Proceedings of the 8th International Conference on Economic Management and Green Development
DOI: 10.54254/2754-1169/2024.MUR17868

Sentiment Analysis in Green Finance with LLMs

Tongfei Chen[1,a,*]

[1]Olin Business School, Washington University in St.louis, One Brookings Drive, St.louis, United States
a. c.tongfei@wustl.edu
**corresponding author*

Abstract: Green finance has gained global significance as governments and financial institutions emphasize sustainable investment. Understanding the sentiment of green finance reports can provide valuable insights into public perception, investor sentiment, and policy reception. This study uses three different models — FinBERT, GPT-3.5 Turbo, and GPT-4o -- to perform sentiment analysis on over 1000 reports obtained from the International Finance Corporation (IFC) website. To assess the accuracy of the models, this paper manually labeled the sentiment of the reports into three categories: Positive, Negative, and Neutral. We compared the models' outputs using standard metrics such as F1-score, Accuracy, Precision, and Recall. The findings indicate that GPT-3.5 Turbo outperforms the other models in terms of accuracy. GPT-4o shows superior performance compared to Finbert which trained on financial texts in extracting sentiment from general text. Even though FinBERT and GPT-4 have stronger financial text processing capabilities, GPT-3.5 Turbo can often capture the true intent and sentiment of the text more quickly and clearly, especially when trained on a relatively small text corpus. Its generalization and speed make it efficient for less complex financial tasks.

Keywords: Green Finance, Sentiment Analysis, FinBert, GPT, Large Language Models (LLMs).

1. Introduction

The rise of green finance has transformed the financial landscape, with investors and institutions prioritizing sustainability, environmental impact, and social governance (ESG) in their investment strategies. In this context, analyzing the sentiment of green finance reports is crucial for understanding how various stakeholders, including investors and policymakers, perceive and react to these initiatives.

In previous literature on sentiment analysis related to Green Finance, most studies have focused on company-based analyses of financial and sustainability reports, often emphasizing the relationship between the digital economy and specific company performance [1]. There has been limited use of large language models to analyze the overall outlook of the Green Finance industry.

Sentiment analysis, powered by large language models (LLMs), has emerged as a key tool in assessing public and institutional sentiment. Nevertheless, considering that LLMs, particularly general-purpose models like GPT, have a wide range of applications, they may still face challenges in fully understanding domain-specific terminology in the green finance sector. This is especially true when more detailed industry knowledge and contextual awareness are required, potentially affecting the accuracy of the sentiment analysis. This paper presents a comparative analysis of sentiment analysis models applied to green finance reports from the International Finance Corporation (IFC) website. The models used include FinBERT, GPT-3.5 Turbo, and GPT-4.0. While FinBERT is specifically fine-tuned on financial text, the GPT models are general-purpose LLMs. By manually categorizing reports into Positive, Negative, and Neutral sentiment categories, this research aims to determine which model provides the most accurate analysis in the context of green finance.

2. Related Work

2.1. Literature Review

Sentiment analysis in the finance sector has traditionally been used to gauge investor sentiment, market trends, and company performance through textual data. Early studies, such as Rational investor sentiment [2], applied Bayesian and related machine learning models, also using unsupervised learning to predict financial market turbulence and volatility has become a important method analyzing finance activities [3]. However, at that time, there were no comprehensive open-source tools or complete text databases available for analysis, requiring large amounts of training sets.

In the following years, to better optimize and utilize corpora, Pre-trained Language Models (PLMs) emerged, with the most prominent being BERT (Bidirectional Encoder Representations from Transformers) based on the Transformer architecture [4]. BERT revolutionized NLP by using a bidirectional encoder that pre-trains text, allowing it to consider both the left and right context of a sentence simultaneously. This innovation eliminated the need to build separate models for tasks such as question answering and language inference, significantly enhancing the efficiency and effectiveness of NLP tasks.

The success of BERT spurred further developments in Transformer-based text processing technologies. From 2020 to 2022, numerous innovations and applications of Language Models (LMs) appeared, including GPT series, T5, and RoBERTa. These models excelled not only in a wide array of natural language processing (NLP) tasks but also expanded into multimodal domains such as text generation, machine translation, sentiment analysis, and conversational systems. During this period, the financial sector, with its high demand for text analysis, led to the emergence of a specialized branch of LLMs known as FinLLMs. In this process, the novel FinLLMs model based on LLMs have been introduced, emphasizing that NLP Transformers demonstrate a significantly higher distinction between positive and negative sentiments in financial text compared to traditional decision trees and Naive Bayes classifiers. The key reason is that certain words, typically classified as negative in traditional corpora (e.g., "Debts"), tend to be more neutral in the financial market. Based on adjusted lexicons, numerous FinLLMs models have been developed. These models, such as FinBERT, FinMA, and FinGPT [5], were developed by fine-tuning language models on financial-specific corpora including financial reports, investment information, and market data.

However, the development of FinLLMs highlighted a gap in corpus availability for green finance, a growing field of interest within the financial sector. While FinLLMs have been trained extensively on financial information, there is limited availability of green finance-related corpora due to the large volume of textual data that LLMs require for effective model training. To date, the only FinLLM that has addressed this gap is the FinBERT-ESG model, designed specifically for analyzing environmental, social, and governance (ESG) factors. Currently, Fintech sector has begun utilizing machine learning and artificial intelligence to conduct quantitative monitoring and analysis of assets related to carbon emissions and sustainable resources in green finance [6]. However, specific financial language models based on machine learning for analyzing texts related to society and governance are still lacking. This situation may stem from the uncertainty regarding whether existing corpora and models are fully adaptable to the green finance domain. More targeted datasets and models are needed to meet the unique requirements of green finance analysis.

2.2. Models for Sentiment Analysis

2.2.1. FinBert

FinBERT is a transformer-based model specifically fine-tuned for financial sentiment analysis. It builds upon the BERT (Bidirectional Encoder Representations from Transformers) architecture and has been trained on a large corpus of financial documents, including financial reports, news, and earnings call transcripts. Its financial domain specificity allows it to capture nuanced sentiment in texts that other general-purpose models might miss.

2.2.2. GPT-3.5 Turbo

GPT-3.5 Turbo, a variant of the GPT-3 model developed by OpenAI, is designed to generate coherent text and perform various natural language processing tasks. While GPT-3.5 Turbo is not specifically trained on financial data, its massive training corpus and language generation capabilities make it a useful tool for general sentiment analysis [7]. GPT-3.5 Turbo is a more streamlined and higher-performance variant of GPT-3, designed to meet the diverse natural language processing needs across different domains and scales. Its improved efficiency and versatility make it particularly well-suited for specific domains like Green Finance, where accurate and efficient language understanding is crucial for analyzing financial reports, sustainability metrics, and environmental, social, and governance (ESG) factors.

2.2.3. GPT-4o

GPT-4o represents a significant advancement over its predecessor, GPT-3.5. It has been trained on a broader and more diverse set of text data and demonstrates improved capabilities in understanding context, generating accurate responses, and performing sentiment analysis across domains. GPT-4o's general understanding of text is particularly useful when applying sentiment analysis to domains like green finance, where explicit sentiment might not always be obvious.

2.3. Dataset Creation and Cleansing

The data used in this analysis consists of over 1000 green finance reports between 2018-2024 and articles scraped from the International Finance Corporation (IFC)(https://www.ifc.org/en/home). In Green Finance, the uniqueness of national economic models and diverse ecosystems is emphasized, and any policies and regulations related to green finance must be tailored to local conditions [8]. The text data varied significantly in length and style, with some reports containing technical financial language and others written in a more general public-facing style. The preprocessing of the text data involved the following steps:

2.3.1. Text Cleaning

Texts were cleaned by removing HTML tags, special characters, and irrelevant symbols. This step ensured that the models focused on the meaningful content without being influenced by noise.

The word cloud highlights several key terms that appear frequently in discussions or reports related to IFC. Here's a simple text analysis based on the most prominent words:

"Investment" and "Private Sector": These two terms are the most dominant, reflecting IFC's core focus on facilitating investments, particularly in the private sector. This aligns with IFC's mission to promote private sector development as a means of fostering sustainable economic growth.

"Emerging Markets": This term is also highly visible, indicating that IFC places a strong emphasis on developing financial solutions and promoting investments in emerging markets, where the need for infrastructure, development, and financial services is often greatest.

"Global" and "Development": These words reflect IFC's broader goal of driving global economic development through targeted investment in sectors like sustainable finance, climate action, and infrastructure.

Overall, the word cloud emphasizes IFC's role in facilitating investment and development in emerging markets, with a growing focus on sustainability, green finance, and climate initiatives, all while supporting private sector growth (see Figure 1).

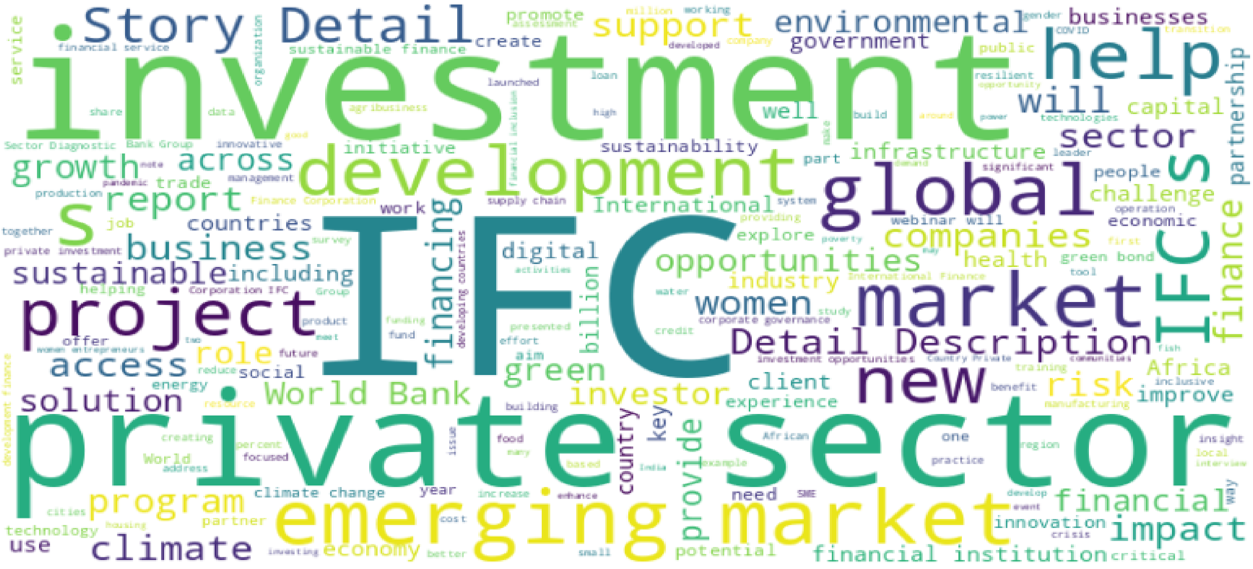

Figure 1: Word Cloud of IFC related Green Finance Report

2.3.2. Manual Sentiment Labeling

To establish a baseline for comparison, this paper manually labeled the reports into three sentiment categories: Positive, Negative, and Neutral. This labeling provided the ground truth for evaluating the accuracy of the models' outputs. To better help training the LLMs models,this paper defined the labels below:

A report categorized as "Positive" indicates optimism and confidence in the discussed projects, investments, or policies. In the context of Green Finance, this label would be applied to reports that highlight successful green initiatives, strong returns on sustainable investments, effective climate action strategies, or favorable policy developments that support sustainability goals. For IFC, a positive sentiment reflects favorable impacts on emerging markets, private sector growth, or successful partnerships that drive green economic growth. Reports on new green bonds generating high investor interest, or a renewable energy project that reduces carbon emissions and creates jobs, would fall under the positive sentiment category.

A report labeled as "Negative" would reflect pessimism or concerns about projects, financial instruments, or policies. In Green Finance, this may include reports discussing failed sustainability projects, financial losses due to climate risks, regulatory challenges, or ineffective environmental policies. For IFC, negative sentiment might indicate concerns about the feasibility of certain green investments, market instability in emerging economies, or risks that hinder green finance growth. A report discussing the underperformance of a sustainable investment fund or challenges in implementing climate change mitigation strategies due to political or economic barriers would be categorized as negative.

A report categorized as "Neutral" contains balanced information without a clear positive or negative tone. This could be purely informational, such as announcements of new sustainability policies, updates on ongoing projects, or discussions of regulatory changes without immediate impact on the market. In the context of Green Finance and IFC, neutral sentiment may apply to discussions that highlight potential opportunities and risks equally or provide factual overviews of green finance initiatives without a strong evaluative judgment. A report detailing a new regulatory framework for green bonds, where the potential impact is yet to be determined, would be considered neutral.

This sentiment labeling helps in assessing public and institutional sentiment towards Green Finance initiatives, guiding decision-making for policy makers, investors, and stakeholders in sustainable development.

2.4. Accuracy Evaluation Standard

To better determine the accuracy of the three models, this paper introduced standard accuracy testing methods, utilizing Accuracy, F1 Score, Precision, and Recall to assess model performance [9]. These metrics provide a comprehensive evaluation of how well each model performs on tasks such as classification or prediction, especially in natural language processing (NLP) tasks like sentiment analysis in Green Finance. Below is an explanation of each metric and its significance, as well as how they are calculated and applied.

2.4.1. Accuracy

Accuracy measures the proportion of correct predictions made by the model out of the total predictions. It is the most straightforward metric, calculated as:

$$\text{Accuracy} = \frac{Number\ of\ Correct\ Predictions}{\text{Total numbers of Predictions}} \tag{1}$$

Accuracy is useful when the dataset is balanced, meaning that all categories (e.g., positive, neutral, negative sentiments) are equally represented. It provides a quick and simple way to assess overall model performance. However, accuracy can be misleading in cases of imbalanced datasets, where one class dominates. In such cases, even if the model performs poorly on minority classes, the accuracy might still appear high.

2.4.2. Precision

Precision is the ratio of correctly predicted positive instances to the total instances that were predicted as positive. It focuses on the accuracy of the positive predictions made by the model and is calculated as:

$$\text{Precision} = \frac{True\ Positives(TP)}{\text{True Positives(TP)+ False Positives(FP)}} \tag{2}$$

Precision is particularly useful in scenarios where the cost of false positives is high. For example, in Green Finance, when determining if an investment is environmentally sustainable (positive class), having a high precision ensures that only truly sustainable investments are identified. Precision is important when you want to minimize incorrect positive classifications.

2.4.3. Recall

Recall measures the model's ability to correctly identify all positive instances. It is calculated as:

$$\text{Recall} = \frac{True\ Positives(TP)}{\text{True Positives(TP)} + \text{False Negatives(FN)}} \tag{3}$$

Recall is critical in situations where identifying all positive instances is important, even at the expense of some false positives. For example, in identifying all relevant ESG factors in Green Finance, recall ensures that the model captures every possible positive instance (e.g., sustainable investments), even if it misclassifies some negative instances. High recall is crucial when missing a positive prediction (false negative) is more costly.

2.4.4. F1 Score

The F1 Score is the harmonic mean of precision and recall, and it provides a balance between the two metrics. It is particularly useful when you need to strike a balance between precision and recall. The formula is:

$$\text{F1 Score} = 2 \ * \ \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The F1 score is highly valuable in cases of imbalanced datasets where a high precision or high recall alone may not fully capture the model's performance. By combining both, the F1 score offers a more holistic view of the model's ability to make accurate predictions. For example, in Green Finance, balancing the model's ability to accurately identify both positive and negative sustainability signals is key, and the F1 score helps evaluate this balance.

3. Result

In the experiment, this paper compared GPT-based models and FinBERT as a classic example of FinLLMs (Financial Large Language Models). By continuously optimizing and training the models, we obtained Fig 2, which illustrates the distribution of sentiment analysis results across different large language models. The results reveal that both the FinBERT model and GPT-4o (which named gpt_sentiment) provided a significant amount of positive sentiment evaluations for Green Finance-related reports, closely aligning with the sentiment distribution obtained from manual classification. In contrast, GPT-3.5 Turbo tended to assign a higher proportion of negative sentiment evaluations (see Figure 2).

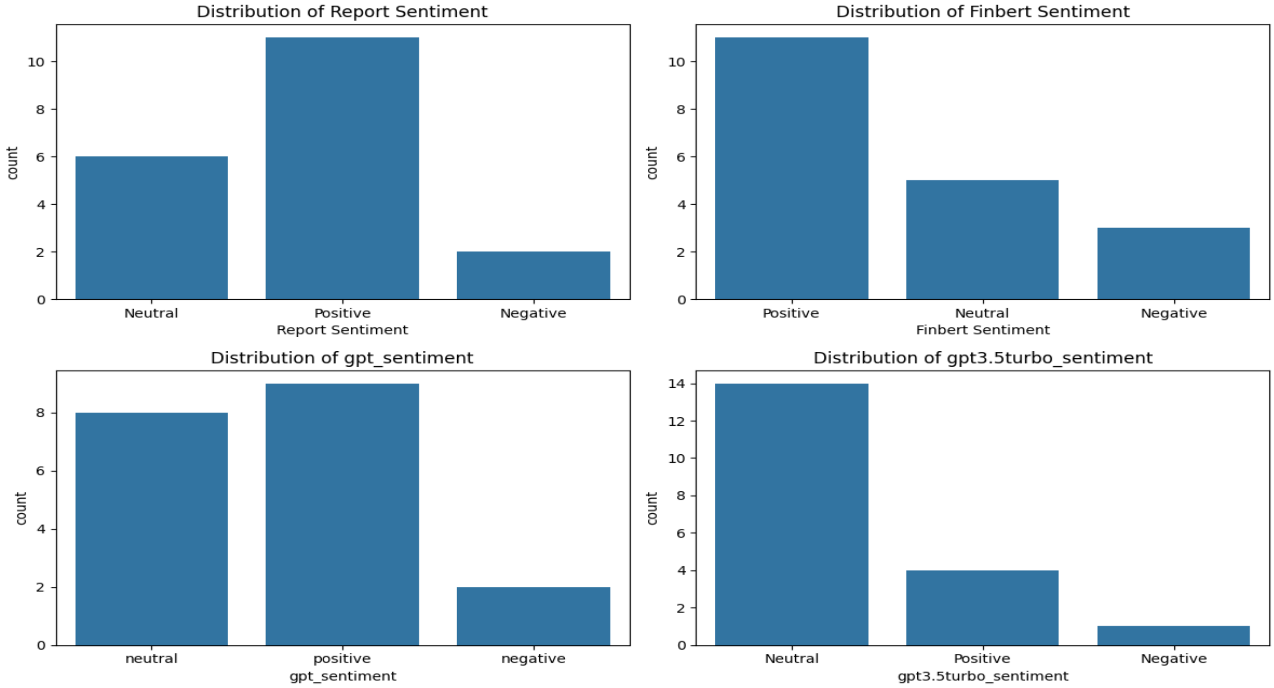

Figure 2: distribution of sentiment analysis results across different large language models

This discrepancy may stem from the fact that FinBERT is specifically fine-tuned for financial text, allowing it to better understand and identify positive indicators in Green Finance contexts. Similarly, GPT-4o is designed to handle complex and nuanced texts, which likely enables it to capture a more balanced and positive sentiment. On the other hand, GPT-3.5 Turbo emphasizes efficiency and text simplification, which may result in a higher sensitivity to negative aspects or more straightforward interpretations, leading to an overestimation of negative sentiment. This highlights how the underlying training data and model design impact sentiment analysis in domain-specific contexts like Green Finance.

From the perspective of accuracy analysis, Table 1 presents a different result. Despite the fact that the distribution of Positive and Negative sentiments in GPT-3.5 Turbo significantly diverges from the manually classified results, the accuracy of this model is actually higher than the other two (FinBERT and GPT-4o). This phenomenon may be attributed to the fact that GPT-3.5 Turbo requires less specific textual analysis compared to the other models.

Table 1: Model Accuracy

Model	Accuracy	Precision	Recall	F1-Score
Finbert	0.37	0.36	0.37	0.36
GPT	0.42	0.46	0.42	0.44
GPT-3.5 Turbo	0.47	0.69	0.47	0.47

When processing inputs, GPT-3.5 Turbo may rely less on intricate, context-heavy analysis and instead focus on quicker, more straightforward interpretations. This allows it to better align with the logic of human analysis, which often emphasizes simplicity and directness in understanding text inputs. While FinBERT and GPT-4 are designed to handle complex language structures and nuanced meanings, which could be advantageous in more detailed tasks, GPT-3.5 Turbo excels in faster, less complicated processing that might mirror the more intuitive judgment humans' use in certain scenarios, resulting in a higher accuracy score in this context. The original paper on GPT-3, discusses the balance between model size and performance, noting that smaller models can often perform tasks more efficiently, especially when the task requires less complexity [10]. Additionally, another research highlights that more specialized models like FinBERT are fine-tuned for detailed, domain-specific tasks but may not always outperform general models in simpler or broader tasks due to their complexity [11].

4. Conclusion

In this study, this paper compared the performance of various language models, including FinBERT, GPT-4o, and GPT-3.5 Turbo, in conducting sentiment analysis on Green Finance-related reports. Our findings reveal some interesting contrasts. While both FinBERT and GPT-4o aligned closely with human-labeled sentiment distributions, particularly by assigning more positive evaluations to Green Finance content, GPT-3.5 Turbo tended to label more reports as negative. However, when we analyzed the models in terms of accuracy, the results, indicated that GPT-3.5 Turbo actually achieved the highest accuracy, despite its differing sentiment distribution.

Overall, the study highlights the importance of considering not just sentiment distribution, but also accuracy and the underlying model architecture when selecting a model for domain-specific tasks like Green Finance sentiment analysis. While FinBERT and GPT-4o are more suited to complex, nuanced analysis, GPT-3.5 Turbo may offer a more efficient solution in contexts where quick, straightforward analysis is required.

While this study highlights the strengths of FinBERT and GPT-series models in green finance sentiment analysis but acknowledges several limitations. Other financial LLMs like FinMA and FinGPT were not explored due to computational constraints and limited text data, though these models may provide additional insights or even outperform those used here. Future research will focus on incorporating these models and testing them on a more diverse dataset. Additionally, human-labeled sentiment may introduce biases, particularly in green finance where sentiment is subtle and hard to categorize, emphasizing the need for larger datasets and more advanced labeling techniques in future studies.

References

[1] Appiah-Kubi, E., Koranteng, F. O., Dura, C. C., Mihăilă, A. A., Drigă, I., & Preda, A. (2024). Green financing and sustainability reporting among SMEs: The role of pro-environmental behavior and digitization. Journal of Cleaner Production, 143939.

[2] Gerber , A., Hens, T., & Vogt, B. (2010). Rational investor sentiment in a repeated stochastic game with imperfect monitoring. Journal of Economic Behavior & Organization, 76(3), 669-704.

[3] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160.

[4] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[5] Lee, J., Stevens, N., Han, S.C., & Song, M. (2024). A Survey of Large Language Models in Finance (FinLLMs). ArXiv, abs/2402.02315.

[6] Wang, C., Zheng, C., Chen, B., & Wang, L. (2024). Mineral wealth to green growth: Navigating FinTech and green finance to reduce ecological footprints in mineral rich developing economies. Resources Policy, 94, 105116.

[7] Johnson, D., Goodman, R., & Patrinely, J. (2023). Assessing the accuracy and reliability of AI-generated medical responses: An evaluation of the Chat-GPT model. Research Square.

[8] Nedopil, C., Dordi, T., & Weber, O. (2021). The nature of global green finance standards—Evolution, differences, and three models. Sustainability, 13(7), 3723.

[9] Poggio, T., & Girosi, F. (1990). Networks for approximation and learning. Proceedings of the IEEE, 78(9), 1481-1497

[10] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J.,

Winter, C., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

[11] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2020). XLNet: Generalized autoregressive pretraining for language understanding. Advances in Neural Information Processing Systems, 32.