Frank Xing**
National University of Singapore
Singapore, 117418
xing@nus.edu.sg

## Abstract

*Social media content has been widely used for financial forecasting and sentiment analysis. However, emojis as a new "lingua franca" on social media are often omitted during standard data pre-processing processes, we thus speculate that they may carry additional useful information. In this research, we study the effect of emojis in facilitating financial sentiment analysis and explore the most effective way to handle them during model training. Experiments are conducted on two datasets from stock and crypto markets. Various machine learning models, deep learning models, and the state-of-the-art GPT-based model are used, and we compare their performances across different emoji encodings. Results show a consistent increase in model performances when emojis are converted to their descriptive phrases, and significant enhancements after refining the descriptive terms of the most important emojis before fitting them into the models. Our research shows that emojis are a valuable source for better understanding financial social media texts that cannot be omitted.*

**Keywords:** Emoji, Financial Sentiment Analysis, OpenAI, ChatGPT, Social Media, NLP

## Introduction

Social media platforms, such as Twitter, StockTwits, and Weibo, have been widely used as data resources for financial forecasting (Xing et al, 2020; Saha et al, 2022; Dong et al, 2022). This is because the market price of a stock is ultimately determined by the dynamics of bid and ask prices from the crowd, and market sentiment is one of the essential factors in driving supply and demand (Gao et al, 2022). Hence, being able to determine the underlying sentiments of the market helps investors and traders handle the problem of information overload and perform a better analysis of the price trends (Uhr et al, 2014). This motivates the study of financial sentiment analysis (FSA), which involves applying natural language processing (NLP) techniques to the financial corpus and classifying sentences into a few predefined sentiment categories, most of the time into positive and negative polarities (Khader et al, 2019; Du et al, 2023).

The emergence of various types of visual communication is remarkable on social media platforms. They have gained widespread popularity and significantly transformed the ways people communicate on social media (Qiu et al, 2023), especially among younger users. The introduction of the ASCII characters ":-)" in 1982 by Scott Fallman marks the inception of emoticons, and enabled people to convey emotions through text-based but non-verbal symbols. Currently, the range and category of visual expressions have considerably expanded to not only emoticons, but also emojis, ideograms, stickers (pictographs), and memes as illustrated in Figure 1 (Suntwal et al, 2021; Wang et al, 2021).

This paper primarily focuses on the effect of emojis, instead of other visual expressions, on financial sentiment analysis, as they represent the most structured visual expression type with a limited number of icons published by the designing organization. The first set of emojis (Figure 2 left side) traces back to Japanese interface designer Shigetaka Kurita, who created 176 unique graphic symbols for NTT DoCoMo devices in 1999. Nowadays, emojis are used by 92% of the world's online population in their communications

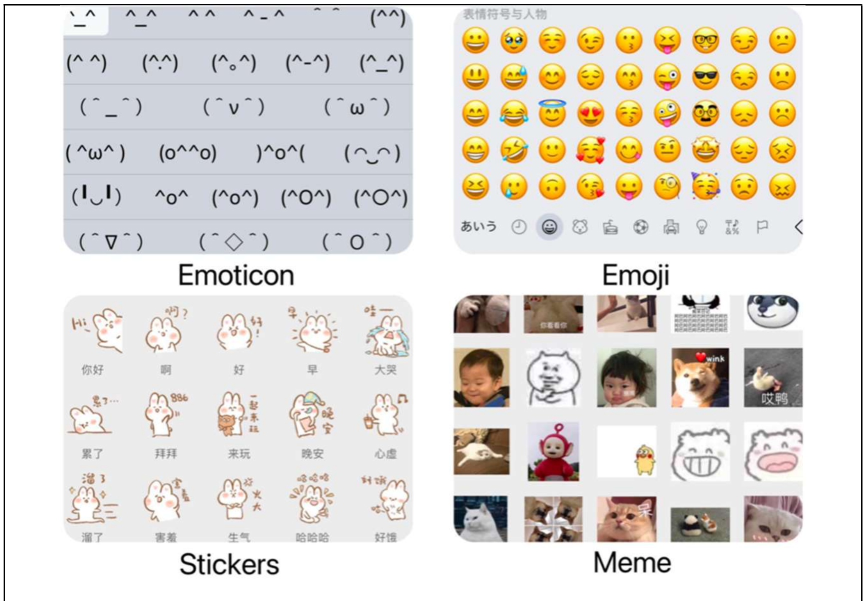


**Figure 1. Visual Expressions Recently Used in Social Media**

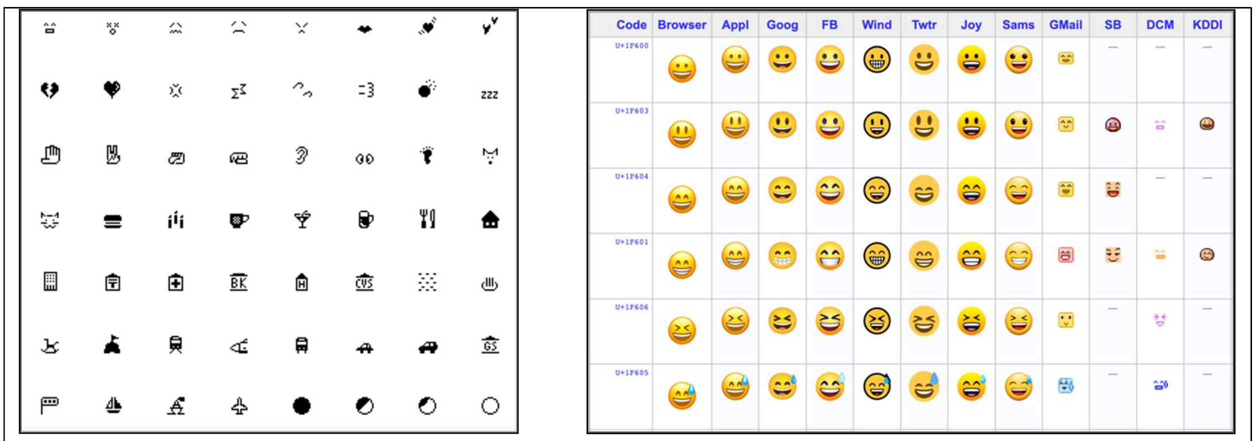


**Figure 2. The Evolution of Emojis (1999 to 2023)**

(The Unicode Consortium, 2022). The most recent Full Emoji List by Unicode (v15.0) contains 3,600+ emoji icons from different vendors such as Apple, Google, and Facebook (Figure 2 right side). According to Emojipedia Statistics in 2021, a famous emoji reference site, over one-fifth of the tweets now contain emojis (21.54%). Research has shown that using emojis may provide extra emotional or contextual significance to communication, adjust the tones of the sentences, as well as increase the message's appeal to recipients (Henriette et al, 2016; Hu et al, 2017; Qiu et al, 2023).

Despite the importance and prevalence of emojis used on social media, sentiment analysis does not always include this feature (Chen, 2023). For example, Deng et al (2018) used SentiStrength (Thelwall et al, 2010) to extract microblog sentiment, but the tool has only considered emoticons, not emojis. Xing et al (2020) analyzed the errors in financial sentiment analysis. However, during the data prepossessing and cleaning steps, only plain text information is reserved, while emojis are simply removed from the corpus. Recent evidence shows that emojis are useful for sentiment analysis on top of the text features (Singh et al, 2019; Yuan et al, 2022). We therefore hypothesize that they will also benefit financial sentiment analysis. Moreover, we are interested in whether emoji features are properly handled and their intrinsic meanings behind the language are fully captured, and the possible effect on financial sentiment analysis. In summary, we ask two research questions (RQs) as below: the first is descriptive and observational, the second is prescriptive and empirical.

**RQ1:** What are the usage patterns of emojis in short and informal financial texts on social media?

**RQ2:** What are the recommended practices for handling emoji features for financial sentiment analysis?

## Literature Review

### *Sentiment Effects of Emojis*

Emojis are gaining popularity in Internet communication as they may convey deep emotional connotations that are difficult to explain with words alone such as humor, sarcasm, and irony (Derks et al., 2008). They are more sensitive than text and able to boost the effectiveness of computer-mediated communication (Qiu et al, 2023). Here we are interested in how these sentiment effects can be revealed with text processing.

There are few standardized methods regarding the treatment of emojis. If keeping emojis in their original form inside the corpus, one common approach of machine learning models when dealing with this situation is to treat emojis as separate tokens, like words or punctuation marks, and to encode them as numerical values using techniques such as one-hot encoding or embedding. These numerical representations can then be fed into the model along with the text input, allowing the model to learn associations between specific emojis and sentiments. For example, a model might learn that the "👍" emoji is often associated with positive sentiment, while the "😠" emoji is often associated with negative sentiment. By incorporating these associations into the predictions, the model's accuracy in detecting sentiment in the text that includes emojis can be improved.

### *Feature Engineering of Emojis*

In existing NLP studies, "distributed embedding" is the main method used to handle emojis. According to the implementation details, they can be classified into three categories: skip-gram embedding, bi-sense emoji embedding, and word representation embedding.

Skip-gram is an unsupervised learning method used in Word2Vec. It is used to search for the most correlated words given a word. The implicit principle is that "a word is characterized by the company it keeps", which was popularized by English linguist John Firth in the 1950s. If the algorithm mis-predicts the context, it will alter its internal weights to produce a better forecast in the next round. Barbieri et al (2016) used several skip-gram word embedding models to map 10 million tweets to the same vectorial space and reduce the embeddings of 100 emojis to 2 dimensions with t-Distributed Stochastic Neighbor Embedding (t-SNE). Using the emoji vectors, the 5 most related tokens that can best describe every emoji are selected, and emojis of similar topics are grouped together using K-Means clustering algorithm (Figure 3). The effectiveness of this embedding method is then verified by a pair similarity and relatedness test, which is a standard approach in embedding evaluations (Mikolov et al, 2013).

Besides, to explore the meanings of the same emoji under different Twitter contexts, a scheme called bi-sense emoji embedding is proposed (Chen et al, 2018). Instead of embedding one emoji in one vector, they first assign two distinct tokens to each emoji and embed it into two sentences with opposite sentiments using fasttext, which is a more computationally efficient embedding method compared to Word2Vec. Both the positive-sense and negative-sense embeddings for each emoji are obtained and concatenated as one special word to feed into the LSTM network as BiE-LSTM. In this manner, the Bi-sense emoji embedding can represent more complex and fine-grained semantics combined with sentences and words and achieve higher performances with LSTM models.

To include more supervision to the embeddings, Hill et al (2016) constructed word representations for words and concepts based on their dictionary definitions to enhance the quality of single-word embeddings. Based on this study, Eisner et al (2016) suggest the meaning of the emojis can be represented by their surrounding words. In their implementation, for every sentence containing emojis, words except the emojis are separately embedded and aggregated to obtain the vector for the description vector for emojis. The embedding of a single word is done by the Word2Vec model, which is trained on Google News Dataset, containing 300-dimensional vectors for 3 million words and phrases. After all emoji embeddings are generated, they are combined with the original dataset for model training. The embedding performance is verified through both intrinsic and extrinsic tasks and projected down to a 2-dimensional space using the t-SNE technique (Figure 3).

| Emoji | Most similar text tokens |
|---|---|
| 😘 | love, babe, youu, awww, bby |
| 😋 | mmm, craving, pancakes, yummmmm, nutella |
| 😳 | dying, wtf, nooo, noo, lmaooo |
| 😭 | mlrt, omfg, blum, lmaoooo, crying ugh, |
| 😢 | sad, stomach, miss, noooooo coffee, |
| ☕ | roasters, caffeine, latte, redeye |
| ⚽ | soccer, futbol, regionals, mckale, muskingum |
| 🗽 | statue, nyc, rockerfeller, barneys, hk snowing, |
| ❄️ | snow, icy, brrrrr, outerwear |
| 👭 | sistas, ily, bunches, sista, ilysm |

| Cluster | Topic |
|---|---|
| 1 | Sports and animals |
| 2 | Nature |
| 3 | Body gestures and positive |
| 4 | Free time |
| 5 | Unclear |
| 6 | Love and parties |
| 7 | Letters |
| 8 | Barber and simbols |
| 9 | Eating and drinking |
| 10 | Music |
| 11 | Sad and tears |

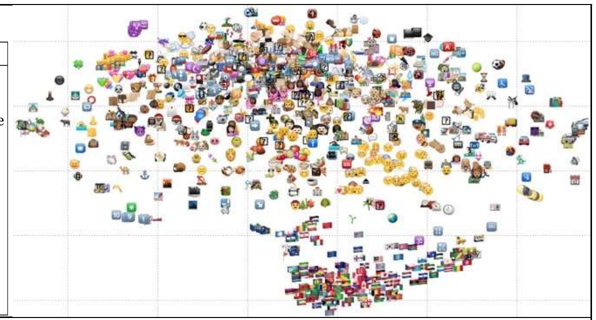

**Figure 3. Emoji Clusters & Embeddings (Barbieri et al, 2016; Eisner et al, 2016)**

### *Emojis in Financial Sentiment Analysis*

The abovementioned studies all focused on the impacts of emojis in sentiment analysis on the general domain corpus. For sentiment analysis in the financial domain, Xing et al (2020) revealed that it becomes a more challenging task due to the lack of labeled training datasets and domain adaptation. The study subsequently presented error analyses, concluding that sentiment analysis techniques are prone to fail for six types of reasons, such as irrealis moods, rhetoric, dependent opinion, etc.

When we dive into the StockSen dataset (Xing et al, 2020), it appears that many emojis in byte codes are not been converted correctly. In the study's original preprocessing steps, these codes are simply deleted from the corpus before fitting into the machine learning models. We found that in financial sentiment analysis, emojis are not as aware and exploited as in the general domain (also evidenced by the Deng et al (2018) study). This research fills this gap by presenting results to answer the two proposed research questions.

## Data and Method

### *Datasets, Preprocessing, and Descriptive Statistics*

Two datasets, i.e., the StockSen dataset (Xing et al, 2020) and the CMC dataset[1], are used in this research.

StockSen is a collection of 20675 financial tweets on the StockTwits platform spanning from June to August 2019. Either a bullish or bearish tag is chosen and self-labelled when posted by the commenter, which both reduces the need for skilled labor in the labeling process and ensures the dataset's quality. Among the 20675 data records, 15100 targets are bullish, while only 5575 targets are bearish. The ratio between the two classes is around 3:1, which indicates the imbalanced nature of the dataset. The mean review length in the corpus is 13 words, indicating that the reviews are generally concise. Nonetheless, the raw dataset contains several types of noise that impede readers from comprehending and analyzing the underlying sentiments of the corpus. These noises can be classified into four categories: (1) improper encodings regarding emojis, e.g., "\\xf0\\x9f\\x92\\xaa", (2) character entity encodings, e.g., "$amp;" for "&" and ">" for ">", (3) ticker codes, e.g., "$AMZN" for the Amazon company, and (4) URL to other webpages.

For these four types of noises above, improper encodings are converted back to the original emoji icons at first. Ticker codes are then removed as they just represent a specific stock and do not contain the sentiment of users. Character entity encodings are converted back to the raw characters. Finally, reviews containing URL links are removed with the assumption that the information inside the URLs is important to study the sentiment. After pre-possessing, 18470 (from 20675) data entries are left, out of which only 2581 pieces contain emojis. Most importantly, all the emojis icons in the dataset are converted successfully. To understand the prevalence of users using emojis on the StockTwits platform, a new column called 'emoji_count' is created to count the number of emojis within each text. The average number of emojis used in one piece of comment in the StockSen dataset is 0.45. Among the 2581 comments that contain at least one emoji and account for 13.97% of the entire dataset, the statistics of emoji count is shown in Table 1.

[1] Available from the authors upon reasonable request and for research purposes.

| **Emoji Count** | 1 | 2 | 3 | 4 | ≥ 5 |
|---|---|---|---|---|---|
| **Number of Reviews** | 1105 | 495 | 404 | 181 | 396 |

**Table 1. Statistics of Emoji Count**

With this small training sample size (2581), machine-learning models may have not sufficiently learned the sentiments behind these emojis. Hence, it is necessary to collect more financial comments datasets containing emojis to conduct comparative experiments and analyze the effects of emojis on the financial sentiment analysis performances.

We further collect data from the crypto community, where the demographic is younger and emoji use is more frequent. Unlike StockSen, which is gathered from StockTwits (a stock marketplace platform), the new cryptocurrency reviews dataset (CMC) is extracted using the Python 'request' package from the CoinMarketCap website, containing 13545 records in total. During the collection process, only records with both self-labeling and usage of emojis are included in the dataset. These reviews cover the period from February 2022 to December 2022 and covers all the reviews of the major coins.

Comparable to StockSen, CMC reviews have a slightly larger mean length of 17 words. For this kind of short informal text, the NLP literature suggests emojis to be important sentiment feature. We then test the balance of the dataset by calculating the ratio between the number of bullish and bearish reviews. The ratio is almost 8:1, which suggests that CMC is more skewed than StockSen. Because of this imbalance, the sentiment analysis accuracy baseline should be set to 83%, and some techniques such as over/undersampling methods, or SMOTE classifier are necessary during the model training step. With the same assumption that the content of the URL webpage will be useful in obtaining the final sentiment, we deleted all 783 data entries with URLs directly from the dataset, accounting for 5.78% of the entire dataset. With that, we are left with 12762 data entries inside the dataset.

## *Data Insights*

Figure 4 shows some records from the StockSen and CMC datasets. Because CMC contains many more emojis, we primarily focus on CMC as a representative for the use of emoji on financial social media. We analyze emoji polarities (bullish/bearish ratios) with specific focus on animal icons, colors, shapes, etc.

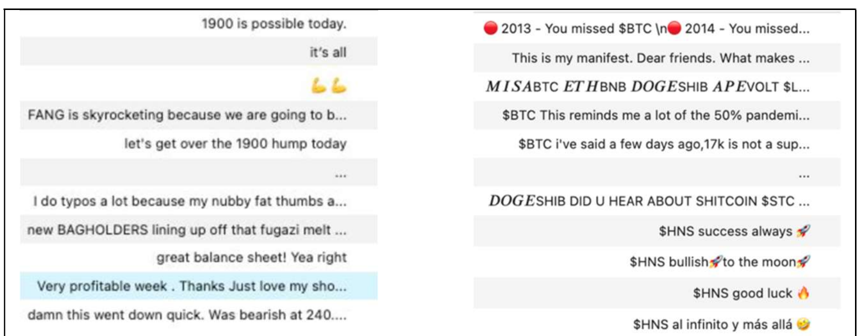


**Figure 4. Records from the StockSen and CMC datasets**

We produce a full list of unique emojis used in CMC and their associated counts are also consolidated into a summary table for manual investigation and analysis. In accordance with the StockSen, the Python emoji library is used to convert each emoji to its descriptive phrase in a new column. Additionally, the likelihood that each emoji appearing in bullish and bearish comments is recorded as 'bull_ratio' and 'bearish_ratio'. To search for the most frequently used emojis in bullish comments, we sort the file by 'total_count' and 'bullish_ratio' in descending order. Table 2 lists the 20 most frequent emojis in the CMC corpus.

| emoji | bull_count | bear_count | bull_ratio | bear_ratio | total_count | description |
|---|---|---|---|---|---|---|
| 🚀 | 15964 | 57 | 0.99644217 | 0.00355783 | 16021 | rocket |
| 🔥 | 5760 | 39 | 0.993274703 | 0.006725297 | 5799 | fire |
| 🟥 | 772 | 1779 | 0.302626421 | 0.697373579 | 2551 | red square |

| | | | | | | |
|---|---|---|---|---|---|---|
| ⬜ | 764 | 788 | 0.492268041 | 0.507731959 | 1552 | white large square |
| 🤣 | 433 | 757 | 0.363865546 | 0.636134454 | 1190 | rolling on the floor laughing |
| 💎 | 1069 | 7 | 0.993494424 | 0.006505576 | 1076 | gem stone |
| ⬛ | 507 | 392 | 0.563959956 | 0.436040044 | 899 | black large square |
| 🤑 | 771 | 49 | 0.940243902 | 0.059756098 | 820 | money-mouth face |
| 💪 | 771 | 7 | 0.991002571 | 0.008997429 | 778 | flexed biceps |
| 😂 | 365 | 358 | 0.504840941 | 0.495159059 | 723 | face with tears of joy |
| 💥 | 700 | 6 | 0.991501416 | 0.008498584 | 706 | collision |
| ✅ | 630 | 56 | 0.918367347 | 0.081632653 | 686 | check mark button |
| 💰 | 643 | 17 | 0.974242424 | 0.025757576 | 660 | money bag |
| 💯 | 618 | 11 | 0.982511924 | 0.017488076 | 629 | hundred points |
| 🤡 | 140 | 445 | 0.239316239 | 0.760683761 | 585 | clown face |
| ⚡ | 552 | 12 | 0.978723404 | 0.021276596 | 564 | high voltage |
| ❤️ | 531 | 10 | 0.981515712 | 0.018484288 | 541 | red heart |
| 👍 | 509 | 26 | 0.951401869 | 0.048598131 | 535 | thumbs up |
| 💸 | 400 | 15 | 0.963855422 | 0.036144578 | 415 | money with wings |
| 🥇 | 401 | 2 | 0.995037221 | 0.004962779 | 403 | 1st place medal |

**Table 2. Top 20 Emojis in CMC Corpus**

Out of the top 20 emojis, 14 are predominantly bullish, with a bull_ratio exceeding 0.9. Three emojis are neutral, with a bull_ratio hovering around 0.5, and three emojis lean more towards being bearish. Analyzing the emoji CSV file produced from the crypto comments further leads to some intriguing findings. To begin with, since "bullish" and "bearish" are derived from the metaphors of two animals, bull and bear, we initially extracted all the relevant emojis pertaining to these animals.

| emoji | bull_count | bear_count | bull_ratio | bear_ratio | total_count | description |
|---|---|---|---|---|---|---|
| 🐂 | 107 | 2 | 0.981651376 | 0.018348624 | 109 | ox |
| 🐮 | 7 | 0 | 1.0 | 0.0 | 7 | cow face |
| 🐻 | 19 | 17 | 0.527777778 | 0.472222222 | 36 | bear |
| 🧸 | 2 | 2 | 0.5 | 0.5 | 4 | teddy bear |
| 🐕 | 251 | 1 | 0.996031746 | 0.003968254 | 252 | dog |

**Table 3. Animal Emojis in CMC Corpus**

It was observed that icons related to bulls, such as "🐂" and "🐮", primarily appeared in bullish comments, whereas icons related to bears, such as "🐻" and "🧸", did not carry strong implications with bearish comments. Furthermore, another animal emoji, "🐕", was frequently used in bullish comments, with a usage rate of 99%. This may due to the fact that dog symbols are associated with a few Meme coins, namely $DOGE and $SHIB, which have experienced a surge in popularity following Elon Musk's tweets endorsing the potential of Dogecoin in 2022.

In term of colors (Table 4), it is surprising to observe that all the green emoji icons have a bull_ratio of 1, meaning that they never appear in bearish comments. The reason behind this is that in most cryptocurrency charts, a green candle represents a bullish move or a rise in price, while a red candle represents a bearish move or a fall in price. However, red emoji icons have more complexity and cannot be used to distinguish bullish or bearish comments directly. It is observed that red icons like "🔴", and "❗" are used more often in bullish comments. Although "🟥 " is used more often in bearish comments, it is also considered as a common emoji in bullish comments with 772 appearances.

Besides the relationship between the most important colors in crypto (green/red) and the sentiments, certain shapes are also closely related to sentiments. Take the example of heart shape in Table 5, it is observed that all emojis containing heart elements are used above 91% in bullish comments compared with bearish ones except the broken heat icon (“💔”).

| emoji | bull_count | bear_count | bull_ratio | bear_ratio | total_count | description |
|---|---|---|---|---|---|---|
| 🟢 | 298 | 0 | 1.0 | 0.0 | 298 | green circle |
| 🟩 | 78 | 0 | 1.0 | 0.0 | 78 | green square |
| 💚 | 373 | 0 | 1.0 | 0.0 | 373 | green heart |
| 🔴 | 249 | 91 | 0.732352941 | 0.267647059 | 340 | red circle |
| 🟥 | 772 | 1779 | 0.302626421 | 0.697373579 | 2551 | red square |
| 🔻 | 12 | 50 | 0.193548387 | 0.806451613 | 62 | red triangle pointed down |
| 🔺 | 3 | 0 | 1.0 | 0.0 | 3 | red triangle pointed up |
| ❗ | 26 | 2 | 0.928571429 | 0.071428571 | 28 | red exclamation mark |
| ❓ | 7 | 2 | 0.777777778 | 0.222222222 | 9 | red question mark |
| ⭕ | 3 | 0 | 1.0 | 0.0 | 3 | hollow red circle |

**Table 4. Green/ Red Emojis in CMC Corpus**

| Emoji | Crypto Dataset (CMC) | | | Stock Dataset (StockSen) | | | Description |
|---|---|---|---|---|---|---|---|
| | bull_count | bear_count | bull_ratio | bull_count | bear_count | bull_ratio | |
| ❤️ | 531 | 10 | 0.981515712 | 55 | 3 | 0.948275862 | red heart |
| 💚 | 373 | 0 | 1.0 | 9 | 0 | 1 | green heart |
| 😍 | 276 | 4 | 0.985714286 | 36 | 1 | 0.972972973 | smiling face with heart-eyes |
| 🥰 | 117 | 1 | 0.991525424 | 6 | 0 | 1 | smiling face with hearts |
| ♥️ | 93 | 3 | 0.96875 | 3 | 0 | 1 | heart suit |
| 💜 | 54 | 5 | 0.915254237 | - | - | - | purple heart |
| 💙 | 31 | 1 | 0.96875 | 6 | 0 | 1 | blue heart |
| 🫶 | 31 | 0 | 1.0 | - | - | - | heart hands |
| 💗 | 24 | 0 | 1.0 | 1 | 0 | 1 | growing heart |
| 💛 | 21 | 1 | 0.954545455 | - | - | - | yellow heart |
| 💕 | 20 | 0 | 1.0 | 9 | 0 | 1 | two hearts |
| 🧡 | 13 | 1 | 0.928571429 | - | - | - | orange heart |
| 💔 | 4 | 10 | 0.285714286 | - | - | - | broken heart |
| 💖 | 12 | 0 | 1.0 | - | - | - | sparkling heart |

**Table 5. Heart-shaped Emojis and Their Usage**

## *Emoji Phrasing*

Although the emoji library used to convert emojis to their corresponding phrases in our experiments has already provided a mapping list from Unicode to their corresponding aliases, they are only general descriptions of the visual image. When it comes to the emotional level, these descriptions are not accurate enough to represent the sentiment behind them. This is understandable since the emoji library is built as a tool to facilitate the usage of emojis instead of understanding their meanings. For example, as we can see in Table 2, the most frequently used emoji is "🚀" and it is transcribed to the description "rocket". While according to Emojipedia, it is used to indicate a fast increase in the stock price in the stock context. Hence, the description can be refined to "rising quickly in price" for sentiment analysis purposes.

Due to this reason, a refined emoji dictionary is created to enhance the relevance of the word representation. Manual checks are performed depending on the financial context of sentences that these emojis always appear. For the portion of the emojis selected to be modified, their refined phrases are referenced from their sentimental part of descriptions on Emojipedia, and their polarity indicated by the bull_ratio inside the emojis summary table (Table 2). The descriptive phrases of 30 emojis are refined during this data preparation step.

Table 6 provides some refining examples. With this approach, we obtained another "Refined Version" dataset, where all the emojis in the original dataset are replaced by the refined descriptive phrases.

| Emoji | Original Phrases | Refined Phrases | Emoji | Original Phrases | Refined Phrases |
|---|---|---|---|---|---|
| 🚀 | rocket | rising quickly in price | 💎 | gem stone | preciousness |
| 💥 | collision | excellent and exciting | ⚡ | high voltage | exciting |
| 🤡 | clown face | foolish | 😎 | smiling face with sunglasses | cool |

**Table 6. Refining Descriptive Phrases for Emojis**

## *Model Training*

The sentiment analysis function may be implemented using various model settings. To have a broad coverage and robust results, three types of settings, i.e., (1) six basic machine learning models, (2) deep learning models, and (3) the state-of-the-art GPT-based model are used as testbeds to observe the effects of different emoji phrasing.

### Basic Machine Learning Models

On this stage, we performed feature engineering and model training on both datasets. To begin with, we split the dataset into training and testing sets using the default ratio of 3:1. To streamline the entire process, we used the Scikit-learn pipeline. Firstly, features are generated using Count Vectorizer and TF-IDF Transformer with ngram_range from 1 to 3. These vectorizers tokenize the text data and convert it into a matrix of token counts containing unigrams (e.g. "apple" and "today"), bigrams (e.g. "red_apple"), and trigrams (e.g. "two_red_apples"), which are then turned into term frequency – inverse document frequency (TF-IDF) representations. Secondly, given the imbalanced nature of both datasets, we employed Synthetic Minority Oversampling Technique (SMOTE) to perform oversampling for the bearish samples in the pipeline. Finally, we fitted the training data to a model classifier, e.g., Multinomial Naive Bayes, to perform sentiment analysis. The six models experimented are: Logistic Regression (LR), Multinomial Naive Bayes (MNB), Support Vector Machines (SVM), CART Tree, Random Forest (RF), and Gradient Boosting (GB). Table 7 provides a brief description to these models.

Models are trained on distinct versions of datasets that had slight variations in the treatment of emojis. In our experiment, we prepared four versions of the datasets. The first version, referred to as the "Baseline Version", removes all emojis from the dataset directly. The second version, the "Emojis Version", retains all emoji icons after their conversion from byte code. The third version, known as the "Phrases Version", utilized the .demojize() function within the Python emoji package to convert emojis into descriptive texts. The final version, referred to as the "Refined Version", uses our modified descriptive texts for sentiment analysis purposes.

### Deep Learning Models

Besides the basic machine learning models, deep learning models such as LSTM (Long Short-Term Memory) and BERT (Bidirectional Encoder Representations from Transformers) are also leveraged to test the different treatments on the emoji datasets.

LSTM is a type of recurrent neural network (RNN) that can learn and retain information over extended time intervals. It is particularly adept at addressing the issue of long-term dependencies in sequence prediction problems, where the output of the network depends not only on the current input but also on past inputs in the sequence. BERT is another recent transformer-based deep learning model often used for natural language processing. The BERT model has already learned about the links between words, sentence structure, and other linguistic aspects since it has been trained on a vast corpus of text data. Because of this pre-training, it can perform well on newly encountered text input with less training samples, making it a powerful model for many NLP tasks.

When fitting the Emoji datasets on both the LSTM and BERT models, we use the standard training losses and optimizers with 10 epochs.

| Model Name | Category | Brief Introduction |
|---|---|---|
| **Logistic Regression (LR)** | Statistical Model | A special case of linear classification algorithm by applying a sigmoid function to map the response variable domain to [0,1] and generate a binary outcome. |
| **Multinomial Naive Bayes (MNB)** | Probabilistic Model | A probabilistic classification algorithm that uses Bayes theorem and is widely used in text classification tasks. Naive Bayes requires a strong assumption that the predicting variables in the model are independent of each other. |
| **Support Vector Machines (SVM)** | Statistical Model | A powerful class of algorithms that can be used to solve a binary classification problem. Different kernels are utilized to find the hyperplane to separate the dataset. |
| **Classification and Regression Trees (CART)** | Decision Tree-based Model | A tree-structure classifier, starting from the root node and continuously judging and classifying from top to bottom. Each node of CART represents an explanatory variable. The criterion for the node selection is to minimize impurity and obtain the maximum information gain and the bottom leaf node will be returned. |
| **Random Forest (RF)** | Ensemble Models | A model that constructs multiple decision trees using bagging algorithms to reduce the risk of overfitting and improve accuracy. |
| **Gradient Boosting (GB)** | Ensemble Models | A model that iteratively improves the performance of a weak learner by adding new models that correct errors made by the previous ones. |

**Table 7. Description of the Six Basic Machine Learning Models**

## Large Language Model (GPT-3.5 Turbo Model)

A chatbot called ChatGPT (Chat Generative Pretrained Transformer) was created by OpenAI and released in November 2022. It immediately grew in popularity for being able to produce human-like real-time writing and precise answers in almost every subject area (Samantha, 2022). The chatbot was developed on top of the OpenAI large language models (LLMs) from the GPT-3.5 and GPT-4 families, and it has been fine-tuned (a method of transfer learning) using both supervised and reinforcement learning strategies on a massive amount of data sources.

Based on our interactions with the model, it is claimed that a great amount of training data is used to provide a broad understanding to the model of how people will use language and emojis in different contexts and situations. A wide range of social media platforms are allegedly included in the training data as well, such as Twitter, Facebook, Instagram, Reddit, LinkedIn, and TikTok. As both the StockSen and CMC datasets are also retrieved from social media platforms, we believe that ChatGPT has some prior understanding of emojis. To assess its understanding of emojis before fine-tuning on a specific task and domain, we ask ChatGPT what are the most frequently used emojis in the crypto reviews in the chat completion mode (Figure 5). It is observed that ChatGPT's understanding of some emojis given the context of crypto review is even better than referring to Emojipedia website during our preparation of the "Refined Version" dataset.

When using the ChatGPT GUI to answer some simple queries regarding financial sentiment analysis as shown in Figure 5, it is observed that it can indeed understand the sentiment behind the short text, and even the meaning of "🌕" and "🐳". We therefore conclude that the underlying GPT-3.5 Turbo model that is optimized for dialogue is powerful in both understanding the emoji meanings and financial sentiment analysis. We further test the GPT-3.5 Turbo model's performances using the Chat Completion API by setting the context and giving some instructional examples together with the real prompt.

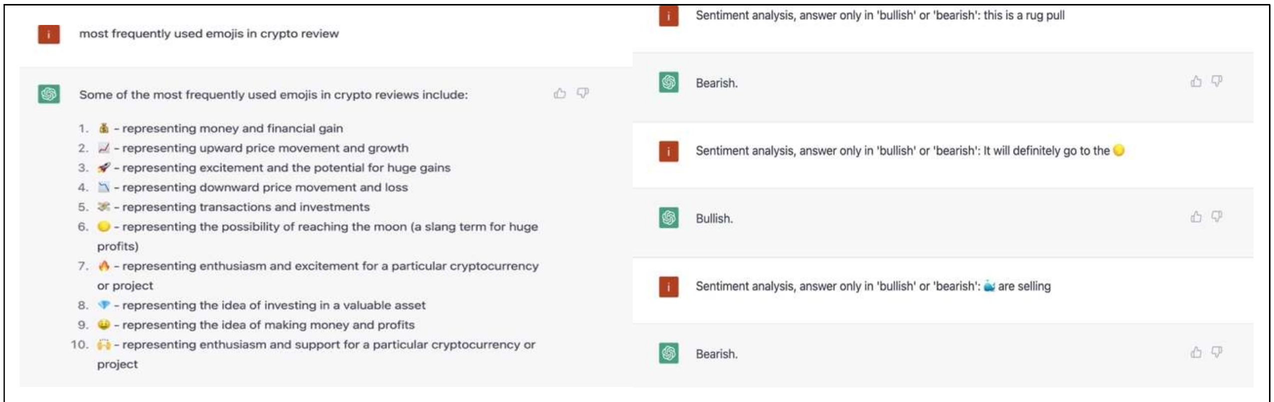


**Figure 5. ChatGPT's Understanding of Emojis**

| | text | target | len | pred |
|---|---|---|---|---|
| 0 | red circle 2013 - You missed red circle 2... | 1 | 60 | 1 |
| 1 | This is my manifest. Dear friends. What makes ... | 1 | 173 | 1 |
| 2 | MISA preciousness preciousness preciousness... | 1 | 21 | 1 |
| 3 | This reminds me a lot of the 50% pandemic cras... | 1 | 107 | 1 |
| 4 | i've said a few days ago,17k is not a support,... | 0 | 17 | 0 |
| 5 | It seems all you bears missed that weekly bull... | 1 | 70 | 1 |
| 6 | I m bullish with best Token with low MarketC... | 1 | 24 | 1 |
| 7 | Make It Happen CZ.. Pump It excellent and exc... | 1 | 16 | 3 x 1, the sentiment score of this review is 1... |
| 8 | i'm waiting for buy 55 coin under ,000 feeling... | 1 | 65 | 1 (The overall sentiment is positive, indicati... |
| 9 | Expect what you dont expect ,Im Pablo Emilio E... | 1 | 50 | 0 |
| 10 | If nukes this month 2023, 2024 and 2025 shoul... | 1 | 31 | 0 |
| 11 | Historical prices for today, December 7th 202... | 1 | 56 | I'm sorry, but I am not able to predict future... |

**Figure 6. GPT-3.5 Turbo Model Sentiment Prediction Result**

Since the output by the Chat Completion API is not standard (return only 0 or 1) and the domains covered by the underlying GPT-3.5 Turbo model are too general, we only test the classification performance on the first 40 rows of the CMC dataset. Out of the 40 reviews, 4 cases produced unexpected predictions, 28 sentences are predicted correctly, while the remaining 8 are predicted incorrectly (Figure 6). We therefore estimate the model accuracy to be slightly below 28/36=78%. Since the baseline accuracy is 83%, we conclude that the underlying GPT model has to be fine-tuned before fair comparisons can be made to the machine learning and deep learning models.

## Large Language Model (Fine-tuning Model)

Fine-tuning is a technique to enable few-shot learning by giving more examples for the LLM to learn in advance so that we do not have to give examples every time in the prompt. Meanwhile, this largely saves the token costs of sending requests with detailed context to the model using Chat Completion API. At the time of this study, there are only 4 base GPT-3 models made available to be fine-tuned by OpenAI, i.e., Davinci, Curie, Babbage, and Ada, in descending order of model quality and cost. In our experiment, to balance FSA performances with the cost for model training, we choose to fine-tune based on the Babbage and Curie model.

The primary preparation of the fine-tuning task is on the dataset, which requires structured input data in JSONL format for the training dataset with two arguments, "prompt" and "completion", and separated by some special characters. Figure 7 provides examples of how to encode training data as prompts.

```
{"prompt": "MISA 💎💎💎 Only token stays strong in bear market. 1000X Soon 🚀🚀🚀===>", "completion": "
1"}
{"prompt": "This reminds me a lot of the 50% pandemic crash in 2020. The bears in these comments were
spreading so much fud when we had that Covid crash saying expect much lower prices we in a pandemic, an
unprecedented time, calling for lower price targets and saying get out of the ponzi, etc and then it
just went on a massive bull run from 3K to 69k. So many of you bears were screaming this time is
different back then and then we pumped so hard on you. This makes me even more bullish now. You bears
have no room to talk at all LOL 👱===>", "completion": " 1"}
{"prompt": "i've said a few days ago,17k is not a support,ofcourse i was called clown again 🤡===>",
"completion": " 0"}
```

**Figure 7. The Preparation of JSONL Training Dataset**

To avoid information leakage, we separate the dataset by the same default 3:1 ratio with stratified sampling and leave aside the validation dataset to check the model performance. The prepared training JSONL file is then uploaded to the cloud before requesting for model training. When the model is trained successfully, the customized model can be quoted in an API call with prompt message and parameters such as “temperature”, which can control the randomness of the output, and “max_token” to limit the returned word count. In our case, as we want to have a more predictable result with only one value 0 or 1 to indicate bearish or bullish, we set temperature to 0.3, and max_token to 1.

With that, we iterate through the testing dataset and fit every row into our fine-tuned base models and obtained the predicted outcome. This whole process is performed on 4 versions of dataset, and the performance metrics of two chosen base models are recorded in Table 11. The advantage of using the Fine-tuning mode over the Chat Completion mode is apparent. Especially, as the cost of calling the API depends directly on the tokens inside each query, using the Fine-tuning API saves the tokens used to describe the problem settings, and also ensures the stability of the output predictions. After fine-tuning, there are no longer predictions generating random text messages other than 0 or 1 as in the previous case.

## Empirical Results and Discussion

In this section, we present model performances in the sequence of (1) six basic machine learning models, (2) deep learning models, and (3) the state-of-the-art GPT-based model. Each part is followed by discussions on the empirical results.

### *Basic Machine Learning Model Performances*

We report model performance metrics including accuracy, precision, recall, and F1-score in Table 8 & Table 9. As shown in Table 8, with only emojis icons added to the dataset in the Emojis Version, their sentiments cannot be learned by the models sufficiently, resulting in even poorer performances compared to the Baseline Version. However, after converting all the emojis into description phrases (Phrases Version), it results in an average increase of the model performance metrics by 1.02% compared with the emojis version dataset, and 0.22% compared with the baseline dataset. Among these, the Multinomial Naive Bayes and Random Forest Models achieve the highest performance compared with other models. Hence, it is shown that converting emojis to their description texts can indeed improve financial sentiment analysis performances. When applying the refined phrases on the StockSen dataset of 30 emojis, however, there aren’t obvious enhancements in the model performances across all models on average, probably due to the scarcity of data containing emojis (13.97%).

This enhancement is particularly noticeable within the CMC dataset, where each data entry includes emojis. We observe from Table 9 that the Phrases Version dataset achieved over 90% accuracy, precision, recall, and F1-score across all the models, which is significantly higher than the Baseline Version (i.e., without emoji) and Emojis Version dataset. When considering the imbalanced dataset feature, the classification result metrics of Phrases Version are all above the 83% baseline accuracy if simply predicting all data entries as bullish. Furthermore, after refining the descriptive phrases, the model accuracy exhibits an additional average enhancement of 1.95% compared with Phrases Version, and 8.45% when contrasted with Baseline Version. This suggests the effectiveness of converting emojis to high quality and context-aware descriptive phrases.

From both the StockSen and CMC dataset results, it is also identified that although most models appear to be incompatible with pure emojis, tree-based models (CART, RF, and GB) have the capacity to capture their intrinsic meanings when exposed to a larger variety of emojis in the training data. Consequently, this leads to a substantial increase in Emojis Version in comparison of Baseline Version.

| **FSA Model Performances on Stock Dataset** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **LR** | Accuracy | Precision | Recall | F1-score | **CART** | Accuracy | Precision | Recall | F1-score |
| Baseline Version | 74.43% | 81.75% | 83.38% | 82.56% | Baseline Version | 72.54% | 78.25% | 86.10% | 81.99% |
| Emojis Version | 74.01% | 81.89% | 82.47% | 82.18% | Emojis Version | 71.65% | 78.78% | 83.46% | 81.05% |
| Phrases Version | 75.08% | 82.37% | 83.58% | 82.97% | Phrases Version | 71.89% | 79.06% | 83.40% | 81.17% |
| Refined Version | 74.95% | 82.29% | 83.49% | 82.88% | Refined Version | 72.35% | 79.45% | 83.55% | 81.45% |
| **MNB** | Accuracy | Precision | Recall | F1-score | **RF** | Accuracy | Precision | Recall | F1-score |
| Baseline Version | 74.91% | 84.04% | 80.79% | 82.38% | Baseline Version | 75.26% | 79.94% | 88.01% | 83.78% |
| Emojis Version | 74.01% | 83.83% | 79.58% | 81.65% | Emojis Version | 74.49% | 79.88% | 86.74% | 83.17% |
| Phrases Version | 75.34% | 84.60% | 80.75% | 82.63% | Phrases Version | 75.16% | 80.58% | 86.71% | 83.53% |
| Refined Version | 75.25% | 84.61% | 80.60% | 82.55% | Refined Version | 75.12% | 80.64% | 86.53% | 83.48% |
| **SVM** | Accuracy | Precision | Recall | F1-score | **GB** | Accuracy | Precision | Recall | F1-score |
| Baseline Version | 73.40% | 80.67% | 83.32% | 81.97% | Baseline Version | 67.41% | 82.20% | 70.33% | 75.81% |
| Emojis Version | 72.39% | 80.55% | 81.73% | 81.14% | Emojis Version | 66.59% | 82.10% | 69.06% | 75.02% |
| Phrases Version | 73.73% | 81.39% | 82.77% | 82.07% | Phrases Version | 68.06% | 82.55% | 71.06% | 76.37% |
| Refined Version | 73.71% | 81.20% | 83.04% | 82.11% | Refined Version | 68.77% | 83.39% | 71.21% | 76.82% |

**Table 8. Model Performances: StockSen Dataset**

| **FSA Model Performance on Crypto Dataset** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **LR** | Accuracy | Precision | Recall | F1-score | **CART** | Accuracy | Precision | Recall | F1-score |
| Baseline Version | 84.95% | 94.89% | 87.66% | 91.13% | Baseline Version | 88.27% | 93.13% | 93.61% | 93.37% |
| Emojis Version | 76.97% | 95.03% | 78.01% | 85.89% | Emojis Version | 88.91% | 92.78% | 94.82% | 93.79% |
| Phrases Version | 91.79% | 97.44% | 93.16% | 95.25% | Phrases Version | 91.35% | 96.09% | 94.04% | 95.05% |
| Refined Version | 94.01% | 97.23% | 95.96% | 96.59% | Refined Version | 92.92% | 96.49% | 95.46% | 95.97% |
| **MNB** | Accuracy | Precision | Recall | F1-score | **RF** | Accuracy | Precision | Recall | F1-score |
| Baseline Version | 87.85% | 95.63% | 90.38% | 92.92% | Baseline Version | 89.22% | 93.27% | 94.61% | 93.93% |
| Emojis Version | 77.37% | 95.14% | 78.40% | 96.96% | Emojis Version | 89.47% | 92.42% | 95.96% | 94.15% |
| Phrases Version | 90.41% | 97.72% | 91.28% | 94.39% | Phrases Version | 92.95% | 98.09% | 95.92% | 96.01% |
| Refined Version | 93.26% | 97.62% | 94.68% | 96.13% | Refined Version | 93.98% | 95.88% | 97.38% | 96.62% |
| **SVM** | Accuracy | Precision | Recall | F1-score | **GB** | Accuracy | Precision | Recall | F1-score |
| Baseline Version | 84.21% | 93.56% | 88.18% | 90.79% | Baseline Version | 76.90% | 93.83% | 79.02% | 85.79% |
| Emojis Version | 76.28% | 94.37% | 77.80% | 85.29% | Emojis Version | 90.69% | 92.63% | 97.20% | 94.86% |
| Phrases Version | 92.64% | 96.91% | 94.68% | 95.78% | Phrases Version | 91.29% | 97.07% | 92.94% | 94.96% |
| Refined Version | 94.33% | 97.01% | 96.56% | 96.78% | Refined Version | 93.61% | 96.42% | 96.35% | 96.38% |

**Table 9. Model Performances: CMC Dataset**

## *Deep Learning Models Performances*

We report accuracies also for the LSTM and BERT models (Table 10), but only on the CMC dataset, as the data portion with emojis in StockSen dataset is too small to train deep learning models. Both models already perform well on the Emoji Version dataset with high validation accuracy of 93.1%. After transforming emojis into descriptive phrases within either the Phrases Version or the Refined Version, there is a further enhancement of the model performances, particularly evident in the case of the BERT model. While the effect on the LSTM model is not as pronounced, it still showcases some improvement. In line with the outcomes observed in basic machine learning models case, the conversion of emojis into high-quality and contextually aware phrases is similarly effective when applied to deep learning models for financial sentiment analysis.

A nuance is that when keeping pure emojis in the training data, LSTM has an increase in accuracy of 2%, while BERT model drops by 0.5% compared with Baseline Version dataset. This observation aligns with findings from Chen's research (2023), revealing that both the base and large versions of BERT encoders encounter challenges with emojis, which is due to the tokenizer replacing emojis with out-of-vocabulary unknown tokens (e.g., "<UNK>") instead of generating distinct representations for these emoji tokens.

| | Baseline Version | Emoji Version | Phrases Version | Refined Version |
|---|---|---|---|---|
| **LSTM Model** | 91.1% | 93.1% | 93.2% | 93.4% |
| **BERT Model** | 93.6% | 93.1% | 94.5% | 94.8% |

**Table 10. Validation Accuracies of 4 Dataset Versions: Deep Learning**

| Fine-tuning Base Model | Data Version | Accuracy | Precision | Recall | F1_Score |
|---|---|---|---|---|---|
| Babbage (GPT-3) | Baseline Version | 93.38% | 95.83% | 96.71% | 96.27% |
| | Emojis Version | 93.48% | 95.95% | 96.70% | 96.33% |
| | Phrases Version | 95.17% | 96.51% | 98.09% | 97.29% |
| | Refined Version | 95.14% | 97.20% | 97.30% | 97.25% |
| Curie (GPT-3) | Baseline Version | 94.15% | 96.01% | 97.42% | 96.71% |
| | Emojis Version | 99.40% | 99.51% | 99.82% | 99.66% |
| | Phrases Version | 95.33% | 97.01% | 97.73% | 97.37% |
| | Refined Version | 95.46% | 97.14% | 97.73% | 97.44% |

**Table 11. Validation Accuracies after Fine-tuning GPT-3 Base Models**

| Confusion Matrix | Fine-tuning Base Model | Baseline Version (2857, scaled) | Emoji Version (3191) | Phrases Version (3191) | Refined Version (3191) |
|---|---|---|---|---|---|
| TN FP<br>FN TP | Babbage | 257 118<br>93 2723 | 256 115<br>93 2727 | 271 100<br>54 2766 | 283 79<br>76 2744 |
| | Curie | 261 114<br>93 2723 | 357 14<br>5 2815 | 286 85<br>64 2756 | 290 81<br>64 2756 |

**Table 12. Confusion Matrices of GPT-3 Fine-tuned Models Predictions**

## *Large Language Models (LLM) Performances*

We report performance metrics when using the training data of our four versions of datasets to fine-tune the base GPT-3 models with 4 epochs, including Babbage and the more advanced Curie model. Babbage is a large model with 3 billion parameters and 300GB of text data. It can handle more complex classification natural language tasks, such as reasoning, logic, arithmetic, and word analogy than Ada. While Curie is an even larger model, with 13 billion parameters and 800GB of text data. In addition to the functionality of Babbage, it's also capable of handling more complex classification tasks and more nuanced tasks like summarization, sentiment analysis, chatbot applications, and question answering (Maeda et al, 2023).

The model performances in Table 11 are measured on the validation datasets. For the Baseline Version dataset, notable improvements in classification performance of the GPT-3 fine-tuned models over the basic machine learning algorithms are readily apparent (Table 9). The Baseline Version accuracy rates of 93.38% with the Babbage and 94.15% with the Curie fine-tuned models closely resemble the accuracies achieved with the Refined Version when utilizing basic machine learning models, as well as with the Phrases Version when leveraging deep learning models.

The accuracy of the Babbage model on Refined Version is 95.11%, which is the highest among all the models expect from Curie model. Surprisingly, when fitting the more advanced and powerful Curie model, although there's also enhancement of the classification performance from Baseline Version to Phrases & Refined Version, it achieves the highest performance on the Emoji Version dataset with 99.4% accuracy rate. This suggests its distinct and remarkable capacity of recognizing sentiments behind pure emojis in financial sentiment analysis task.

Confusion matrices are also reported in Table 12 to provide a more straightforward illustration of the percentages of type I and type II errors when emojis are handled differently. The confusion matrices of the Baseline Version dataset are scaled up to facilitate the comparison, as clearing emojis resulting in lost data entries. From these matrices, it is observed the prediction errors are basically balanced except from the Phrases Version with Babbage model. Nevertheless, the matrix resumes balance in Refined Version. Despite the outstanding performance by Curie model in Emoji Version, this re-confirms the benefits using high quality and context-aware descriptive phrases to replace emojis in financial sentiment analysis.

## *Significance Analysis*

To further prove the validity of the empirical results, we applied significance test to the model results focusing on CMC dataset, treating the difference in model accuracy between treated data versions and the baseline version as a set of random variables, r1 (Emoji Version vs. Baseline Version), r2 (Phrases Version vs. Baseline Version), and r3 (Refined Version vs. Baseline Version).

The computed 95% confidence intervals provide insightful interpretation. For r1, ranging from -0.056 to 0.046, the span crossing zero signifies that the treatment yielded no uniform accuracy enhancement or degradation. Conversely, the confidence interval for r2 (0.015, 0.075) stands entirely above zero, indicating a statistically significant and positive impact of the treatment on accuracy in the phrases version compared to the Baseline Version. Similarly, the confidence interval for r3 (0.022, 0.093) also excludes zero, underscoring a statistically significant improvement in accuracy for the refined version relative to the baseline version. In recapitulation, the assessment of significance test outcomes indicates the absence of statistical significance in accuracy improvement for the Emoji Version, while confirming the statistically significant and coherent accuracy enhancements observed in both the Phrases Version and Refined Version, when contrasted with the Baseline Version.

With the significance test applying to other model metrics like precision, recall, and F1 score, we noticed the similar results achieved shown in Table 13. These results consistently highlight enhanced accuracy through the treatment of converting emojis to their high quality and context-aware descriptive phrases.

| Treatment Effects | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| **r1** (Emoji vs. Baseline) | (-0.056, 0.046) | (-0.011, 0.004) | (-0.139, 0.102) | (-0.077, 0.050) |
| **r2** (Phrases vs. Baseline) | (0.015, 0.075) | (0.024, 0.033) | (-0.007, 0.102) | (0.008, 0.070) |
| **r3** (Refined vs. Baseline) | (0.022, 0.093) | (0.021, 0.033) | (0.012, 0.131) | (0.019, 0.083) |

**Table 13. The 95% Confidence Intervals of Treatment Effects (CMC Dataset)**

## Conclusion and Future Work

In this paper, we discussed the best practice when handling emojis for financial sentiment analysis. Experiments are conducted through four data-prepossessing methods on emojis, and the model robustness is tested across a variety of machine learning and deep learning models, such as LSTM and BERT. Moreover, the recent GPT-based large language model (LLM) is leveraged to test its understanding of emojis and to perform the financial sentiment analysis task after fine-tuning.

Instead of simply keeping emojis in their original forms or deleting them from the corpus, an experiment of mapping emojis to phrases is carried out and it indeed helps to improve the model performance. Two datasets are used in the experiments: one is StockSen dataset from general stock reviews, and the other is CoinMarketCap dataset from cryptocurrency reviews filtered with emojis and self-labels by commentors.

Out of the four pre-processed dataset versions (Baseline Version, Emoji Version, Phrases Version, and Refined Version), the Phrases Version and Refined Version result in a noteworthy consistent improvement in accuracies on both the stock and cryptocurrency datasets by converting emojis to corresponding explanatory phrases. On average, the increase is between 2-4% when all reviews inside the corpus include at least one emoji. With the refined descriptive phrases, the model performance is further enhanced by 2-3%. Compared with the Baseline Version without emojis, there is an overall 8.45% increase in the model performance metrics on average.

The emergence of the ChatGPT tool and its underlying GPT model series provides a powerful foundation for general language understanding and generation and has pushed the development of natural language processing technology to a new level. After verifying its ability of understanding emoji icons in the crypto review context, the Chat Completion and Fine-tuning APIs are employed. Out of all the models used in this research, the GPT-3 Curie model achieved the best performance after fine-tuning on the Emoji Version dataset, followed by Refined and Phrases Version datasets. This performance improvement may benefit many downstream tasks such as financial time series prediction.

In summary, we made several contributions to the IS literature: (1) We discovered that emojis are important non-verbal feature for financial sentiment analysis by studying the StockSen dataset and we recovered all the emojis to enhance the data quality; (2) We collected a novel crypto review dataset with over 10000 records with both self-labels and rich emojis features from the CoinMarketCap website; (3) We conducted different treatments on the emojis and concluded the most effective approach in financial sentiment analysis is by using high-quality descriptive phrases within financial context (Refined version). (4) We identified LLM starting from Curie has exceptional capacity to perform FSA task with emojis in original form.

The differences in (3) illustrate the importance of properly encoding emojis to fully leverage and understand the information content. Our study alerts that the predictive power of social media data depends heavily on rigorous data treatments, and clearing out emojis will result in information loss and poorer FSA results.

There are certain limitations to our study. (1) Due to budget and time constraints, we did not test all the available models from OpenAI. For example, Davinci, which is the largest and most powerful model available for fine-tuning at the time of this research, with 175 billion parameters and 45TB of text data. (2) When converting to phrases, each emoji is treated as a single token and the problem of polysemy within the finance domain is not considered. For this reason, our qualitative analyses of the emojis are only statistically meaningful and concerns the primary sense. In future work, we will conduct analysis on those more “neutral” emojis in a finer granularity to investigate multi-sense emojis.

## Acknowledgements

The authors would like to thank the three anonymous reviewers for their helpful suggestions for improvement. This research has been partially supported by Singapore MOE research grant number 251RES2107.

## References

Barbieri, F., Ronzano, F. & Saggion, H. (2016). What does this Emoji Mean? A Vector Space Skip-Gram Model for Twitter Emojis. *Proceedings of LREC*, 3967-3972. (https://aclanthology.org/L16-1626)

Broni, K. (2021). Emoji use at all-time high. *Web resource*.(https://blog.emojipedia.org/emoji-use-at-all-time-high)

Chen, B. (2023). Emojis Aid Social Media Sentiment Analysis: Stop cleaning them out. *Towards Data Science*. (https://towardsdatascience.com/emojis-aid-social-media-sentiment-analysis-stop-cleaning-them-out-bb32a1e5fc8e)

Chen, Y., Yuan, J., You, Q., & Lu, J. (2018). Twitter Sentiment Analysis via Bi-sense Emoji Embedding and Attention-based LSTM. *Proceedings of the 26th ACM International Conference on Multimedia*, 117-125.

Cramer, H., de Juan, P., & Tetreault, J. (2016). Sender-intended functions of emojis in US messaging. *Proceedings of MobileHCI*, 504-509. (https://doi.org/10.1145/2935334.2935370)

Deng, S., & Huang, Z.J., Sinha, A., & Zhao, H. (2018). The interaction between microblog sentiment and stock returns: an empirical examination. *MIS Quarterly*, 42(3), 895-918.

Derks, D., Bos, A.E.R., & von Grumbkow, J. (2008). Emoticons and social interaction on the Internet: the importance of social context. *Computers in Human Behavior*, 23(1), 842–849.

Dong, H., Ren, J., Padmanabhan, B., & Nickerson, J.V. (2022). How are social and mass media different in relation to the stock market? A study on topic coverage and predictive value. *Information & Management*, 59, 103588. (https://doi.org/10.1016/j.im.2021.103588)

Du, K., Xing, F., & Cambria, E. (2023). Incorporating Multiple Knowledge Sources for Targeted Aspect-based Financial Sentiment Analysis. *ACM Transactions on Management Information Systems*, 14(3), 1-24.

Eisner, B., Rocktäschel, T., Augenstein, i., Bošnjak, M., & Riedel, S. (2016). emoji2vec: Learning emoji representations from their description. *Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media*, 48-54. (https://doi.org/10.18653/v1/W16-6208)

Gao, Y., Zhao, C., Sun, B., & Zhao, W. (2022). Effects of investor sentiment on stock volatility: New evidences from multi-source data in China's green stock markets. *Financial Innovation*, 8(77), 1-30.

Hill, F., Cho, K., Korhonen, A., & Bengio., Y. (2016). Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4, 17-30.

Hu, T., Guo, H., Sun, H., Nguyen, T.T. & Luo, J. (2017). Spice up your chat: The intentions and sentiment effects of using emoji. *Proceedings of ICWSM*, 102-111.

Khader, M, Awajan, A., & Al-Naymat, G. (2019). The impact of natural language preprocessing on big data sentiment analysis. *The International Arab Journal of Information Technology*, 16(3A), 506-513.

Lock, S. (2022). What is AI chatbot phenomenon ChatGPT and could it replace humans? *The Guardian*. (https://www.theguardian.com/technology/2022/dec/05/what-is-ai-chatbot-phenomenon-chatgpt-and-could-it-replace-humans)

Ma, X., Xu, P., Wang, Z., Nallapati, R., & Xiang, B. (2019). Domain adaptation with BERT-based domain classification and data selection. *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP*, 76–83.

Maeda, J., & Bolanos, M. (2023). What are Models? Choosing an LLM model. *Microsoft Learn*. (https://learn.microsoft.com/en-us/semantic-kernel/prompt-engineering/llm-models)

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ICLR (Workshop Poster)*, 1-12. (https://arxiv.org/abs/1301.3781)

Qiu, L., & Wang, W., & Pang, J. (2023). The Persuasive Power of Emoticons in Electronic Word-of-Mouth Communication on Social Networking Services. *MIS Quarterly*, 47(2), 511-534.

Saha, J., Patel, S., Xing, F., & Cambria, E. (2022). Does social media sentiment predict Bitcoin trading volume? *ICIS 2022 Proceedings*, 3, 1-9. (https://aisel.aisnet.org/icis2022/blockchain/blockchain/3)

Singh, A., Blanco, E., & Jin, W. (2019). Incorporating Emoji Descriptions Improves Tweet Classification. *Proceedings of NAACL*, 2096–2101.

Suntwal, S., Brown, S., & Brandimarte, L. (2021). Pictographs, Ideograms, and Emojis (PIE): A Framework for Empirical Research Using Non-verbal Cues. *Proceedings of Hawaii International Conference on System Sciences*, 6400-6409.

Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12), 2544–2558. (https://doi.org/10.1002/asi.21416)

Uhr, P., Zenkert, J. & Fathi, M. (2014). Sentiment analysis in financial markets: A framework to utilize the human ability of word association for analyzing stock market news reports. *IEEE International Conference on Systems, Man, and Cybernetics*, 912-917.

Wang, Y. (2021). Emoji for Natural Language Prepossessing. *Master's Thesis, ETH Zürich Distributed Computing Group*. (https://pub.tik.ee.ethz.ch/students/2020-HS/MA-2020-31.pdf)

Xing, F., Malandri, L., Zhang, Y. & Cambria, E. (2020). Financial sentiment analysis: An investigation into common mistakes and silver bullets. *Proceedings of COLING*, 978 - 987.

Yuan, X., Hu, J., Zhang, X., & Lv, H. (2022). Pay attention to emoji: Feature Fusion Network with EmoGraph2vec Model for Sentiment Analysis. *Proceedings of ICPR*, 1529 – 1535.