Analysis of the relationship between traditional markets and commodities trade

Ana Todorovska[1], Stefan Milev[1], Irena Vodenska[2], Lubomir Chitkushev[3] and Dimitar Trajanov[1,3]
[1] *Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, N. Macedonia*
[2] *Administrative Sciences Department, Boston University, Metropolitan College, USA*
[3] *Department of Computer Science, Boston University, Metropolitan College, USA*

ana.todorovska@finki.ukim.mk, stefan.milev@students.finki.ukim.mk,
vodenska@bu.edu, ltc@bu.edu, dimitar.trajanov@finki.ukim.mk

***Abstract*—In a global world, no country, market, or economy is isolated and can function independently. Interconnectivity is a fundamental feature of economic systems, including both traditional financial markets and novel markets. This study aims to explore the relationships between traditional markets and commodities trade. We develop a methodology for analyzing the relationships between five stock market indexes (S&P500, Dow Jones, BSE, Hang Seng, FTSE) and three commodities (crude oil, natural gas, gold), based on multimodal publicly available datasets incorporating structured numerical and unstructured news and social network data. To find the existence of directional associations, we develop an Explainable ML model that first learns the dependencies between different assets and then explains them in a form understandable by humans. We apply our methodology to analyze connectivity networks between the assets and discuss our conclusions.**

***Index Terms*—stock market indexes, gold, crude oil, natural gas, networks, NLP, sentiment**

I. Introduction

We live in a world that does not allow for economic isolation, as globalization leaves no country, market or economy segregated, evident in both the development and structure of markets. This interdependence is fundamental in global economic systems, including macroeconomic trends, traditional markets, and novel markets [1].

Classical economic indicators are used to study macroeconomic trends and traditional markets. They allow for an interpretation of the current and future investment opportunities, as well as the entire health of an economy. They include stock market indexes, gross domestic product, crude oil price, gold price, etc.

Stock market indexes follow the ups and downs of a group of stocks or other economic assets. They are most often grouped depending on the industry or country they refer to. Stock market indexes allow for an insight into the market and economic movements across the globe.

Crude oil is a commodity that is consistently in demand, as it supplements economic growth. However, there is a limit to oil production, making this commodity's price a fluctuating one, depending both on economic and political influences. Nevertheless, its importance is omnipresent.

Gold plays an important role as a store of value and a medium of exchange. Unlike other commodities, gold does not get used up, imbuing the precious metal with a sense of everlasting value. Gold serves as a hedge against the declining value of traditional currencies through inflation, making it an invaluable asset for storing massive wealth resources, for banks, countries, and companies.

Natural gas is a clean, affordable, abundant commodity that adds billions of dollars to economies annually. It is a more eco-friendly commodity than crude oil, which still supports massive economic growth, making it a win-win commodity in today's society.

The objective of this research is to study the relationships between five of the most influential stock market indexes (S&P500, Dow Jones, FTSE, Hang Seng, BSE) and three important commodities (crude oil, natural gas, and gold).

Traditional and social media platforms have a profound impact on several aspects of human life, including the economy, as they disseminate and circulate an extensive amount of data. Therefore, in this research, we use data relating to the price of each asset, as well as data gathered from the social network Twitter and news databases Google News and GDELT. We utilize the power of advanced Natural Language Processing, Machine Learning, and Explainable Artificial Intelligence [2]–[4] to retrieve information from the data and use the relationships between the stock market indexes and commodities. The relationship networks we obtain are analyzed in detail to provide ground for further research and understanding of global economic dynamics.

This paper is organized as follows. Section II details the data used in this research. Section III explains our methodology. Section IV gives a comprehensive look into the results. Section V gives our conclusion.

II. Data

In this paper, we use both structured (numerical) and unstructured (text) publicly available data for each of the selected assets: S&P500, Dow Jones, BSE, FTSE, Hang Seng, crude oil, natural gas, and gold. For each asset and source, we look at the time period between March 2019 and March 2023.

We use Investing [1] as a resource to gather all the structured data for each asset. For each asset, we obtain a dataset

[1] https://www.investing.com/

containing the date, price, opening price, high price, low volume, and percentage of change. We use the date and price in the following steps of our research.

Furthermore, we collect data from Twitter[2], Google News[3] and GDELT[4] using their publicly available APIs.

To obtain data from these sources, we send requests to their APIs that allow us to gather data with desired specifics. Therefore, we look only at articles and tweets written in English, published or posted in the selected time period, containing one of the assets' names in their title. Accordingly, we search each API 8 times, once per each asset, as we need their name to appear in the title of the article or tweet. In the end, we obtain 24 datasets, 3 for each asset, containing the date of publishing/posting, the title of the article/text of the tweet, and the link. Some sources provide additional data like the id or name of the user that posted the content, however, we choose not to use this data in our research.

III. Methodology

The first step in our methodology is to obtain the sentiment from all the text data available for each asset and source. For this purpose, we use the DeBERTa model[5], based on a transformer architecture, that has the goal to improve the BERT and RoBERTa models. We extract the sentiment only from relevant text data, after filtering out the articles and tweets that are not of interest.

We use the extracted sentiments to find the average daily sentiment for each asset in each dataset. Therefore, we obtain three time series per asset, one for the Twitter data, one for the Google News data, and one for the GDELT data.

The following steps build on top of our previous working methodology [1], [5]. The first goal is to find all the possible correlations between the assets of interest. To do this, we use the Pearson Correlation Index. We find correlations between prices, and sentiments from Twitter, Google News, and GDELT. Table I shows one of the resulting information obtained from this procedure - the correlation between each pair of assets based on their GDELT average daily sentiment.

We then use these results to obtain networks of the relationships between assets based on each of the data sources used in this research. To create these networks, we represent each stock market index and each commodity as a node. We portray each correlation value as a weight of the link between the two appropriate assets/ nodes. In this manner, we obtain 4 networks, detailed in IV.

The next step in the methodology utilizes the latest advancements in the fields of Machine Learning and Explainable AI. We create a ML model that forecasts the price of an asset and then we apply an Explainable AI model to understand how the prediction was obtained. We feed the ML model data for all assets, thus allowing the Explainable AI model to explain the influence assets have one over another, when forecasting their price.

TABLE I
Correlation based on GDELT average daily sentiments

Keywords	Correlation
Dow Jones, FTSE	0.750
Dow Jones, Hang Seng	0.683
FTSE, Hang Seng	0.664
BSE, Dow Jones	0.469
Crude Oil, Natural Gas	0.442
BSE, FTSE	0.422
Dow Jones, S&P500	0.367
BSE, S&P500	0.364
BSE, Hang Seng	0.362
Crude Oil, FTSE	0.305
FTSE, S&P500	0.272
Crude Oil, Dow Jones	0.254
Hang Seng, S&P500	0.249
Crude Oil, Hang Seng	0.221

To achieve our goal, we utilize XGBoost[3] as our ML model and SHAP[4] as our Explainable AI model. The combination of these two models is shown to be effective in [6] for explaining the influence of atoms, in [7] for real-time accident analysis and in [8].

XGBoost is a ML model that can be used to solve regression, classification, and ranking problems. It utilizes a highly accurate Gradient Boosting Decision Trees (GBDT) algorithm that combines multiple machine learning algorithms to enhance the model's accuracy.

SHAP is an AI model that is specifically designed to explain the results of other machine learning models. This is achieved through use of Game Theory concepts, allowing the model to calculate the contribution of each input to the final result.

In order to analyze the relationships between assets, we first construct the data that we feed to the XGBoost model. We create a dataset that aligns temporally all the features for the assets of interest across the selected time period. The resulting dataset includes the following data for each asset on a given date: price, average daily Twitter sentiment, average daily GDELT sentiment, and average daily Google News sentiment. A sample subset of the resulting dataset can be seen in Figure 1.

Date	Dow Jones Price	GDELT Dow Jones	Google News Dow Jones	Twitter Dow Jones	BSE Price
2019-03-01	26026.3200	0.605613	0.289022	0.231264	36063.810000
2019-03-02	26026.3200	0.580290	0.250733	0.231264	36063.810000
2019-03-03	26026.3200	0.350027	0.476049	0.312400	36063.810000
2019-03-04	26026.3200	0.485935	0.266913	0.244231	36063.810000
2019-03-05	25922.9850	0.416681	0.337445	0.244231	36063.810000
2019-03-06	25884.2000	0.362130	0.383295	0.244231	36253.175000
2019-03-07	25766.5800	0.298587	0.321388	0.244231	36539.320000
2019-03-08	25693.2425	0.227552	0.389360	0.657948	36601.353333
2019-03-09	25644.6420	0.301930	0.443945	0.657948	36618.872500
2019-03-10	25600.8900	0.408526	0.407779	0.657948	36618.872500

Fig. 1. Subsection of the dataset used as input to the XGBoost model

We feed the XGBoost model with this dataset, utilizing all prices and sentiments from the previous day as input features

[2] https://api.twitter.com/1.1/search/tweets.json
[3] https://news.google.com/rss
[4] https://api.gdeltproject.org/api/v2/doc/doc
[5] https://huggingface.co/mrm8488/deberta-v3-small-finetuned-sst2

to predict the current asset price. We repeat this process 8 times to obtain the results for each analyzed stock market index and commodity.

We use the SHAP Explainable ML model to interpret the results obtained from the XGBoost model. We follow two different approaches for calculating the Explainable ML values: full conditional values (observational conditional expectation) and interventional values (interventional conditional expectation) [9].

Interventional values are obtained by breaking down the dependence between features, which allows us to understand how the model would behave if certain inputs were changed. In contrast, full conditional values take into account the correlations between input characteristics. If the model depends on a particular input, then both that input and any other inputs correlated with it are given some credit. The interventional option is ”true to the model,” meaning it recognizes the features used by the model. On the other hand, the conditional value option is ”true to the data” as it considers how the model would behave while respecting the correlations present in the input data.

To incorporate both approaches in our analysis, we begin by using the full conditional values approach with a linear regression model to explain the impact of input variables. We then use the intervention values approach with the created ML model to determine how the input variables affect the predicted variable. Next, we compare the global Explainable ML value with the direction of both approaches for each input variable. Any input variables that have different directions are discarded. Finally, we calculate the average of the first and second approach to obtain the intensity of the impact.

We use these results to obtain a network of the relationship between the assets. This network is shown and analyzed in IV.

IV. Results

In our research, we create weighted networks representing both the intra and inter-dependencies between select stock market indexes and commodities. We accomplish this by using correlation or Explainable AI values to represent the relationships' strength, as detailed in section III.

All the resulting networks have the following nodes: S&P500, Dow Jones, FTSE, Hang Seng, BSE, crude oil, natural gas, and gold. For each network visualization that we show, the link thickness is determined by the link weight, whereas the link color is randomly generated.

First, we analyze the networks obtained using the correlation between assets, based on one of their features, as explained in III. These 4 networks have undirected links, as per the nature of the relationship they represent. For each network, we use only the top 1/2 of all correlations in its creation. Therefore, there are 14 links obtained per network. If a node is not connected, it is representative of the fact that this node has insignificantly small correlation values to the others.

In the network created using the daily price correlations, the link with the highest weight is between the nodes Dow Jones and S&P500, with a weight of 0.976, and the link with the lowest weight is between the nodes BSE and FTSE, with a link weight 0.559. Figure 2 shows this network.

The node degrees are as follows:

- BSE: 5
- Crude Oil: 5
- Natural Gas: 5
- S&P500: 5
- Dow Jones: 4
- FTSE: 2
- Gold: 1
- Hang Seng: 1

We can observe significantly high correlations, which indicate a very strong relationship between the prices of all 8 assets of interest.

We notice that the strongest links are all between nodes that represent stock market indexes, supporting the assumption of powerful relationships between the traditional markets. Another interesting observation is the strong relationship between crude oil and natural gas.

Next, we look at the network created using the average daily sentiment obtained from GDELT data. The link with the highest weight is between the nodes Dow Jones and FTSE, with a weight of 0.750, and the link with the lowest weight is between the nodes Crude Oil and Hang Seng, with a weight of 0.221. Figure 3 shows this network.

The node degrees are as follows:

- Dow Jones: 5
- FTSE: 5
- Hang Seng: 5
- BSE: 4
- Crude Oil: 4
- S&P500: 4
- Natural Gas: 1
- Gold: 1

We observe that the indexes Dow Jones, FTSE, and Hang Seng are central to this network and most connected.

The next network is created using the average daily sentiment from Google News. The link with the highest weight is between the nodes Dow Jones and S&P500, with a weight of 0.338, and the link with the lowest weight is between the nodes Dow Jones and Natural Gas, with a weight of 0.063. Figure 4 shows this network.

The node degrees are as follows:

- Hang Seng: 6
- Dow Jones: 5
- BSE: 4
- FTSE: 4
- S&P500: 4
- Natural Gas: 3
- Crude Oil: 2
- Gold: 0

We can observe that in this network, the stock market indexes are more related and central to the network, leaving the commodities on the ends.

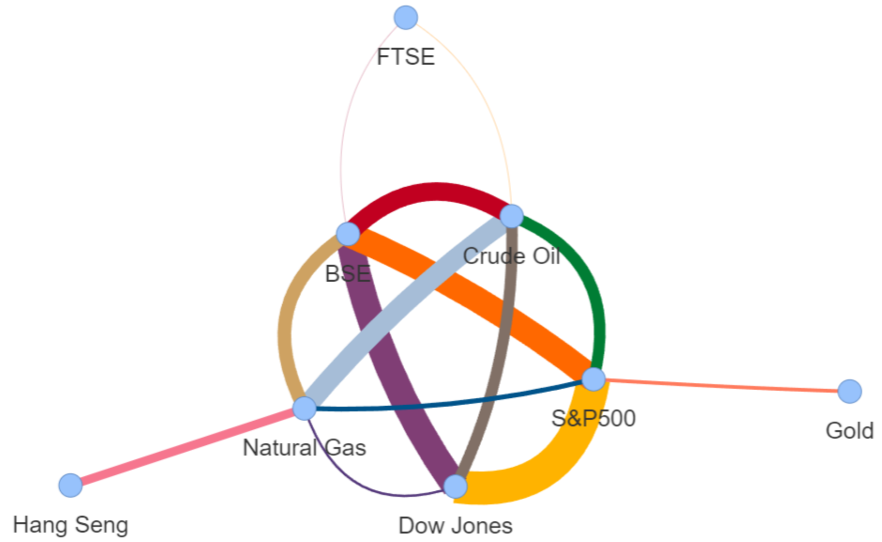

Fig. 2. Network created using the daily prices Pearson correlations between March 2019 and March 2023 for S&P500, Dow Jones, FTSE, Hang Seng, BSE, Crude Oil, Natural Gas, Gold

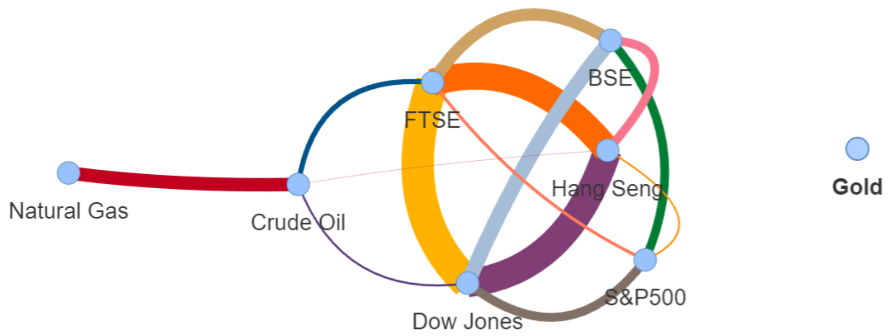

Fig. 3. Network created using the average daily sentiment from GDELT between March 2019 and March 2023 for S&P500, Dow Jones, FTSE, Hang Seng, BSE, Crude Oil, Natural Gas, Gold

The final network obtained in this manner is created using the average daily sentiment from Twitter. The link with the highest weight is between the nodes Dow Jones and S&P500, with weight 0.261, and the link with the lowest weight is between the nodes Natural Gas and S&P500, with weight -0.061. Figure 5 shows this network.

The node degrees are as follows:

- Hang Seng: 6
- S&P500: 5
- Dow Jones: 4
- FTSE: 4
- Natural Gas: 2
- BSE: 2
- Crude Oil: 2
- Gold: 1

We can observe that a link between crude oil and natural gas does not exist in this network, meaning that their correlation with respect to this data source is very low.

Next, we analyze the network obtained using the Explainable AI approach, as explained in III. This network has directed links, as it represents the influence of one asset over an other. This network uses the top 1/4 of all resulting SHAP values. Therefore, there are 14 links obtained in this network. Alike the previous networks, if a node is not connected, it is representative of the fact that the node has insignificantly small influence over the others.

In this network, the link with the highest weight is from the node S&P500 to the node BSE, with weight 1, and the link with the lowest weight is from the node Dow Jones to the node S&P500, with weight 0.060. Figure 6 shows this network.

The node in-degrees are as follows:

- BSE: 6
- Hang Seng: 4
- Dow Jones: 2
- FTSE: 1
- S&P500: 1
- Crude Oil: 0
- Gold: 0

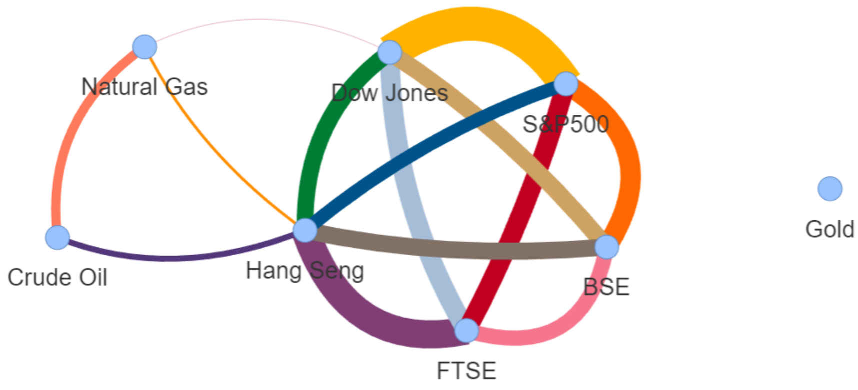

Fig. 4. Network created using the average daily sentiment from Google News between March 2019 and March 2023 for S&P500, Dow Jones, FTSE, Hang Seng, BSE, Crude Oil, Natural Gas, Gold

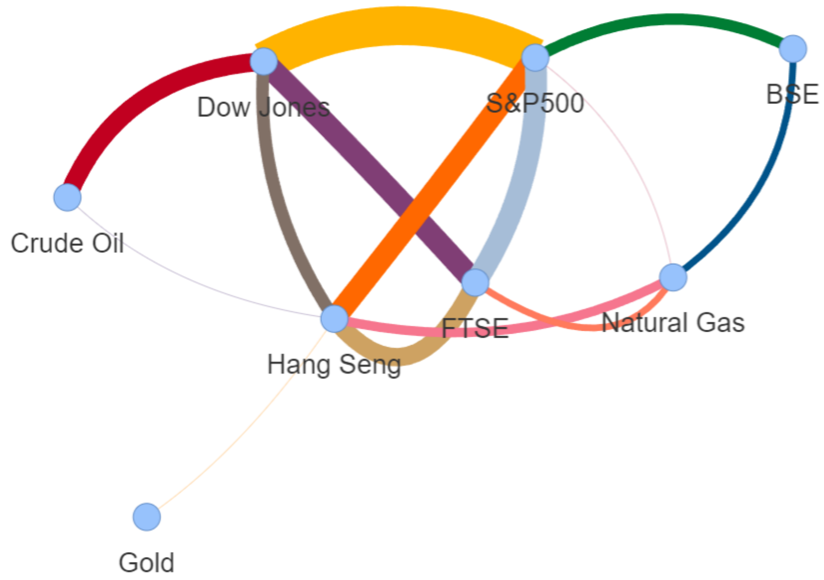

Fig. 5. Network created using the average daily sentiment from Twitter between March 2019 and March 2023 for S&P500, Dow Jones, FTSE, Hang Seng, BSE, Crude Oil, Natural Gas, Gold

- Natural Gas: 0

The node out-degrees are as follows:

- Crude Oil: 3
- S&P500: 3
- BSE: 2
- Dow Jones: 2
- Natural Gas: 2
- FTSE: 1
- Hang Seng: 1
- Gold: 0

We can observe that the node BSE has way more links going into it than all the other nodes which suggests that a lot of other nodes are important for determining its price, while most other nodes have just one or no links going into them.

From all of the different networks, we make some additional conclusions. The strongest links within all of the networks are between stock market indexes. This is supported by the fact that all networks except the Explainable ML one, feature the node Dow Jones on one side of the link with the highest weight, and either S&P500 or FTSE as the node on the other side.

The Explainable ML network also features stock market indexes on both sides of the link with the highest weight, namely S&P500 and BSE. In all networks, at least the top 3 links with highest weights are all between two stock market indexes, and there is usually a strong link between crude oil and natural gas. Gold is the node with the fewest links connected to it, often having none or just one link, which suggests that the relationship between gold and all other nodes is not as significant.

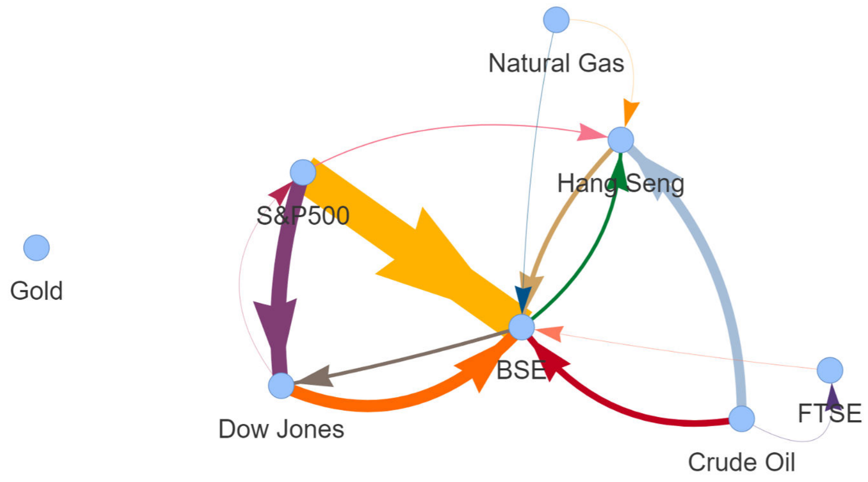

Fig. 6. Network created using the SHAP importance values between March 2019 and March 2023 for S&P500, Dow Jones, FTSE, Hang Seng, BSE, Crude Oil, Natural Gas, Gold

V. Discussion and Conclusion

This paper presents an analysis of the relationships between traditional markets and commodities trade by utilizing publicly available multimodal datasets that include both structured numerical data and unstructured text data.

By incorporating news sources, the methodology used in this paper creates models that consider both quantitative and qualitative data about stock market indexes and commodities.

Using multimodal data fusion our research develops a model that first learns the dependencies between assets and then explains them in a form understandable by humans.

Our paper examines the dependencies between five stock market indexes (S&P500, Dow Jones, BSE, Hang Seng, FTSE) and three commodities (crude oil, natural gas, gold) based on their prices and sentiment from Twitter, Google News, and GDELT. The networks formed using the correlations and interdependencies reveal strong relationships within one type of asset, but weaker relationships between stock market indexes and commodities.

Our research serves as a foundation for understanding the complexities of trade and can be extended to include more data and multiple different models and tools to confirm or refute the results and conclusions.

The findings shown in this paper could be of significant interest for both research and investment purposes, as well as in times of economic crisis.

VI. Acknowledgment

We thank Andrej Sterjev, Aleksandar Radulovski, Stefani Kulebanova and Nikolina Petrovic for their help with data processing and data analytics.

References

[1] A. Todorovska, E. Spirovska, G. Angelovski, *et al.*, “Analysis of cryptocurrency interdependencies,” in *Proceedings of Blockchain in Kyoto 2021 (BCK21)*, 2021, p. 011 004.

[2] P. He, X. Liu, J. Gao, and W. Chen, “Deberta: Decoding-enhanced bert with disentangled attention,” *arXiv preprint arXiv:2006.03654*, 2020.

[3] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[4] S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *arXiv preprint arXiv:1705.07874*, 2017.

[5] G. Angelovski, A. Todorovska, I. Rusevski, *et al.*, “The influence of stock market indexes (s&p500 and dow jones) on cryptocurrencies prices,” 2022.

[6] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. K. Mohammadian, “Toward safer highways, application of xgboost and shap for real-time accident detection and feature analysis,” *Accident Analysis & Prevention*, vol. 136, p. 105 405, 2020.

[7] Y. Meng, N. Yang, Z. Qian, and G. Zhang, “What makes an online review more helpful: An interpretation framework using xgboost and shap values,” *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 16, no. 3, pp. 466–490, 2021.

[8] Y. Bi, D. Xiang, Z. Ge, F. Li, C. Jia, and J. Song, “An interpretable prediction model for identifying n7-methylguanosine sites based on xgboost and shap,” *Molecular Therapy-Nucleic Acids*, vol. 22, pp. 362–372, 2020.

[9] H. Chen, J. D. Janizek, S. Lundberg, and S.-I. Lee, “True to the model or true to the data?” *arXiv preprint arXiv:2006.16234*, 2020.