ISSN (Online): 3033-0106

BUSINESS & MANAGAMENT COMPASS
Volume: 68 (2024), Issue: 2
Doi: https://doi.org/10.56065/crmc1136, pp. 35-45

Issues since 2024: https://bi.ue-varna.bg/ojs/index.php/bmc
Previous issues: http://journal.ue-varna.bg/en/home

Comparing ChatGPT and LSTM in predicting changes in quarterly financial metrics

Vitali CHAIKO[1]

[1] University of Library Studies and Information Technologies Sofia, Bulgaria, chaikovitali.cv@gmail.com

JEL: O14, P24, C41, E13

Abstract

In the financial industry, the ability to predict financial metrics accurately and in a timely manner can significantly impact investment decisions, risk management, and strategic planning. In recent years, machine learning has emerged as a powerful tool for such predictions. This study aims to explore the heretofore underexplored predictive potential of ChatGPT by predicting positive or negative changes in quarterly financial metrics, such as revenue and sales numbers, using textual data from social media. The performance of ChatGPT is compared against Long Short-Term Memory (LSTM) neural network models developed as part of this research. The methodology involves preprocessing large datasets from Twitter concerning major companies such as Amazon, Google, and Tesla, training LSTM models, and prompt engineering for ChatGPT-based predictions. Initial findings indicate that LSTM models can predict quarterly financial metric changes with up to 87% accuracy, significantly outperforming ChatGPT, which achieves a maximum accuracy of 67%. Therefore, at the current time, ChatGPT cannot be considered a reliable predictive tool for changes in quarterly financial metrics.

Key words:
ChatGPT, financial metrics prediction, LSTM, twitter

Published by University of Economics – Varna

Citation: Chaiko, V. (2024). Comparing ChatGPT and LSTM in predicting changes in quarterly financial metrics. *Business & Management Compass.* University of Economics Varna, 68 (2), pp. 35 - 45. DOI: 10.56065/crmc1136

Received: 09.04.2024
Revised: 17.06.2024
Accepted: 28.06.2024
Published: 30.06.2024

1. Introduction

In the financial industry, the accuracy and timeliness of financial metric predictions can crucially impact investment decisions, risk management strategies, and strategic planning. As the financial landscape continues to evolve rapidly, machine learning algorithms have become significant in forecasting various financial metrics. These algorithms, powered by complex neural network models such as Long Short-Term Memory (LSTM) networks, offer the potential to assist in strategic and investment decision-making processes by providing automated, data-driven insights. For instance, stock changes can be predicted quite accurately when applying such algorithms (Pramod & Mallikarjuna, 2021). Moreover, in recent times, large language models (LLMs) have emerged as a new generative artificial intelligence technology, with ChatGPT standing out as a versatile and transformative tool across an array of applications. The performance of ChatGPT in predictive cases remains under-researched nowadays. This study aims to approach the predictive task using social media data. Social media platforms, such as Twitter, are valuable sources of information that reflect the social and economic events of the real world today and several machine learning predictive models have been developed for forecasting financial metrics using data from social media (Albahli, et al., 2022; Memiş, et al., 2024). The challenge of such predictions is that textual data is unstructured and chaotic in nature. Careful selection and preparation of such data are often required to achieve decent predictive performance. Therefore, this study aims to explore the predictive potential and compare the performance of ChatGPT with that of established prediction models in classification tasks related to positive or negative changes in various quarterly financial metrics, such as revenue and sales numbers. The financial metrics will be predicted based on social media textual data, more precisely, Twitter data, regarding three major US companies: Google, Amazon, and Tesla. The structure of this study is as follows: initially, existing literature and solutions are reviewed and presented. The methodology then advances with preprocessing of large Twitter datasets. The LSTM models are subsequently trained to accurately predict changes in financial metrics. Concurrently, prompt engineering is applied to ChatGPT, aiming for the same goal of predicting such changes. Finally, the resulting predictions are compared and discussed, using standardized metrics, such as accuracy and the Matthews Correlation Coefficient (MCC).

2. Literature Review

2.1. LSTM

LSTM networks (Brownlee, 2018; Hochreiter & Schmidhuber, 1997) belong to the family of recurrent neural network architectures (RNN). LSTM based models have demonstrated their effectiveness in a variety of applications, notably in the areas of language modeling (Sundermeyer, et al., 2012) and speech recognition (Graves, et al., 2013). These networks are particularly adept at processing sequential data, a capability that comes from their use of past input and output information in current calculations. This feature is important in tasks that require understanding the temporal dynamics within input data elements and is particularly useful for working with textual data in financial context (Usmani & Shamsi, 2023). Each LSTM layer consists of several memory cells, which are integral to controlling the flow of information through neurons and cells. These cells are characterized by three unique multiplicative gates: the input, forget, and output gates. The ability of LSTM networks to learn long-term dependencies in input data features can be attributed to these gates' functionality in selectively adding, removing, or retaining information within the memory cell. The input gate decides what part of the input data is significant enough for memory storage, while the forget gate plays a role in determining which aspects of the previous state should be discarded. The function of the output gate is to merge the state of the memory cell with the final output layer, a crucial step in the network's decision-making process.

2.2. LLMs

LLMs (Bender, et al., 2021) are types of self-supervised machine learning algorithms that are trained on massive textual datasets, often comprising hundreds of billions of text tokens. The training involves predicting the next word in a sequence based on all previous words. Prominent LLMs such as ChatGPT (Bahrini, et al., 2023; OpenAI, 2023) and Google Bard (Google AI, 2023) demonstrate the ability to generate text that closely mirrors human language. Beyond mere text generation, the models excel in tasks like translation, summarization, and question-answering. These models are built on the generative pre-trained transformer architecture by training them on a diverse range of internet-sourced data. The models are accessible through web-based chat applications in which users interact with them using prompts. Additionally, an application programming interface (API) is available, allowing developers to access the models' capabilities and integrate them into their own applications.

2.3. Financial Metrics Prediction with LLMs

The possibility of predicting financial metrics, especially stock price movements and returns, with ChatGPT has been explored by several studies The use of GPT models in predicting stock returns using news sentiment analysis has been demonstrated (Lopez-Lira & Tang, 2023), achieving cumulative returns of over 550% and a Sharpe ratio of 3.1 with ChatGPT 3.5, and cumulative returns of over 350% and a Sharpe ratio of 3.8 with ChatGPT 4. However, the authors also outline possible overfitting of the models and limited real-word applicability because of certain assumptions of fixed costs per trade and no large trades. ChatGPT and other large language models were utilized for predicting Chinese stock returns through sentiment analysis of news summaries (Zhang, et al., 2023), achieving annual excess returns of 23.17%, 19.88%, and 24.14% and Sharpe ratios of 0.6444, 0.4836, and 0.6841 for ChatGPT, Chinese-FinBERT, and Erlangshen-110M models, respectively, highlighting the superior performance of the language-specific models over ChatGPT in this context. Another study (Ding, et al., 2023) demonstrates the integration of LLMs with stock features significantly improves stock return predictions, achieving an annual return of 28.8% and a Sharpe ratio of 1.24 with the SCRL-LG model, highlighting the potential of LLMs in enhancing financial forecasting despite concerns about model complexity and assumptions of fixed costs per trade. Moreover, the predictive ability of ChatGPT has been demonstrated in predicting future rate decisions of the Bank of England committee by analyzing the textual content of each member's speeches (Woodhouse & Charlesworth, 2023). The content of the speeches is classified by ChatGPT into categories such as inflation hawk, and a policy rate model is constructed based on those classifications. The resulting model score shows a significant statistical and economic relationship to future interest rate decisions with regression coefficients of up to 0.168 and R^2 values up to 0.32. A survey of LLMs in the finance sector evaluates the performance of various natural language processing tasks, including text classification and stock price movement prediction (Lee, et al., 2024). The LLMs, especially GPT-4 and specific financial LLMs such as BloombergGPT and FinBERT, trained on financial data, achieve accuracy values of 0.54 in stock movement prediction and F1-Scores of 0.98 in financial text classification demonstrating prediction capabilities in the financial sector. To summarize the findings of the related studies, the performance of ChatGPT in prediction tasks, especially those related to stock returns, is promising but varies significantly across different data sets and financial metrics.

3. Material and Methods

3.1. Dataset

This work leverages a subset from a collection of historical tweets from the years 2015 to 2020, focusing on top NASDAQ companies. The dataset was created as part of a stock market research study (Doğan, et al., 2020) and is made available in two parts on a public data science & code platform Kaggle (Kaggle, 2023): the

tweet data and the companies' market data, for purpose of use by other researchers. The dataset consists of more than 3 million tweets. Each tweet record in the dataset is detailed with unique identifiers, author information, the tweet's creation date, its textual content, the related NASDAQ company, and metrics of social engagement such as comments, likes, and retweets. The excerpt of the dataset structure is presented in Table 1. (In accordance with Twitter's content distribution policy, the exact contents of the tweets cannot be displayed in this work. The tweet contents presented in this study are not from real tweets of the dataset but are semantically similar). The tweet records related to Google, Amazon, and Tesla are selected from the dataset for the subsequent training and evaluation of the prediction models. This selection is made because these companies represent a major part of dataset's data and precise quarterly financial metrics are readily available online for them. These financial metrics are obtained from public statistical data: Amazon quarterly revenue (Statista, 2024a), Tesla quarterly car sales numbers (Statista, 2024b) and Google quarterly search engine market share numbers (Statista, 2024c). As an example, the excerpt of Amazon's quarterly revenue financial metrics is illustrated in

Table ***2***. The hypothesis formed from the gathered data suggests that the contents of multiple tweets—specifically their words, semantics, and syntax—are related to the increase or decrease of financial metrics at the end of the quarter in which the tweets were generated. Throughout this study, such collections of tweets are referred to as 'tweet groups.' For instance, in Table 1, the tweets with tweet IDs from 551234509 to 551234518 are identified as the first tweet group of size five.

Table 1

Excerpt of twitter dataset

tweet_id	writer	post_date	body	comment_num	retweet_num	like_num	ticker_symbol
551234509	Global Stocks	10/01/2015 14:32:10	$AMZN acquisitions & #onlineactivity are at risk - @Amazon declines to adopt #SecureWeb across all pages #privacyconcerns	11	1	4	AMZN
551234511	News24	10/01/2015 17:01:14	Risk Retailers Operating Cash Flow1. $WMT2. $AMZN3. $TGTFull Overview: #MarketTrend	5	0	0	AMZN
551234514	Tim101	10/03/2015 10:16:27	$AMZN $MSFT $LYFT What's the direction for tomorrow? - Insights from Equity Trading Forum - Investor Discussions	0	0	0	AMZN
551234516	Trader Stop	10/03/2015 11:12:21	Not a supporter of @MarketMaverick on @CNBCMarketWatch, consistently pessimistic. Relieved I dismissed your advice, Mark, and invested in $amzn when it was under $350 #shiftfrombearishbullish!;	3	1	12	AMZN
551234518	Sarah Gord	10/04/2015 12:09:05	$AMZN Popular Series Decline in Appeal with Network Repeats ;Expanding Cloud Strategy: Aiming to be an AWS Consulting Partner? $VMW $AMZ	0	0	0	AMZN
551234519	News25	10/04/2015 12:14:01	$BND $AMZN $MSFT $LYFT What's the direction for tomorrow? - Insights from Equity Trading Forum	2	1	5	AMZN

Source: (Kaggle, 2023)

Table 2

Excerpt of Amazon revenue in billion U.S dollar

from_date	to_date	value
01/10/2014 00:00:00	12/31/2014 23:59:59	20.58
01/01/2015 00:00:00	03/31/2015 23:59:59	29.33
01/04/2015 00:00:00	06/30/2015 23:59:59	22.72
01/07/2015 00:00:00	09/30/2015 23:59:59	23.18
01/10/2015 00:00:00	12/31/2015 23:59:59	25.36
01/01/2016 00:00:00	03/31/2016 23:59:59	35.75

Source: (Statista, 2024a)

3.2. LSTM model training

To predict changes in financial metrics from tweets, custom LSTM models are trained using word contents encoded as word embeddings (Mikolov, et al., 2013) of each tweet to perform the prediction. No other data transformations, such as sentiment analysis, are performed, as word embeddings are able to capture nuanced and detailed information from the text. Firstly, the dataset column, which contains the text of the tweets, is preprocessed by removing all stop words, hyperlinks, and numbers. Then, custom word embeddings that represent each word as a 300-dimensional vector were trained for every word in the dataset using Gensim Python library (Rehurek & Sojka, 2010). Each training input sample is then represented by a list of word embeddings, created by taking a group of tweets of size *N* and encoding every subsequent word in this group as a word embedding. Each training output target class is represented by one of two values: decrease (0) or increase (1), depending on the change in the financial metric at the time the tweet group was created compared to the previous financial metric value. For instance, if the revenue value was 20 billion dollars in the first quarter of a year and increased to 29 billion in the second quarter, the target class for tweet groups written in the second quarter would be 1. The LSTM architecture is then built with the PyTorch Python framework (Paszke, et al., 2019), and the LSTM model is trained with the data by applying the ten-fold cross-validation method (Scikit-Learn, 2024). The exact training parameters of LSTM model are presented in Table 3. Multiple models are trained on different subsets of tweets and financial metrics for tweet group sizes of *N* of 5, 10, and 20:

1) *AmazonLSTM@N*: Prediction of the increase or decrease in Amazon's revenue from 2015-2020, using word embeddings of tweet groups corresponding to the company Amazon as input variables and the quarterly revenue changes during that time period as target prediction class.
2) *TeslaLSTM@N*: Prediction of the increase or decrease in Tesla car sales numbers from 2015-2020, using word embeddings of tweet groups corresponding to the company Tesla and the quarterly car sales numbers changes during that time period as target prediction class.
3) *GoogleLSTM@N*: Prediction of the increase or decrease in Google's search engine market share from 2015-2020, using word embeddings of tweet groups corresponding to the company Google and the quarterly search engine market share changes during that time period as target prediction class.

Table 3

LSTM training parameters

Parameter	Value
Embedding layer	torch.nn.Embedding.from_pretrained(word_embeddings, freeze=False)
LSTM Layer	torch.nn.LSTM(emb_size=300, hidden_size=512, num_layers=2, batch_first=True, dropout=0.2)
Linear layer 1	torch.nn.Linear(512, 512)
Linear layer 2	torch.nn.Linear(512, 256)
Linear layer 3	torch.nn.Linear(256, 2)
Activation function	torch.nn.functional.relu
Loss function	torch.nn.functional.cross_entropy
Optimizer	torch.optim.Adam
Batch size	100
Epochs	10

Source: Author's Compilation, 2023

3.3. Prompt engineering

Prediction prompts for ChatGPT are developed to predict changes in financial metrics using the content of tweet groups of size N, aiming to achieve a procedure similar to LSTM-based prediction. Initially, an instruction prompt is provided that defines the rules of the prediction task. No other information is provided to ChatGPT before the instruction prompt. The instruction prompt is deliberately chosen to be as simple as possible, not providing further metadata such as temporal information, to quickly perform a prediction task and to avoid overloading ChatGPT with instructions, as ChatGPT tends to lose focus when following complex tasks (Chen, et al., 2024). This instruction prompt is created for each of the three observed companies:

1) *AmazonChatGPT@N*:

"Given a group of tweets about the company Amazon in my next input, separated by '<SEP>', for each group, output 1 if the text of this group would indicate an increase in Amazon's quarterly revenue, or 0 if it would indicate a decrease. Output the numbers as a list separated by commas. Provide no further explanation."

2) *TeslaChatGPT@N*:

"Given a group of tweets about the company Tesla in my next input, separated by '<SEP>', for each group, output 1 if the text of this group would indicate an increase in Tesla's quarterly car sales numbers, or 0 if it would indicate a decrease. Output the numbers as a list separated by commas. Provide no further explanation."

3) *GoogleChatGPT@N*:

"Given a group of tweets about the company Google in my next input, separated by '<SEP>', for each group, output 1 if the text of this group would indicate an increase in Google's quarterly search engine market share, or 0 if it would indicate a decrease. Output the numbers as a list separated by commas. Provide no further explanation."

After the instruction prompt, 100 tweet groups of size *N* are selected randomly from the dataset and used subsequently as input for prediction prompts. In response, ChatGPT generates output prediction prompts that contain 'increase' (1) or 'decrease' (0) values, indicating the predicted changes in financial metrics. This process is repeated for all three companies, and for each company, tweet groups of sizes 5, 10, and 20 are used. The interactions with ChatGPT 4, as of 12.14.2023, for the first three input tweet groups of size 5 related to the company Amazon are documented in Table 4.

Table 4

Interactions with ChatGPT for prediction of Amazon revenue

Input Instruction prompt:
Given a group of tweets about the company Amazon in my next input, separated by '<SEP>', for each group, output 1 if the text of this group would indicate an increase in Amazon's quarterly revenue, or 0 if it would indicate a decrease. Output the numbers as a list separated by commas. Provide no further explanation.
ChatGPT output:
Sure, I'm ready. Please provide the group of tweets.
Input prediction prompt:
$TRIP Forum Updates Available: View Here: $T $VWO $PFE $AMZN ;$QCOM Price Movement Recap Thursday, April 2, 2015 10:07:16 PM $MSFT $CVX $DWT $AMZN ;Not a supporter of @MarketMaverick on @CNBCMarketWatch, consistently pessimistic. Relieved I dismissed your advice, Mark, and invested in $amzn when it was under $350 #shiftfrombearishbullish!;4/2/15 - US: $DIA -0.33% $XLK -0.50% $VTI +0.05% RED: $MSFT $GOOGL $AMZN $PYPL $DIS $NVDA $BABA $SQ $TWTR $SNAP $FB $CRM BLUE: $LYFT $NFLX;$AG Stock Community Update Thursday, April 2, 2015 11:07:16 PM $INTC $DELL $AMZN $CGC <SEP>* *$AMZN acquisitions & #onlineactivity are at risk - @Amazon declines to adopt #SecureWeb across all pages #privacyconcerns;$OIL Business Profile Refreshed Sunday, April 5, 2015 7:15:51 PM $PFE $YUM $AMZN $GLUU ;Risk Retailers Operating Cash Flow1. $WMT2. $AMZN3. $TGTFull Overview: #MarketTrends;$AMZN Latest Updates Sunday, April 5, 2015 7:08:11 PM $GILD $BMY $IYE $SLX ;$MTD Latest Updates Sunday, April 5, 2015 7:02:20 PM $GSK $SPY $AMZN $JNK<SEP>* *$GPRO $NVAX $DAL $CELG $AMZN $CTIC $AGEN $FXI ;Reviewing the upcoming trading session charts - $DIA $MDY $XLK $XBI $NVDA $SNAP $LYFT $AMZN $REGN $VRTX $JD $MSFT ;$BND $AMZN $MSFT $LYFT What's the direction for tomorrow? - Insights from Equity Trading Forum - Investor Discussions ;$AMZN Popular Series Decline in Appeal with Network Repeats ;Expanding Cloud Strategy: Aiming to be an AWS Consulting Partner? $VMW $AMZN<SEP>
ChatGPT prediction output:
1,0,1

Source: Author's Compilation, 2023

3.4. Evaluation metrics

For the evaluation of financial metric predictions, predefined metrics are applied that are widely used in the evaluation of machine learning prediction models. The metric 'precision' indicates how many of the samples predicted as true are actually true (Scikit-learn, 2023e). Recall indicates the total number of correct classifications made out of all actual true instances (Scikit-learn, 2023e). Accuracy is a metric that calculates the overall performance of a classification model across all classes; it measures the proportion of correctly predicted instances out of the total instances (Scikit-learn, 2023e). The F1-score (Murphy, 2012) is the harmonic mean of precision and recall, and the value range of all the above metrics is {0,1}. The closer the value is to 1, the better the prediction performance. Another evaluation metric for the performance of binary classifiers is MCC (Chicco & Jurman, 2020): MCC takes into account true and false positives and negatives and is generally regarded as a balanced measure, which can be used if the classes provided are of very different

sizes. The MCC value ranges from -1 to +1. A value of +1 represents a perfect prediction, 0 represents an average random prediction, and -1 represents an inverse prediction.

4. Results and Discussion

4.1. Evaluation

The evaluation of LSTM models and ChatGPT classification prompts, using specific metrics, is shown in Table 5. For the evaluation of LSTM models, the metrics are applied to a holdout sample of 100 tweet groups that were not used for training the model. This process is repeated 10 times for each of the training cross folds. For the evaluation of ChatGPT, the metrics are applied directly to the output predictions of the same 100 tweet group prompts. The metrics are calculated across both prediction classes—increase and decrease in changes of financial figures—and are averaged for the total result value. For the LSTM models, the average values of all folds are used. All LSTM models perform fairly well and can predict long-term changes in quarterly financial metrics. The best LSTM model, *AmazonLSTM@10*, has an accuracy of 0.89 and an MCC of 0.77. Conversely, the worst-performing LSTM model, *TeslaLSTM@20*, has an accuracy of 0.70 and an MCC of 0.39, yet it still performs better than the best ChatGPT-based prediction, *AmazonChatGPT@10*, which has an accuracy of 0.67 and an MCC of 0.29. The least effective ChatGPT-based prediction, GoogleChatGPT@20, performs slightly worse than a random prediction, with an accuracy of 0.46 and an MCC of -0.08. Notably, as the tweet group size exceeds 10, the prediction performance decreases across all LSTM models. The prediction metrics with ChatGPT are considerably lower but, in several cases, higher than a random prediction and follow the same trend as LSTM prediction metrics regarding tweet group size, with Amazon data yielding the best predictions. Especially with larger groups, ChatGPT struggles to make accurate predictions. Additionally, the prediction performance varies significantly among different companies.

Table 5

Evaluation results

Model	Precision	Recall	F1-Score	Accuracy	MCC
AmazonLSTM@5	0.83	0.92	0.85	0.88	0.74
AmazonLSTM@10	**0.89**	**0.88**	**0.87**	**0.87**	**0.77**
AmazonLSTM@20	0.85	0.84	0.84	0.84	0.67
TeslaLSTM@5	0.81	0.79	0.79	0.79	0.59
TeslaLSTM@10	0.80	0.83	0.82	0.81	0.62
TeslaLSTM@20	0.71	0.67	0.69	0.70	0.39
GoogleLSTM@5	0.76	0.83	0.80	0.76	0.50
GoogleLSTM@10	0.85	0.78	0.81	0.82	0.64
GoogleLSTM@20	0.79	0.85	0.81	0.76	0.48
AmazonChatGPT@5	0.61	0.56	0.58	0.59	0.20
AmazonChatGPT@10	**0.67**	**0.66**	**0.67**	**0.67**	**0.29**
AmazonChatGPT@20	0.45	0.62	0.53	0.57	0.15
TeslaChatGPT@5	0.62	0.56	0.59	0.60	0.21
TeslaChatGPT@10	0.57	0.64	0.61	0.59	0.18
TeslaChatGPT@20	0.47	0.43	0.45	0.46	-0.07
GoogleChatGPT@5	0.50	0.60	0.55	0.57	0.13
GoogleChatGPT@10	0.50	0.75	0.60	0.62	0.28
GoogleChatGPT@20	0.46	0.48	0.47	0.46	-0.08

Source: Author's Compilation, 2023

4.3. Discussion

The evaluation results show that ChatGPT possesses predictive capabilities for quarterly financial metrics, which are significantly lower compared to those of the trained specialized LSTM models. LSTM models can predict different quarterly metrics fairly well, performing best with tweet groups of size 10. Notably, both ChatGPT and LSTM models exhibit a consistent trend, demonstrating similar changes in evaluation metrics across tweet groups; for instance, both exhibit poorer performance with tweet groups of size 20 compared to other sizes. Therefore, when it comes to predicting quarterly financial metrics based on tweet textual data and rudimentary prompts, ChatGPT cannot be relied upon to make accurate predictions. The results of the evaluation in this study could provide decision assistance for companies or further research attempting to utilize ChatGPT for similar tasks, given its current prominent use and reputation as a versatile tool. The results of this study also reflect and strengthen the hypotheses of the studies in the literature review, as they show that the performance of ChatGPT varies significantly between different use cases, datasets, and similar predictions. For the ChatGPT models in the literature review that perform significantly better than in this study, these models were not only prompt-based but also had certain assumptions, such as fixed trade rates for stock returns. The internal training data of ChatGPT could also play a role in its performance differences compared to LSTM, because LSTM was trained on data up to 2020, while the current ChatGPT model was trained on data up to 2023, during which major financial events, such as the COVID-19 pandemic and the invasion of Ukraine, took place. To improve the effectiveness of predictions made by ChatGPT, several strategies can be considered. It could be beneficial to experiment with a wider range of prompt inputs. Providing different information for the prediction task in the prompt, such as more metadata like the exact date of the tweet, tweet authors, correctly labeled tweet group samples, or even defining a custom ChatGPT agent for the prediction, can enable the LLM to delve deeper into the context of the task. However, it is important not to overcomplicate ChatGPT prompts, as ChatGPT can lose focus, which is especially problematic for complex, longer tasks like predicting a large number of tweets. Moreover, the quality of the text data is crucial, as shown by the gap in the performance of prediction metrics for ChatGPT across tweets from different companies. Therefore, different datasets or more sophisticated prefiltering of the data could improve prediction performance. The evolving nature of LLMs also plays a significant part in the discussion. Currently, LLMs are continuously being developed and changing rapidly; their user interfaces can significantly influence how they process user prompts, which in turn impacts prediction performance.

5. Conclusion

This study compared the predictive performance of specialized LSTM models and ChatGPT in forecasting positive or negative changes in quarterly financial metrics from social media text data. Utilizing standardized evaluation metrics—such as precision, recall, F1-score, accuracy, and MCC—the results demonstrated that LSTM models significantly outperform ChatGPT-based predictions. LSTM models can predict quarterly financial metric changes with up to 87% accuracy, significantly outperforming ChatGPT, which achieves a maximum accuracy of 67%. Therefore, at the current time, ChatGPT cannot be considered a reliable predictive tool for changes in quarterly financial metrics. Further findings also indicate a notable decline in predictive performance for both approaches as the size of the observed tweet groups increases, especially when exceeding ten tweets. Additionally, the predictive performance of both methods exhibited similar trends across various companies and financial metrics. The possible reasons for the lower performance of ChatGPT include the nature of its internal training data, which considers different aspects of the tweets compared to the LSTM models, and the design of its prompts, which influences complex and long prediction tasks with chaotic data like social media text. Despite ChatGPT's lower predictive performance, its evolving nature holds potential for future enhancement. Strategies like further prompt engineering and data quality improvement could boost its predictive capabilities. In summary, while traditional machine learning models, such as LSTM currently offer a reliable solution for predicting financial metric changes from tweet data, the continuous development of ChatGPT shows potential for future applications in this domain.

References

1. Albahli, S. et al. (2022). A Machine Learning Method for Prediction of Stock Market Using Real-Time Twitter Data. *Electronics*, Issue 11, 3414-3420.

2. Bahrini, A. et al. (2023). ChatGPT: Applications, Opportunities, and Threats. *arXiv,* Issue arXiv:2304.09103.

3. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big.* Virtual Event Canada, Association for Computing Machinery, 610–623.

4. Brownlee, J. (2018). *Long Short-Term Memory Networks With Python.* 1.5 ed. San Juan: Machine Learning Mastery.

5. Chen, L., Zaharia, M. & Zou, J. (2024). How Is ChatGPT's Behavior Changing Over Time?. *Harvard Data Science Review*, 6(2).

6. Chicco, D. & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, Issue 21(1), 6.

7. Ding, Y. et al. (2023). Integrating Stock Features and Global Information via Large Language Models for Enhanced Stock Return Prediction. Issue arXiv:2310.05627.

8. Doğan, M. et al. (2020). *Speculator and Influencer Evaluation in Stock Market by Using Social Media.* Altanta, IEEE, 4559-4566.

9. Google AI (2023). *Google AI.* [Online] Available at: https://ai.google/static/documents/google-about-bard.pdf [Accessed 01 12 2023].

10. Graves, A., Mohamed, A. & & Hinton, G. (2013). *Speech recognition with deep recurrent neural networks.* Vancouver, IEEE, 6645-6649.

11. Hochreiter, S. & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation,* Issue 9 (8), 1735-1780.

12. Kaggle (2023). *Tweets about the Top Companies from 2015 to 2020.* [Online] Available at: https://www.kaggle.com/datasets/omermetinn/tweets-about-the-top-companies-from-2015-to-2020 [Accessed 20 01 2024].

13. Lee, J., Stevens, N., Han, S. C. & Song, M. (2024). A Survey of Large Language Models in Finance. *arXiv,* Issue arXiv:2402.02315.

14. Lopez-Lira, A. & Tang, Y. (2023). Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. *arXiv,* Issue arXiv:2304.07619.

15. Memiş, E. et al. (2024). Comparative Study for Sentiment Analysis of Financial Tweets with Deep Learning Methods. *Applied Sciences,* 14(2).

16. Mikolov, T., Chen, K., Corrado, G. & & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv*, Issue arXiv:1301.3781.

17. Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective*.. Massachusetts: MIT Press.

18. OpenAI (2023). *OpenAI.* [Online] Available at: https://openai.com/blog/introducing-chatgpt-and-whisper-apis?ref=blog.orgspace.io [Accessed 28 07 2023].

19. Paszke, A. et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv* , Issue arXiv:1912.01703.

20. Pramod, B. S. & Mallikarjuna, S. P. M. (2021). Stock Price Prediction Using LSTM. *Test Engineering and Management,* Issue 83, 5246-5251.

21. Rehurek, R. & Sojka, P. (2010). *Software framework for topic modelling with large corpora.* Valletta, ELRA, 45-50.

22. Scikit-learn (2023e). *Metrics and scoring: quantifying the quality of predictions.* [Online] Available at: https://scikit-learn.org/stable/modules/model_evaluation.html [Accessed 13 04 2023].

23. Scikit-Learn (2024). *Scikit-Learn Cross-validation.* [Online] Available at: https://scikit-learn.org/stable/modules/cross_validation.html [Accessed 14 02 2024].

24. Statista (2024a). *Net sales revenue of Amazon from 1st quarter 2007 to 4th quarter 2023.* [Online] Available at: https://www.statista.com/statistics/273963/quarterly-revenue-of-amazoncom/ [Accessed 12 02 2024].

25. Statista (2024b). *Number of Tesla vehicles delivered worldwide from 1st quarter 2016 to 3rd quarter 2023.* [Online] Available at: https://www.statista.com/statistics/502208/tesla-quarterly-vehicle-deliveries/ [Accessed 13 02 2024].

26. Statista (2024c). *Market share of leading desktop search engines worldwide from January 2015 to January 2024.* [Online] Available at: https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/[Accessed 13 02 2024].

27. Sundermeyer, M., Schlüter, R. & Ney, H. (2012). *LSTM neural networks for language modeling.* Portland, ISCA.

28. Usmani, S. & Shamsi, J. (2023). LSTM based stock prediction using weighted and categorized financial news. *PLoS ONE,* 18(3).

29. Woodhouse, D. & Charlesworth, A. (2023). Can ChatGPT Predict Future Interest Rate Decisions?. *SSRN Electronic journal,* Issue 4572831.

30. Zhang, H. et al. (2023). Unveiling the Potential of Sentiment: Can Large Language Models Predict Chinese Stock Price Movements?. *arXiv,* Issue arXiv:2306.14222.