C-SHAP: Collocation-Aware Explanations for Financial NLP

Martina Menzio
Intesa Sanpaolo
University of Milano-Bicocca
martina.menzio@intesasanpaolo.com

Elisabetta Fersini
University of Milano-Bicocca
elisabetta.fersini@unimib.it

Davide Paris
Intesa Sanpaolo
University of Milano-Bicocca
davide.paris@intesasanpaolo.com

Abstract

Understanding the internal decision-making process of NLP models in high-stakes domains such as the financial sector is particularly challenging due to the complexity of domain-specific terminology and the need for transparency and accountability. Although SHAP is a widely used model-agnostic method for attributing model predictions to input features, its standard formulation treats input tokens as independent units, failing to capture the influence of collocations that often carry non-compositional meaning, instead modeled by the current language models. We introduce C-SHAP, an extension of SHAP that incorporates collocational dependencies into the explanation process to account for word combinations in the financial sector. C-SHAP dynamically groups tokens into significant collocations using a financial glossary and computes Shapley values over these structured units. The proposed approach has been evaluated to explain sentiment classification of Federal Reserve Minutes, demonstrating improved alignment with human rationales and better association to model behaviour compared to the standard token-level approach. [1]

1 Introduction

In recent years, the adaptation of language models, particularly BERT-like (Devlin et al., 2019) architectures, to the financial domain has transformed text analysis, allowing a more nuanced understanding of financial documents. However, while these models excel at capturing the complex relationships within textual data, explaining their decisions remains a significant challenge. SHAP (Lundberg and Lee, 2017) (SHapley Additive exPlanations) has emerged as a popular method to interpret model outputs by quantifying the contribution of individual features, such as words or tokens, to the overall prediction. However, SHAP considers each feature independently, failing to explicitly model the influence of collocations that often carry non-compositional meaning, which are instead effectively captured by modern language models. In this paper, we investigate the enhancements of SHAP-based explainability methods when applied to BERT-like models in the financial domain, with a particular focus on sentiment classification using the FinBERT model (Liu et al., 2021). In financial texts, where meaning is often conveyed through domain-specific collocations and non-compositional expressions, understanding the internal decision-making process of language models requires methods that go beyond token-level attributions. To address this issue, we develop a novel extension of SHAP that accounts for collocational dependencies by grouping tokens into meaningful word combinations, guided by a financial glossary. In particular, we introduce C-SHAP (Collocation-based SHAP) as an extension of SHAP that incorporates collocations into the explanation process to account for word combinations that are peculiar in the financial sector.

2 Related Works

In the foundational paper (Lundberg and Lee, 2017) Scott Lundberg and Su-In Lee introduce SHAP (SHapley Additive exPlanations) as a unified framework for interpreting model predictions. SHAP leverages concepts from cooperative game theory, specifically Shapley values, to attribute the output of a model to its input features in a manner that satisfies certain desirable properties. Although SHAP has been widely applied across domains, its application to text has focused primarily on token-level attributions (Kokalj et al., 2021; Chen et al., 2023; Amara et al., 2024), which do not always align with human interpretability, especially in cases involving multi-word expressions or domain-specific terminology.

[1]The views and opinions expressed are those of the authors and do not necessarily reflect the views of Intesa Sanpaolo, its affiliates or its employees.

Proceedings of Recent Advances in Natural Language Processing, pages 711–717
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-082

Despite this limitation, the integration of SHAP and BERT (Bidirectional Encoder Representations from Transformers) in the financial domain has become an increasingly prominent area of research. In particular, BERT has been widely adopted in various financial applications, such as sentiment analysis (Bhadouria et al., 2025), credit risk assessment (Mienye et al., 2024), and stock price prediction (jun Gu et al., 2024), due to its ability to capture nuanced language patterns in financial texts, coupled with SHAP to explain the models' predictions. FinBERT (Liu et al., 2021), i.e. a BERT-based model trained for financial sentiment analysis, has also been interpreted using SHAP to validate model behavior against domain expectations. Some effort has been devoted to combine SHAP with hierarchical structures (Chen et al., 2020), but the modeling of multi-word financial terms remains underexplored. Our work addresses this gap by extending SHAP to consider significant collocations, which often convey non-compositional meaning vital in financial texts.

3 Original SHAP

SHAP values are used to understand which tokens have a greater impact on the prediction of the model: the positive SHAP values have a positive effect on predicting the considered label, i.e. they push the classification of the given sentence towards that label, while negatively contributing features influence the model's output in a direction opposite to the target label.

In the context of text classification, each word or token in the text can be considered a feature. The SHAP value ϕ_{t_i}, for a token t_i of a word w_i in sentence i, represents its contribution to the difference between the prediction of the model for the given text and the baseline prediction. The SHAP values are computed on the following principles:

- Efficiency: The sum of the SHAP values for all the features of an instance (in this case, for all the tokens composing a sentence) equals the difference between the model's prediction for the instance and the average prediction over all instances.
- Symmetry: If two features contribute equally to all possible subsets of features, they receive the same SHAP value.
- Dummy: Words that do not affect the prediction have a SHAP value of zero.
- Additivity: For models combined linearly, the SHAP values of the combined model are the sum of the SHAP values of the individual models.

From a practical point of view, computing the exact SHAP values requires evaluating the model on all possible subsets of features, which is computationally infeasible for texts with many words. To address this, approximation methods such as Kernel SHAP are used. Kernel SHAP estimates SHAP values by sampling subsets of features and weighing them appropriately. In text classification, this involves:

1. *Sampling Subsets*: Randomly selecting subsets of words from the text
2. *Perturbation*: Creating perturbed versions of the text by masking tokens or words
3. *Clustering*: Create clusters of tokens (in order to improve efficiency, semantic coherence and quality of explanations) to group related features and mask them together during coalition creation
4. *Model Evaluation*: Passing these perturbed texts through the model to observe changes in predictions
5. *Weighting and Regression*: Using the observed changes to solve a weighted linear regression that estimates the SHAP values

All of these steps have the main goal of balancing computational efficiency with the need for accurate feature attribution. Since each token is considered independent with respect to the others, it is necessary to sum up the SHAP values obtained by tokenizing a single word. Given a word w_i composed of tokens $t_{i_1}, ..., t_{i_n}$, and the SHAP values $\phi_{t_{ij}}$ estimated for each token $t_{ij} \in w_i$, the SHAP value of w_i is estimated as:

$$\Phi_i = \sum_{j=1}^{n} \phi_{t_{ij}} \tag{1}$$

4 C-SHAP

In order to take into account specific (financial) collocations, we reformulate the original SHAP problem. In particular, we revised the two main steps of the original Kernel SHAP, i.e. *Perturbation* and *Clustering*.

Perturbation The core idea is that tokens that compose each collocation have to be considered together, i.e. in the masking phase, they have to be masked or unmasked always together.

More formally, let $x = (x_1, x_2, \ldots, x_M) \in \mathcal{X} \subset \mathbb{R}^M$ be the original input vector with M atomic features (e.g., words/tokens) and $G = \{g_1, g_2, \ldots, g_K\}$ be a partition of the index set $\{1, 2, \ldots, M\}$, where each $g_k \subset \{1, \ldots, M\}$ is a collocation group. Each g_k corresponds to each collocation and $x^{(G)} = (x_{g_1}, x_{g_2}, \ldots, x_{g_K})$, where each $x_{g_k} = \{x_j \mid j \in g_k\}$.

We define a coalition vector $z \in \{0,1\}^K$, where each $z_k = 1$ means group g_k is included and $z_k = 0$ means it is excluded. Let $h_z : \{0,1\}^K \to \mathcal{X}$ be the mapping that reconstructs the full input $\tilde{x} = h_z(z)$, such that:

$$\tilde{x}_j = \begin{cases} x_j, & \text{if } \exists k \text{ such that } j \in g_k \text{ and } z_k = 1 \\ x_j^{\text{baseline}}, & \text{otherwise} \end{cases} \quad (2)$$

where if $z_k = 1$ then x_{gk} is included, while if $z_k = 0$ then the values in x_{gk} are replaced with a baseline $x_j^{baseline}$ (e.g., zero embedding).

Now, let $f : \mathcal{X} \to \mathbb{R}$ be the prediction function. The C-SHAP value ϕ_k for the collocation group g_k is:

$$\phi_k = \sum_{S \subseteq G \setminus \{g_k\}} \frac{|S|!(K - |S| - 1)!}{K!} \times \\ \times \left[f(h_z(S \cup \{g_k\})) - f(h_z(S)) \right] \quad (3)$$

This quantifies the marginal contribution of the collocation group g_k to the prediction $f(x)$.

The C-SHAP values are obtained by modifying the *Perturbation* phase: every time a token that composes a collocation is unmasked, we unmask all the other masked tokens of the same collocation, so that at each iteration the whole collocation we are considering is fully either masked or unmasked. Once we obtain the C-SHAP value for each token, the same procedure as in the original SHAP can be used, i.e. we sum up the C-SHAP values of the tokens belonging to the same word as reported in Equation 1. This method, where only the perturbation step has been revised (i.e., the clustering step corresponds to the original one), will be denoted as **C-SHAP (P)**.

Clustering To emphasize that tokens belonging to a collocation should be considered together, we also revised the *Clustering* phase, whose main goal is to mask blocks of correlated features instead of generating random masks on individual features. In particular, we ensure that tokens belonging to the same collocations are clustered together first, followed by individual tokens, and finally the entire text.

More formally, let $x = (x_1, x_2, \ldots, x_M)$ be an input sequence composed of M tokens, $\phi = (\phi_1, \phi_2, \ldots, \phi_M)$ be the corresponding SHAP values, and $C = \{c_1, c_2, \ldots, c_K\}$ be a set of collocation groups, where each $c_k \subset \{1, \ldots, M\}$ and $c_i \cap c_j = \emptyset$. The goal is to build a binary hierarchical clustering tree $\mathcal{T}$ over the tokens x_i, subject to merging priorities: (1) tokens belonging to the same collocation c_k must be merged first, (2) individual non-collocation tokens merge next, and (3) the entire sequence of tokens is finally clustered.

To this purpose, we define a clustering state $\mathcal{P}_t$ as a partition of $\{1, \ldots, M\}$ at time t. For candidate merges $A, B \subseteq \{1, \ldots, M\}$:

$$\mathcal{H}(A, B) = \begin{cases} 0, & \text{if } A \cup B \subseteq c_k, \forall c_k \in C \\ 1, & \text{if } A \cup B \subseteq \{1, \ldots, M\} \setminus \bigcup_{k=1}^{K} c_k \\ 2, & \text{otherwise} \end{cases} \quad (4)$$

A clustering cost function $\pi(A, B)$ is therefore defined to include a penalty term to respect the hierarchy of the merging priorities:

$$\pi(A, B) = \lambda^{\mathcal{H}(A,B)} \cdot d(A, B) \quad (5)$$

where and $\lambda > 1$ is a scaling constant enforcing merge priority, and $d(A, B)$ is the absolute SHAP difference defined as:

$$d(A, B) = \left| \frac{1}{|A|} \sum_{i \in A} \phi_i - \frac{1}{|B|} \sum_{j \in B} \phi_j \right| \quad (6)$$

Finally, as before, once the C-SHAP value for each token is obtained, we sum up the coefficients of the tokens belonging to the same word. This approach, where both the perturbation and the clustering step have been modified, will be denoted as **C-SHAP (P+C)**.

5 Experimental Settings

Financial Glossary In order to use real-world collocations that are relevant for the financial domain, we download a financial glossary, keeping just the terms, without their explanations. At this point, we divide the abbreviations from their long forms, e.g., we start from *APR (Annual Percentage*

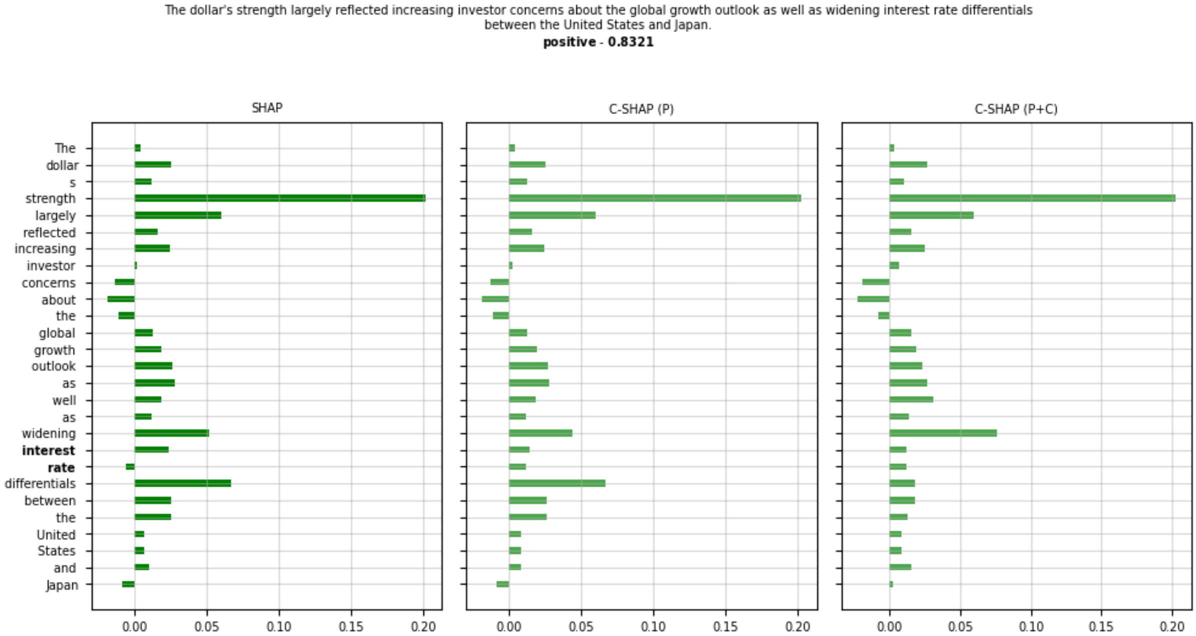

Figure 1: Example of plot for SHAP values computed with different methodologies.

Rate), getting separately *APR* and *Annual Percentage Rate*. Finally, we obtain plural forms starting from singular terms, obtaining a final dataset composed of 492 financial collocations.

FinBert To demonstrate the effect of the proposed C-SHAP, we adopted the FinBERT model to predict the sentiment of the FED communications over time presented in (Menzio et al., 2024). The model classifies each sentence belonging to any Federal Reserve System (FED) minutes and speeches mentioning a given Forex as positive, negative and neutral. In particular, we maintain only those sentences that contain at least one financial collocation and that mention the term *dollar*.

Model Comparison In order to evaluate the explainability, we compared:

- ***SHAP***: the original Kernel SHAP, where both Perturbation and Clustering correspond to the standard implementation;
- ***C-SHAP (P)***: Kernel SHAP where only the masking strategy in the Perturbation step is constrained to deal with collocations, while maintaining the original Clustering;
- ***C-SHAP (P+C)***: Kernel SHAP where both masking and clustering in the Perturbation step have been revised to deal with collocations.

6 Results

In order to verify that the proposed C-SHAP better captures the impact of tokens and collocations when FinBert makes the prediction, we have quantitatively and qualitatively evaluated the obtained coefficients. In Figure 1, we report one of the processed sentences:

> *The dollar's strength largely reflected increasing investor concerns about the global growth outlook as well as widening interest rate differentials between the United States and Japan.*

which sentiment predicted by FinBERT is *positive* with a probability of 83.21%.

We can easily note that the original SHAP values associated with the terms that contribute positively towards the positive predicted class do not change significantly, i.e. "*strength*" and "*widening*" still have the highest impact. On the other hand, the SHAP values associated with the terms belonging to the collocation suffer changes, in particular:

- ***SHAP***: *interest* contributes positively to the positive sentiment prediction while *rates* is reported to have a negative influence, denoting a counterintuitive explanation;
- ***C-SHAP (P)***: *interest* and *rates* are characterized by the same positive direction, showing

a more natural view that reflects what is expected by human beings when considering explanations of sentences with multi-word expressions;

- ***C-SHAP (P+C)***: *interest* and *rates* are characterized by the same positive direction, they assume the same value and the contribution of other relevant terms (e.g. *widening*) has been correctly increased to explain the positive class prediction.

To summarize, the value of the term *interest* is high in the original SHAP estimation, but decreases in the two C-SHAP variants, confirming that it is itself a word that positively contributes to the predicted class, but loses some of its nuance when it is in the context of *interest rate*. In Table 1, we report the computed values per word in the example sentence, emphasizing the positive effect of considering collocations in a real scenario. In particular, in green are highlighted the SHAP values associated with words that contribute positively from a financial point of view (i.e. *strength*, *increasing* and *widening*) and in blue those of the considered collocation (*interest rates*).

To perform a more in-depth analysis, we performed a human evaluation. In particular, we engaged three experts in the financial domain, focusing in particular on **SHAP** and **C-SHAP (P+C)** evaluation. For each sentence to be evaluated, we provided both a summary plot and the importance values estimated according to the two methodologies. In particular, 75 sentences containing the collocation *interest rates* or *interest rate* and mentioning either USD or dollars have been selected and qualitatively evaluated. For each sentence, we computed the FinBERT sentiment and proposed a comparative evaluation between SHAP and C-SHAP (P+C) attribution to each expert. More precisely, each expert had to verify:

1. If both terms *interest* and *rate(s)*, in each considered sentence, are characterized by the same direction and magnitude;

2. If the C-SHAP values are equal to or better than the original SHAP to explain the FinBERT prediction for positive, negative or neutral sentences.

In most cases, participants agreed that the C-SHAP (P+C) improved the explainability of the FinBERT sentiment predictions. Specifically, the terms identified as most influential for the predicted sentiment were more consistent with domain-expert expectations in the financial context. The following summarizes the expert evaluation conducted on the set of 75 proposed sentences:

Token	SHAP	C-SHAP (P)	C-SHAP (P+C)
The	0.003948897	0.003948897	0.003948897
dollar	0.025454779	0.025454779	0.026967511
s	0.012090307	0.012090307	0.010853442
strength	0.202259242	0.202259242	0.202259242
largely	0.059940234	0.059940234	0.059940234
reflected	0.016159069	0.016159069	0.015453395
increasing	0.024084724	0.024084724	0.024790398
investor	0.001884158	0.001884158	0.006974796
concerns	-0.013412408	-0.013412408	-0.018503046
about	-0.018805247	-0.018805247	-0.022663266
the	-0.011388901	-0.011388901	-0.007530882
global	0.012222551	0.012222551	0.0160888678
growth	0.018790853	0.018790853	0.019432703
outlook	0.026559501	0.026559501	0.023421457
as	0.027899526	0.027899526	0.027052584
well	0.018600295	0.018600295	0.031671853
as	0.011720734	0.011720734	0.013915665
widening	0.051761419	0.043687571	0.07646786
interest	0.023654868	0.014355783	0.011946663
rate	-0.00580821	0.011564722	0.011946663
differentials	0.06677462	0.06677462	0.017884071
between	0.025588451	0.025588451	0.018577721
the	0.025588451	0.025588451	0.013350865
United	0.00689604	0.00808703	0.008385395
States	0.00689604	0.00808703	0.008385395
and	0.010469012	0.00808703	0.015912918
Japan	-0.008443049	-0.008443049	0.002357103

Table 1: SHAP values computed for the words in the example sentence with different methodologies.

- In 16 out of 75 cases, the interpretations produced by SHAP and C-SHAP (P+C) are equivalent.

- In the remaining 59 sentences, differences between the two methods are observed. Specifically:
  - In 49 of these 59 cases, C-SHAP (P+C) are preferred over SHAP by 3 annotators.
  - In 54 of the 59 cases, C-SHAP (P+C) are judged superior by at least 2 out of 3 annotators. In only 5 cases, SHAP is favored by the majority.

Among the 54 cases where C-SHAP (P+C) is considered better than SHAP, 21 exhibited a substantial shift in the importance of non-collocation tokens, therefore enhancing the overall interpretability of the explanations. This indicates that the proposed method not only redistributes attention among relevant tokens but also surfaces previously underemphasized elements that are relevant to the model's decision-making process. Such shifts are crucial, as they can provide more comprehensive and human-aligned explanations, especially in domains where interpretability is essential for trust and accountability.

In addition to the qualitative evaluation, we also performed a comparative analysis of the most important words identified by both SHAP and C-SHAP (P+C). We report in Table 2 and Table 3 the top-10 words that contribute more to the positive and negative labels[2]. The results show that both methods are effective in identifying the most relevant terms, which align with general expectations regarding their association with positive and negative sentiment. In fact, despite small differences in the importance scores associated with each token, the top-ranked words remained consistent between the two methods. This equivalence in the highest-ranked token suggests that C-SHAP (P+C) preserves the core attribution signal of the original SHAP method while enhancing the interpretability of the remaining token contributions. This is a signal that C-SHAP (P+C) revises the original interpretive structure, ensuring that the most impacting tokens remain identifiable globally while offering a more context-aware explanation at the sentence level.

WORD	SHAP	C-SHAP (C+P)
strengthened	0.323008513	0.323008513
rose	0.263572386	0.263556974
increased	0.184566060	0.178510886
gain	0.175477574	0.175477574
appreciated	0.167780896	0.164011725
lifted	0.166353529	0.166353529
improving	0.152232815	0.152232815
supported	0.151366703	0.151366703
strength	0.134751284	0.132725641
up	0.134126606	0.134126606

Table 2: Top-10 of words that mostly contribute to the **positive class** label. The SHAP and C-SHAP values have been estimated as the mean over the sentences where the term occurs.

The complete dataset, including sentences, references to the Federal Reserve minutes and importance scores, is available at the following link https://github.com/MIND-Lab/C-SHAP.

[2]For each term, we estimated the mean of the corresponding relevance score across sentences.

WORD	SHAP	C-SHAP (C+P)
pessimistic	0.446069581	0.446069581
declined	0.445957771	0.433646462
fell	0.418518330	0.418518330
depreciated	0.303336474	0.303336474
down	0.296005961	0.295835777
decline	0.291008095	0.291008095
mixed	0.267927359	0.26792735
weaker	0.221566260	0.221566260
declining	0.213053765	0.213053765
depressed	0.176661988	0.176661988

Table 3: Top-10 of words that mostly contribute to the **negative class** label. The SHAP and C-SHAP values have been estimated as the mean over the sentences where the term occurs.

7 Conclusions and Future Work

In this paper, we addressed the challenge of interpreting the decision-making processes of NLP models in high-stakes domains such as finance, where domain-specific terminology and non-compositional language structures are prevalent. While SHAP remains a standard model-agnostic method for attributing predictions to input features, its token-level independence assumption limits its effectiveness in capturing the semantic impact of collocations, i.e., structures that are often critical in financial language and well-represented in large language models. To overcome this limitation, we proposed C-SHAP, an extension of SHAP that integrates collocation constraints into the explanation process. When applied to the sentiment classification task of Federal Reserve Minutes, C-SHAP demonstrated improved interpretability and stronger alignment with expert expectations. These results highlight the potential of incorporating domain-aware linguistic structures into explanation methods for enhancing the accountability of NLP systems in specialized domain. Future work will be devoted to enhance the C-SHAP estimation by incorporating probabilistic information about word sequences to produce more accurate linguistically plausible explanations. We also plan to explore a broader range of explainability techniques (e.g. LIME or Integrated Gradients) as well as alternative financial NLP models. Finally, integrating context-aware collocation detection methods that can learn collocations directly from large corpora could reduce dependence on manually curated glossaries and improve adaptability across subdomains.

References

Kenza Amara, Rita Sevastjanova, and Mennatallah El-Assady. 2024. SyntaxShap: Syntax-aware explainability method for text generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4551–4566, Bangkok, Thailand. Association for Computational Linguistics.

Aashi Singh Bhadouria, Anamika Ahirwar, and Mahendra Singh Panwar. 2025. Analyzing financial sentiments using bert model: A deep dive into market perception and investor behavior for informed investment decisions. In *Utilizing AI and Machine Learning in Financial Analysis*, pages 213–244. IGI Global Scientific Publishing.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. 2020. Generating hierarchical explanations on text classification via feature interaction detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5578–5593, Online. Association for Computational Linguistics.

Hugh Chen, Ian C Covert, Scott M Lundberg, and Su-In Lee. 2023. Algorithms to estimate shapley value feature attributions. *Nature Machine Intelligence*, 5(6):590–601.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Wen jun Gu, Yi hao Zhong, Shi zun Li, Chang song Wei, Li ting Dong, Zhuo yue Wang, and Chao Yan. 2024. Predicting stock prices with finbert-lstm: Integrating news sentiment analysis. In *Proceedings of the 2024 8th International Conference on Cloud and Big Data Computing*, pages 67–72.

Enja Kokalj, Blaž Škrlj, Nada Lavrač, Senja Pollak, and Marko Robnik-Šikonja. 2021. BERT meets shapley: Extending SHAP explanations to transformer-based classifiers. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 16–21, Online. Association for Computational Linguistics.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2021. Finbert: a pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Martina Menzio, Davide Paris, and Elisabetta Fersini. 2024. Unveiling currency market dynamics: Leveraging federal reserve communications for strategic investment insights. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing@ LREC-COLING 2024*, pages 94–102.

Ebikella Mienye, Nobert Jere, George Obaido, Ibomoiye Domor Mienye, and Kehinde Aruleba. 2024. Deep learning in finance: A survey of applications and techniques. *AI*, 5(4):2066.