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

Emerging from the Pandemic and Reshaping Human Endeavors with Digital Technologies

Dec 11th, 12:00 AM

Can Conversations on Reddit Forecast Future Economic Uncertainty? An Interpretable Machine Learning Approach

Jinhang Jiang
Walmart Inc., jinhang.jiang@walmart.com

Mei Feng
University of Kansas, fengmei0520@gmail.com

Karthik Srinivasan
University of Kansas, karthiks@ku.edu

Can Conversations on Reddit Forecast Future Economic Uncertainty? An Interpretable Machine Learning Approach

Completed Research Paper

Jinhang Jiang
Walmart Inc.,
702 SW 8th St. Bentonville,
AR 72716, US
Jinhang.Jiang@walmart.com

Mei Feng
Biopharmaceutical Innovation &
Optimization Center,
University of Kansas
2097 Constant Ave., Lawrence,
KS 66047, US
meifeng@ku.edu

Karthik Srinivasan
School of Business,
University of Kansas
1654 Naismith Dr, Lawrence,
KS 66045, US
karthiks@ku.edu

Abstract

In recent years, social media has become an indispensable source of information through which public attitudes, opinions, and concerns can be studied and quantified. This paper proposes an interpretable machine learning framework for predicting the Equity Market-related Economic Uncertainty Index using features generated from a popular discussion forum on Reddit. Our framework consists of a series of custom pre-processing and analytics methods, including BERTopic for latent topic identification and regularized linear models. Using our framework, we evaluate explanatory models with different configurations over a large corpus of Reddit posts belonging to the personal finance category. Our analysis generates valuable insights about discussion topics on Reddit and their efficacy in accurately predicting future economic uncertainty. The study demonstrates the potential of using social media data and interpretable machine learning to inform economic forecasting research.

Keywords: Economic Uncertainty Analysis; Topic Modeling; Social Media Analytics; Interpretable Machine Learning; Multimodal Modeling.

Introduction

Social media platforms have become an invaluable source of information for studying various social, economic, and political phenomena in modern society (Kurumathur et al., 2022; Sachin et al., 2022; Wang et al., 2022). By analyzing the content of social media posts, we can gain valuable insights into the attitudes, behaviors, and concerns of a wide range of internet users. For example, social media data has been previously leveraged to examine the sentiments of investors toward public stocks and market indices (Deng et al., 2018; Maqsood et al., 2020; Singh Chauhan et al., 2022). Social media data has also been studied to characterize individual and organizational outcomes such as consumer behavior and firm equity value (Ashish et al., 2018; Velichety & Shrivastava, 2022). In a similar light, social media data can be considered a vital source of information for economic policymakers and financial analysts. This paper specifically explores the potential of using social media data to forecast economic uncertainty. We focus our interests on the *Personal Finance* subreddit and develop a framework for analyzing and predicting economic uncertainty. Through a series of experiments, we demonstrate the effectiveness of our approach and identify key topics that influence prediction performance.

To assess economic uncertainty, we model the Economic Policy Uncertainty Index (EUI) introduced by Scott Baker, Nicholas Bloom, and Steven Davis (Baker et al., 2016) as a regression problem. The motivation of this study is that while Baker et al. computed EUI by parsing global news articles, few studies have examined the association between public sentiment and the economic uncertainty index. It is prudent to determine if and how online public sentiment is related to future economic uncertainty. Compared to news articles and microblogging platforms, discussion platforms such as Reddit provide a richer source of information that captures the public's concerns and offers insights into future economic uncertainty. Furthermore, learning the relationship between the social media discourse and EUI variations can contribute to the growing body of research on the role of social media in predicting economic trends and therefore be helpful for macroeconomists, policymakers, firms, and investors.

We collected and analyzed a set of Reddit conversations that specifically talks about personal financial-related issues from the end of 2021 till the mid of 2022. The analysis shows that the volume of particular topics from the conversations can be predictive of future economic uncertainty. Our study contributes to the existing literature in the following ways: (1) We introduce a unique method framework for extracting topics from online discussion platforms using interpretable machine learning (IML) that can be used for modeling economic uncertainty index. (2) We identify the topics whose daily volume on social media platforms could potentially be predictive for future economic uncertainty assessment and interpretation. (3) As part of the predictive evaluation of our proposed framework, we explore a novel multimodal modeling approach that combines the generated topics identified by our framework with a state-of-the-art time series model. Our study contributes to information systems (IS) research in the area of uncertainty modeling using digital technologies. It uses an IML approach to understanding the novel phenomenon of social perception of economic uncertainty, thus contributing to Type II ML research in IS (Padmanabhan et al., 2022). Our analysis informs individuals, businesses, and governments about social media predictors of economic uncertainty. Finally, it updates uncertainty forecasting analysts by introducing text features that can be readily generated from a public data source.

The rest of the sections are presented as follows. Section 2 describes the research background. Section 3 introduces the data used in this study. In Section 4, we explain the analytics methods in detail. Section 5 discusses the analysis. In Section 6, the discussion and conclusions are presented.

Research Background

The EUI index is based on the idea that newspaper text searches can provide useful proxies for economic and policy conditions, particularly in countries with limited data sources or in earlier periods. The authors showed that the EUI can effectively capture economic uncertainty stretching back several decades (Baker et al., 2016). In particular, Equity Market-related Economic Uncertainty Index is a type of EUI that specifically measures the level of uncertainty present in the market. High levels of economic uncertainty are often associated with increased volatility in the stock market, as investors may be more hesitant to make decisions in an uncertain environment. By tracking the index, economists and financial analysts can get a sense of the level of uncertainty that is present in the market and use this information to make more informed decisions about investing and trading.

Social media platforms are rich with information about individual, organizational, and social behavior across time and geographies. Studies have looked at causal as well as associative links between social media and stock market uncertainty (Chaudhary et al., 2020; John & Li, 2021; Ortiz, 2022). Studies have also extracted timely economic signals beyond social media such as newspaper articles to improve forecasts of macroeconomic variables, such as inflation and unemployment (Kalamara et al., 2022; Ryu, 2018). An economic policy uncertainty index was also proposed by using a select set of Twitter accounts whose tweets are considered to reflect an expert opinion on economic policy issues (Yeşiltaş et al., 2022). Reddit is a discussion forum with users focusing on detailed conversations based on recent events and news articles (Proferes et al., 2021). Prior studies either analyze data from microblogging platforms such as Twitter or examine the role of news articles in changes in macroeconomic variables. Our analysis of Reddit and the

Economic Policy Uncertainty is the first to determine and model a potential relationship between the social media discourse and economic uncertainty.

Facebook, LinkedIn, and similar platforms primarily emphasize personal and business networking and not public broadcasts of personal socio-political opinions, thus making them unsuitable for our problem. Twitter is a widely used source of data in which users commonly voice opinions about social and political issues. But the micro-blogs (i.e., tweets) generated by users are often brief and conversational, making it challenging to gather a large volume of data on a specific topic. In contrast, the posts on Reddit tend to be longer and more descriptive, and therefore more informative. The Reddit user community consists of different interest groups invested in numerous forums commonly known as subreddits. Literature in personal finance indicates a relationship between economic conditions and individual financial success (Garman & Forgue, 2014).We identified the "Personal Finance" subreddit as particularly relevant for our research, given its focus on topics such as budgeting, saving, debt, credit, investing, and retirement planning. We believe that these user topics may be directly or indirectly linked to wider economic issues, as individual investment behavior and sentiments are impacted by the economy and vice-versa. This subreddit has a membership of 17.1 million users and is currently ranked twenty-second in size across all of Reddit, making it a large and active community serving as a valuable source of data for our study. In comparison to text analysis on expert opinions in news articles and public announcements, text analysis on discussion forums such as Reddit provides the opportunity to analyze dynamic social perception of economic health and uncertainty. This is mainly because information on microblogs and news articles typically surfaces as a result of systematic journalistic procedures. Journalists often conduct interviews with diverse stakeholders, aggregate expert opinions, and meticulously curate content. Consequently, the content presented in such mediums often reflects events or issues that have already permeated the broader societal consciousness. Traditional news articles and microblogging data sources provide comprehensive on established topics of public interest. In contrast, online discussion platforms such as Reddit offer the opportunity to uncover the collective consciousness of individuals at the grassroots level. Discussions on platforms like Reddit precede the mainstream recognition of emerging issues, enabling the exploration of latent concerns and sentiments that may eventually coalesce into prominent societal phenomena. For instance, consider the scenario where a major economic event such as a wave of mortgage foreclosures crystallizes in the aftermath of a pandemic-induced economic downturn. Before such an event gains universal attention, individuals may have already grappled with personal financial challenges about which they may be discussing online for brainstorming solutions. Excluding atypical factors such as global disaster such as pandemics or wars, shifts in economic health and public market indices mirror public perception and discourse on the internet. To summarize, Reddit discussions could provide real-time insights into how individuals perceive and react to economic developments, government policies, and market trends, which may be invaluable signals for tracking and forecasting of economic uncertainty at a daily level.

Interpretable machine learning (IML) refers to the practice of designing and implementing machine learning models and algorithms in a way that makes their predictions and decision-making processes interpretable by users. It is a subset of explainable artificial intelligence (XAI) that aims to make AI systems more transparent and accountable. While fully parametric statistical models such as linear regression are widely used for explanatory modeling, their predictive power is lower due to high bias (Shmueli & Koppius, 2011). In comparison, XAI and IML methods are gaining importance for explanatory modeling as they use complex multi-layered algorithmic models such as ensemble learners and deep neural networks to first predict a phenomenon followed by using ingenious approaches to interpret the underlying phenomenon by exploiting the structure of the trained model (Molnar, 2020). While IML methods focusing on feature contribution (Lundberg et al., 2017), local prediction explanation (B. Kim et al., 2019), and embedded interpretability (Rudin, 2019) are more common, novel methodology frameworks also plan an important role in facilitating user understanding and deployability of complex ML applications (Doshi-Velez & Kim, 2017). Concurrently, multimodal learning is a modeling approach focusing on combining different modalities of data, i.e., different types of input features, to improve predictive modeling (G. Kress, 2009). Multimodal learning for ranking text attributes used as features in a complex predictive model has the potential to explain complex associations between repeated text patterns and outcomes. For EUI modeling using social media data, IML with multimodal learning is suitable for multiple reasons. First, predictive modeling of EUI is an upcoming research area and therefore interpretability can help promote transparency and trust in complex ML systems developed for this application. Second, online discussion forum data is highly unstructured, with diverse language use, slang, and implicit sentiments. Therefore, capturing text content features in the ML model helps understanding the data generation mechanism as well as underlying economic uncertainty social perception phenomenon. Lastly, the EUI model based on our framework uses information from a single source Reddit, and future work can extend it with additional sources and method improvements. Therefore, an IML approach aids the model adoption and future enhancement processes.

Data

Raw Data Generation

From November 23rd, 2021, to June 25th, 2022, we leveraged the Python Reddit API Wrapper, 'PRAW' (Boe et al., 2014), to scrape the posts and replies within the "Personal Finance" subreddit at one-hour intervals. This resulted in a data set spanning a 215-day period. The data set consists of features including Submission_Id, Reply_Id, Submission_title, Author, Date, Vote, and Text. The Submission_Id and Reply_Id will be the same for the main post, while the remaining features pertain to the corresponding submission or reply. Note that we experienced inconsistencies in the collection process due to internet and API outages on December 30th, 2021, and February 3rd, 2022, resulting in gaps in data extraction. A summary of the raw data is shown in Table 1.

Total Records	**Total Tokens**	**Daily Avg. # Users**	**Daily Avg. # Records**	**Daily Avg. # Unique Posts**
1,228,571	71,316,874	2,717	5,714	455

Table 1. Data Summary

Data Preprocessing

We performed multiple steps of data cleaning to prepare the dataset for modeling. These steps included removing documents with fewer than 10 tokens, which tend to be less informative, eliminating all stop words and web links, dividing long documents into sentence-level units, and ensuring that no records contain more than 500 tokens. The resultant corpus contained 1,028,333 sentences, with an average word count of 42.57 per sentence.

Feature Engineering

We first identified topics across reddits (i.e., Reddit posts), using BERTopic (Grootendorst, 2022), a deep learning method for topic modeling, described in detail in the next section. To determine the relationship between the identified topics in "Personal Finance" and the future value of EUI, we computed the EUI for a period of n days after the reddits were posted. We consider three values of *n* in our study, one, three, and seven, to compare the different potential lagged effects of topics on economic uncertainty predictions. To account for endogeneity in the model, we include the following covariates in our model - day of the week, month of the year, and level of user activity at the time each topic was posted. The level of user activity is measured by the daily number of conversations captured from the subreddit forum and the number of active users on that given day. The data was collected over a period of 215 days, resulting in 215 observations used for training and validation.

Methods

In this section, we provide details of our method framework. First, we use BERTopic (Grootendorst, 2022) to extract relevant topics and their frequency for each corresponding date across the Reddit posts. Then we propose a new topic modeling framework for making reliable EUI predictions. Finally, we demonstrate the applicability of our framework by presenting a model combining the top features extracted using previous steps.

Topic Extraction

BERTopic is a topic modeling method based on word embedding (Mikolov et al., 2013) and transformer technologies (Devlin et al., 2019). It builds upon legacy topic modeling methods such as Latent Dirichlet Allocation (LDA) that are designed to extract coherent topic representations by identifying explicit relationships between words or phrases in the text (Blei et al., 2003). In contrast, BERTopic generates document embeddings using pre-trained transformer-based language models, after reducing the dimensions and clustering of the embeddings, to generate topic representations using a class-based variation of the term frequency-inverse document frequency (c-TF-IDF) procedure (Grootendorst, 2022). c-TF-IDF can best be explained as a TF-IDF formula adopted for multiple classes by joining all documents per class and it can be seen as the importance scores for words within a cluster. It has been shown to generate coherent topics and to be competitive with other classical and more recent topic modeling approaches across various benchmarks (Devlin et al., 2019).

BERTopic was chosen as the topic modeling method for our study for its unique ability to capture the semantic richness of text data. Unlike traditional methods such as LDA or Latent Semantic Analysis (LSA), BERTopic utilizes pre-trained embeddings. This allows it to understand the context and meaning of words in a more sophisticated manner, which is particularly beneficial when dealing with user-generated content on discussion platforms like Reddit. Furthermore, BERTopic leverages contextual embeddings to identify topics, enabling it to uncover subtle nuances and associations within the text. This contextual understanding is especially relevant when analyzing discussions related to economic uncertainty, where topics can be multidimensional and context dependent.

Identifying the optimal number of topics using metrics such as perplexity or coherence (Newman et al., 2010) can be computationally intensive. Therefore, in our framework, we propose to compare the performance of model configurations with different topic counts.

Feature Selection

The BERTopic method can potentially identify thousands of topics from text like Reddit discussions. In our model, each topic is considered as an input feature. Therefore, feature selection becomes necessary to develop a meaningful model for economic uncertainty. Least Absolute Shrinkage and Selection Operator, i.e., Lasso, is a popular regularization-based feature selection method (Tibshirani, 1996). Lasso accomplishes variable selection by shrinking insignificant predictor variables' coefficients to zero, particularly in high-dimensional data.

Cross Validation

While the lasso is helpful for eliminating features that do not contribute to characterizing the outcome, the method can still result in a large number of unranked features. Therefore, to gauge the external validity of topics as predictors, a machine learning model such as the random forest regressor is trained using N-fold cross-validation to derive feature importance rank. These features can then be included in the multimodal model described in the next sub-section or can be used as the final predictive model.

Figure 1 summarizes our proposed framework for using topics to measure and predict future economic uncertainties. The illustration streamlines the process of collecting the data, extracting the topic, filtering the irrelevant information, and forecasting future economic uncertainties.

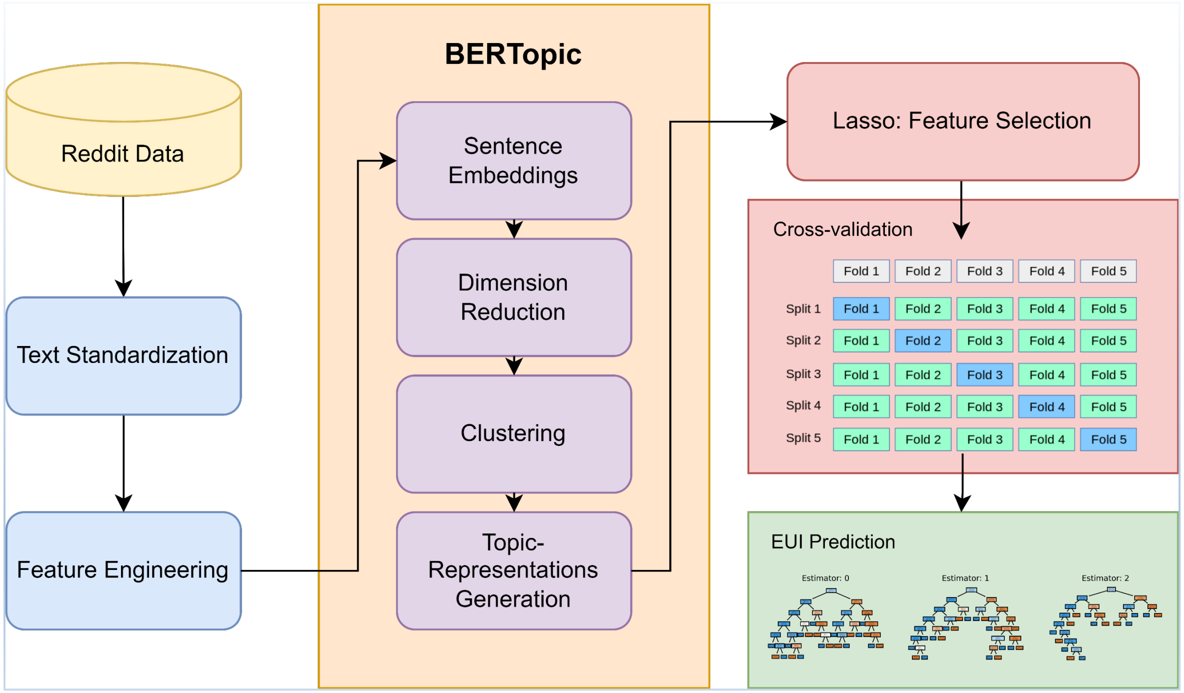
Reddit Data
Text Standardization
Feature Engineering
BERTopic
Sentence Embeddings
Dimension Reduction
Clustering
Topic-Representations Generation
Lasso: Feature Selection
Cross-validation
Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 1 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 2 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 3 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 4 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 5 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
EUI Prediction
Estimator: 0
Estimator: 1
Estimator: 2

Figure 1. Method Framework for Determining Economic Uncertainty Using Reddit Text

Baseline Model

It is important to establish prediction as a necessary scientific endeavor for developing newer explanatory models (Breiman, 2001b; Shmueli, 2010). Assessing predictive performance over test data not only validates the quality of model fit but also helps uncover potential new causal mechanisms and leads to the generation of new hypotheses. Predictive power assessment offers a straightforward way for benchmarking explanatory modeling methods (Shmueli & Koppius, 2011). Therefore, even though our study's emphasis is on explanatory modeling, we test the validity of our proposed framework using predictive performance evaluation. We first train a strong baseline model for EUI prediction. Therefore, we use a multimodal modeling strategy to assess the predictive power of top-10 features generated using our framework.

Prophet (Taylor & Letham, 2018) is an additive forecasting method designed for modeling time series with non-linear trends, seasonality, and holiday information. It is a state-of-the-art time series model that can incorporate a wide range of input types and is easily interpretable(Ensafi et al., 2022). Prophet has been shown to be better than data mining models for time-series data. This is because the latter leads to data leakage as it ignores feature autocorrelations.The Prophet model can be represented as follows:

$$y_t = g(t) + s(t) + h(t) + \epsilon_t \tag{1}$$

In Equation (1), y_t is the value of the time series at time t, $g(t)$ is the trend component modeling non-periodic changes, $s(t)$ is the component for capturing the seasonal changes, $h(t)$ represents the effects of holidays that occur on potentially irregular schedules over one or more days, and ϵ_t represents any idiosyncratic changes.

Multimodal Modeling

To improve the forecasting accuracy of the FB Prophet model, we incorporate the top-K most important features generated from feature importance analysis into the baseline model and conduct another round of simulation. This process is called as a multimodal modeling approach (G. R. Kress, 2010). In this process, we use multiple modalities of data to improve the accuracy and robustness of the baseline model. The equation of the transformed model with the additional ten features is as follows:

$$y_t = g(t) + s(t) + h(t) + \sum_{i \in T} v(i_t) + \epsilon_t \tag{2}$$

In Equation (2), $v(i_t)$ is the volume of a given topic specific topic i on a given day t, and T is the set of all topics. Equation (2) is similar to ARIMAX models, where the time series is modeled as an auto-regressive integrated moving average with exogenous variables X (Friedman et al., 2001). However, our model uses PROPHET which has added benefits over the traditional ARIMA family of time-series models. It is more interpretable as it decomposes the time series into trend, seasonal, irregular scheduled interruptions (e.g., public holidays), and a random component. Further, PROPHET has also been shown to have a better fit compared to ARIMA classes of models, thus enhancing the internal validity of our final feature importance estimates.

Analysis

Following the raw data generation, data processing, and feature engineering, described in the data section, we applied our method framework to develop an explainable model over the Reddit dataset for economic uncertainty forecasting. For the topic modeling step of our framework, we generated sentence embeddings using the pre-trained model *all-mpnet-base-v2* from SentenceTransformers (Reimers & Gurevych, 2019), reduced the embedding dimensions using the UMAP procedure (McInnes et al., 2020), clustered the documents using the HDBSCAN method (Ester et al., 1996; McInnes et al., 2017) and finally derived the topic representations using the c-TF-IDF method (Grootendorst, 2022). Note that the base method of BERTopic (Grootendorst, 2022) offers the flexibility to change the components of our framework to any widely used software (i.e., K-means can be instead of HDBSCAN for clustering, PCA can be used for dimension reduction instead of UMAP, etc.). Our choices of methods were based on interim experiments conducted toward improving our overall model's performance.

Experimental Setup

To analyze the daily frequency of each topic, we obtained the topic group for each corresponding date and aggregated them. However, due to computational constraints, we were unable to determine the optimal number of topics solely based on perplexity, as calculating probabilities using BERTopic is computationally intensive. Therefore, we compared the performance of the model with a set of fixed numbers of topics, including 10, 30, 50, 100, and 200. In addition, we also obtained a set of 2504 topics by allowing the model to automatically generate the optimal number of topics based on the distance strength set in the clustering algorithm. This resulted in a total of six sets of data.

Our experimental setup consists of six sets of data labeled with three future periods of EUI, resulting in a total of 18 combinations of models for comparison. Our dataset contains a large number of features, potentially reaching up to 2500+. Using the Lasso method, we retained only contributing predictors in the input feature set. Thereafter, a Random Forest Regressor (Breiman, 2001a) with default parameters in Python's *scikit-learn* package is trained using 10-fold cross-validation for the feature ranking step of our framework.

For evaluating the baseline model and the multimodal model, a training period of 28 days is utilized to forecast the outcome for the next 7 days. A walk-forward approach can be employed with a step size of 1 day per observation of the training data. For instance, the walk-forward approach results in 181 time series for 215 observations in our study. Each time series is fitted with Prophet using 3,000 Markov chain Monte Carlo (MCMC) draws. By comparing the predicted and actual values, the performance of the FB Prophet model can be benchmarked. All the experiments are conducted in Google Colab with an NVIDIA A100 Tensor Core GPU.

To evaluate the final predictions for each combination, we chose Symmetric Mean Absolute Percentage Error (SMAPE) and Spearman Correlation. SMAPE is a suitable metric for evaluating the accuracy of forecasting methods because it is symmetric, meaning that it treats both positive and negative errors. Moreover, SMAPE, by its nature, favors over-forecasting compared to under-forecasting, which aligns with our objective of presenting a more conservative and safer projection of the EUI index to the audience. The percentage scale of SMAPE makes it easy to understand and compare results, while its reliance on the mean absolute percentage error makes it resistant to the influence of extreme values. We chose Spearman Correlation as another evaluation method because it can detect the monotonic relationship between two variables. The variables in a monotonic relationship frequently change together, albeit perhaps not always at the same rate. Furthermore, an assessment that considers this kind of relationship can be helpful for benchmarking EUI predictions.

Topic Modeling

The results of topic modeling are shown in Table 2 and Table 3. The estimators with the most topics yield the best in predicting the EUI one day after the posts were made, with a SMAPE of 45.65% and a Spearman Correlation of 0.313. We performed feature selection using Lasso and found that the combination with the best performance resulted in a final count of selected features as 194, representing approximately 92.27% of the total number of features being filtered out.

Number of Original Topics	1-day after	3-days after	7-days after
10 topics	49.20 ± 4.80	53.46 ± 2.41	53.98 ± 2.01
30 topics	49.64 ± 5.42	54.04 ± 0.78	52.81 ± 2.21
50 topics	49.52 ± 1.76	54.32 ± 3.65	54.42 ± 5.07
100 topics	50.22 ± 0.49	52.42 ± 0.32	53.40 ± 3.06
200 topics	49.50 ± 5.79	51.50 ± 2.68	52.16 ± 2.52
2504 topics	**45.65 ± 1.19**	**50.53 ± 1.84**	**50.93 ± 3.59**

Table 2. Cross Validation Results - SMAPE

Number of Original Topics	1-day after	3-days after	7-days after
10 topics	0.2233 ± 0.0893	-0.0626 ± 0.0898	0.0815 ± 0.0395
30 topics	0.2064 ± 0.1251	-0.0395 ± 0.0978	0.1299 ± 0.1216
50 topics	0.2601 ± 0.0912	**-0.0966 ± 0.0991**	-0.0264 ± 0.1467
100 topics	0.2542 ± 0.0966	0.0175 ± 0.1130	-0.0052 ± 0.1074
200 topics	0.2615 ± 0.0900	0.0509 ± 0.0743	0.0637 ± 0.0622
2504 topics	**0.3130 ± 0.0459**	0.0164 ± 0.2049	**0.2041 ± 0.1421**

Table 3. Cross Validation Results – Spearman Correlation

We implemented permutation importance to measure the importance of selected features in relation to EUI one day after posts are made with the help of a Python library called Eli5 (Korobov & Lopuhin, 2016). Permutation feature importance is a model evaluation method that can be applied to any fitted estimator with tabular data, particularly useful for non-linear or complex models. It is defined as the decrease in model performance, which is measured by evaluation metrics like F1 or R2, when a single feature value is randomly shuffled (Breiman, 2001a). By disrupting the relationship between the feature and the target, the decrease in model performance reflects the extent to which the model relies on the feature. The output from Eli5 is known in the literature as "Mean Decrease Accuracy (MDA)" or "permutation importance" (Korobov & Lopuhin, 2016). Figure 2 displays the permutation importance for the top 10 topics after 100 iterations with 10-fold cross-validation.

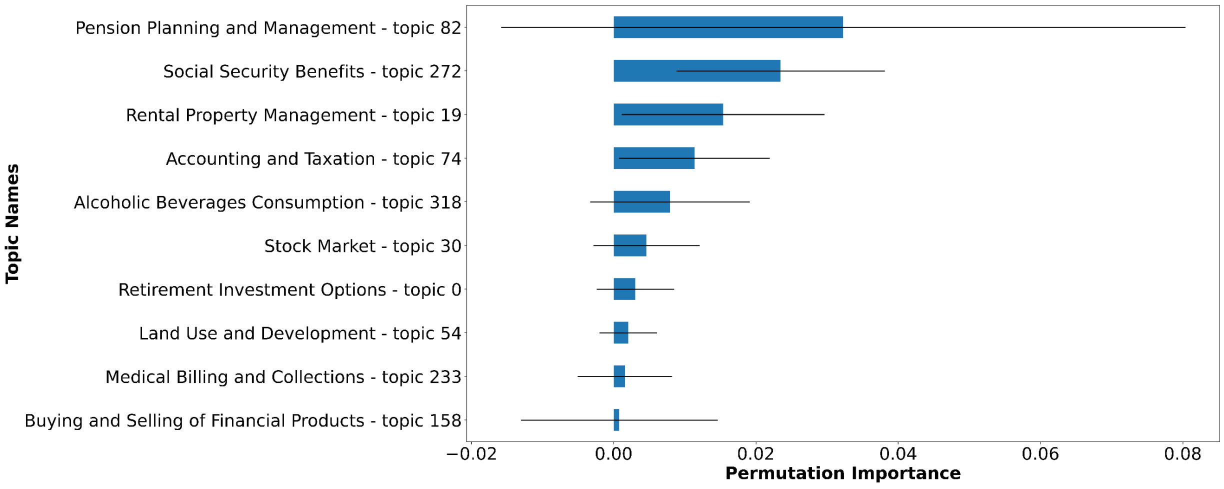

Figure 2. Permutation results for the top ten features

Figure 3 shows the temporal variation in topic frequency for the eight topics that are statistically significant in our model.

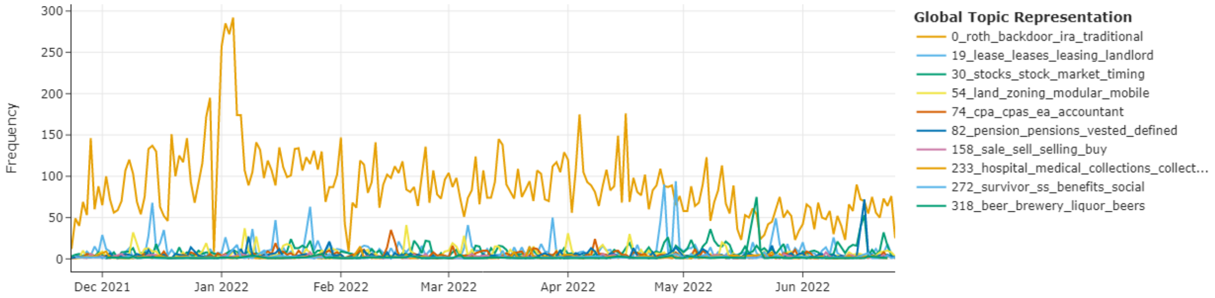

Figure 3. Variation in topic frequency over time for selected features

Prediction Evaluation

The model represented by Equation (2) is estimated using Monte Carlo simulations. As the model structure becomes more complex, MCMC sampling tends to reach saturation and may become trapped in local maxima. To prevent the model from becoming saturated, we set the *adapt_delta* parameter to 0.95 (default is 0.8) and the *maxtreedepth* parameter to 15 (default is 10) for each time series. In Bayesian inference, the No-U-Turn Sampler (NUTS) is a widely used method for sampling from the posterior distribution (Hoffman & Gelman, 2014). To achieve efficient sampling, a key tuning parameter is *adapt_delta*, which controls the step size for NUTS by adjusting the acceptance probability for each candidate point in the Markov chain. Choosing a high *adapt_delta* allows for larger steps but may result in slower convergence (Gelman et al., 2014). In contrast, selecting a lower value leads to a smaller number of steps yet may converge faster. In addition, we also tuned the *maxtreedepth* parameter, which defines the maximum depth of the binary tree used in the sampling process. By setting a higher *maxtreedepth* value, the sampler can explore a more complex model space, but this may lead to longer computational times and higher memory requirements. Conversely, setting a lower value restricts the exploration of complex models but can result in faster sampling.

Figure 4 displays the 7-day average SMAPE from both, the baseline model as well as our multi-modal model that includes the daily volume of the additional 10 topics as regressors. It is clear that after adding the daily volume of these topics, the model's performance became more stable. The average 7-day average SMAPE of the baseline model is roughly 63.46 with a standard deviation of 20.97 and the corresponding average SMAPE of the final model is reduced to 54.77 with a standard deviation of 18.92. We conducted a one-way ANOVA to assess the differences in the performance of those two models. The results indicated a significant difference between the results of the two models (F-statistic = 17.13, p < 0.01).

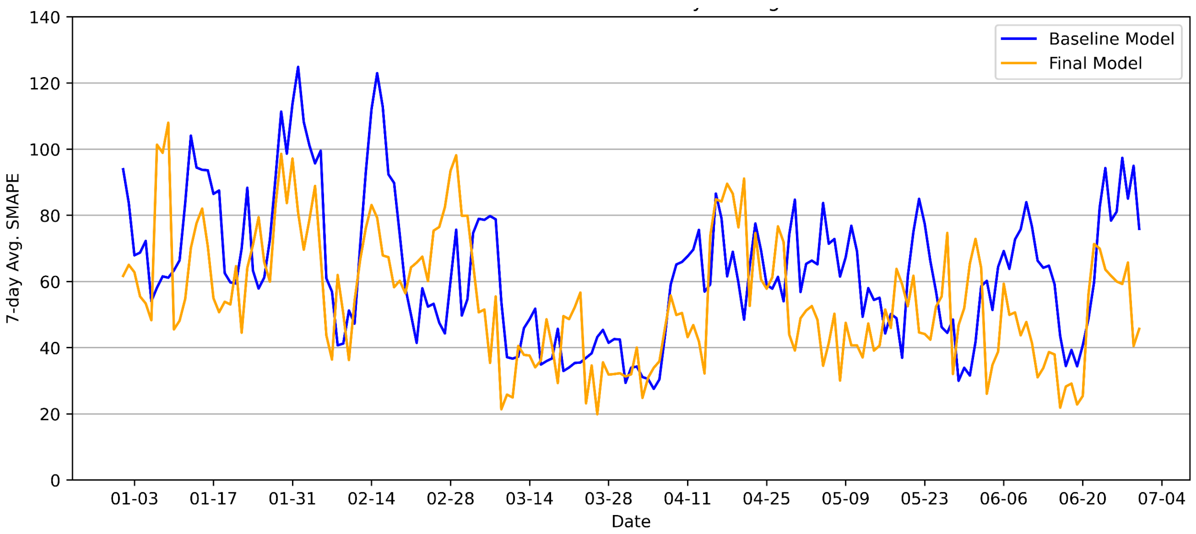

Figure 4. 7-day average SMAPE of baseline model and multimodal model

Figure 5 shows the correlation between EUI and each topic. Based on the analysis presented in Figure 5, we observe a positive correlation between the EUI and discussions related to Rental Property Management, Pension Planning and Management, Social Security Benefits, and Alcoholic Beverages Consumption. Conversely, we observe a negative correlation between the EUI and discussions pertaining to Retirement Investment Options, Stock Market, Land Use and Development, Accounting and Taxation, Buying and Selling of Financial Products, and Medical Billing and Collections.

Discussion and Conclusions

Our proposed framework exhibits the best performance in terms of SMAPE and Spearman Correlation for predicting EUI scores one day after posts or replies are made on the "Personal Finance" subreddit. Among the base training sets, the data set with the most topics automatically optimized by the BERTopic algorithm outperforms the other datasets in five out of six instances. The Lasso algorithm effectively filters out 92.27% of the total features. After conducting permutation analysis, we identified ten topics that have the greatest impact on predictions for the next day's EUI scores. Following the findings from Figure 2 and Figure 3, we dive into the details and analyze the topics making contributions to the predictions. The topics are manually categorized into the following: Retirement Investment Options (Topic 0), Rental Property Management (Topic 19), Stock Market (Topic 30), Land Use and Development (Topic 54), Accounting and Taxation (Topic 74), Pension Planning and Management (Topic 82), Buying and Selling of Financial Products (Topic 158), Medical Billing and Collections (Topic 233), Social Security Benefits (Topic 272), and Alcoholic Beverages Consumption (Topic 318). The listed topics are often related to financial stress and insecurity. For instance, an increase in discussions about renting or leasing could suggest that people are unable to afford home ownership, having issues with on-time payments, or seeking alternatives for housing, which potentially indicates economic strain. Similarly, increased discussions about tax-related issues may indicate that people are facing financial challenges or uncertainty and seeking professional assistance in financial planning and management. Pension management, medical billing and collections, and social security benefits are often related to financial insecurity in the context of healthcare, employment, and retirement. An increase in the volume of such conversations is likely to reflect increased economic uncertainty and financial insecurity among the people discussing them. For example, more frequent discussions about pension or social benefits may suggest concerns about financial security in retirement or difficulties related to unemployment, while an increase in discussions about hospital discharge bills or early withdrawal from 401(k) may indicate financial challenges related to healthcare costs or uncertainty about future sources of income. Conversations over buying and selling of financial products and stock markets are also the core topics as people's behavior is influenced by their perception of the current economic conditions and their expectations of future economic conditions. When individuals are uncertain about the direction of the economy, they may become hesitant to invest in the stock market or purchase financial products. This uncertainty can lead to changes in conversation and behavior, such as increased discussion about the risks and benefits of different investments, or a reluctance to make large financial decisions. The topic regarding the consumption of alcoholic beverages is an interesting one on the list. The topic may be indicative of the level of personal economic stress and well-being because individuals who are experiencing financial stress may turn to alcohol as a means of coping with their financial problems. Therefore, when conversations over these topics pile up in the subreddit, the sentiment of an individual's perception towards the current and future economy is captured.

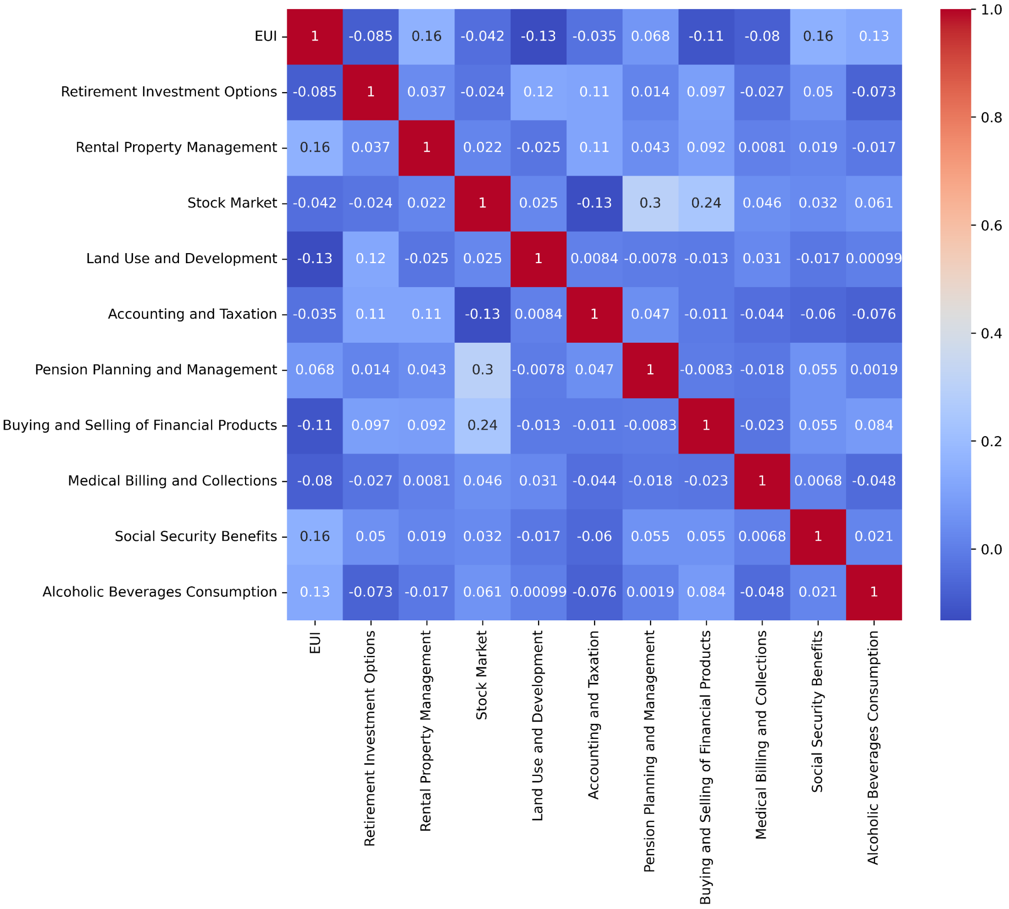

Figure 5. Correlation heatmap for EUI and the relevant topics

Managerial Implications

Our analysis results inform businesses, individuals, and policymakers. The 7-day average SMAPE reduced by 8.69 from the baseline model when we incorporated the volume of the topics. Our analysis suggests that discussions related to money retrieval, management of pensions and social security, or consumption of alcohol within the Personal Finance subreddit may reflect a buildup of concern among individuals towards future economics and possibly be predictive of an increase in economic uncertainty. Conversely, an increase in conversations about spending money or investment planning may indicate a greater faith in the current economic situation and suggest a potential decrease in the EUI. Policymakers could monitor financial forums from time to time to discern early signals of economic slowdowns.

Our analysis started with an exploration of the EUI target values across different scenarios. The original data set showed a high level of volatility, with an average EUI of 103.41 and a standard deviation of 66.70. Further investigation into the 56 trials where the final model failed to improve performance over the baseline revealed an average EUI of 119.49 with a standard deviation of 69.84. However, for trials where the model's performance increased from the baseline, the average EUI was 97.71 with a standard deviation of 64.83. These findings suggest that the integration of the volume of relevant topics into the time series could not significantly improve forecasting accuracy when the target values were highly volatile. This shows that businesses and investors should take caution while forecasting economic activity and stock indices during volatile conditions.

Integration of the volume change in relevant topics did not consistently improve forecasting accuracy. Specifically, the final model failed to outperform the baseline in 56 out of 181 trials. To better understand the underlying factors that impede the effectiveness of topic features, we conducted a detailed analysis of the data. Our investigation highlights the significance of several factors that may influence the relationship between topic features and forecasting accuracy. Our investigation also revealed an intriguing pattern regarding the volume of conversations related to alcoholic beverages consumption. Specifically, we found that the average number of discussions in trials where the final model failed to beat the baseline was nearly four times higher than in trials where the model's performance increased. And we didn’t observe any major differences for other topics under different scenarios. This finding raises the possibility that the ambiguous nature of alcohol consumption, which can be associated with both positive and negative connotations such as stress relief or celebration, could create noise and mislead the forecasting model. Economic modelers should consider such complexity in discussions while incorporating social media features, which go beyond aggregate sentiment analysis and emotion extraction.

Theoretical Contributions

We hypothesized that the aggregated personal financial decisions and concerns may be reflected in the overall economic conditions. We found that topic content in a finance-related sub-forum of a discussion platform were predictive of economic uncertainty measured by a well-established index. Our model not only forecasts the economic uncertainty index for different look-ahead intervals, but our IML approach also uncover knowledge about topics in discussion forums linked to economic uncertainty. This solicits the inclusion of social perception features in economic uncertainty forecasting models. The text features can be easily generated from Reddit using our proposed framework for economic-forecasting modelers, financial analysts, and researchers.

Digital technologies such as ML and AI have taken center stage in many research disciplines including IS. IS Researchers have been active in exploring innovative ideas related to AI and ML methods development as well as their applications in a variety of business domains (Benjamin & Raghu, 2023; B. (Raymond) Kim et al., 2023; Xie et al., 2023). Out of the three types of ML contributions in IS, our study contributes to IS discipline in terms of understanding an important business phenomenon using ML. Our IML framework presents an online discussion-oriented lens for short-term economic uncertainty forecasting. Unlike ML methods contribution studies (i.e., Type 1 ML), our study introduces a novel data source for the economic uncertainty forecasting problem and delineates the predictive modeling process with interpretability in mind.

Our study finds that the changes in the volume of the topics mentioned in this paper are not the direct cause of the changes in EUI. Rather, these signals could be predictive of the future EUI as they reflect individuals' sentiments towards the current and future economy. These sentiments may be influenced by a variety of factors, including news, personal life events, employment changes, and investment losses. Therefore, even though the topical features did not significantly improve the forecasting performance, our study demonstrates that the conversations taking place on this subreddit can be both descriptive and predictive of the future economic direction. We do not make any causal claims in this study while examining the value of social media signals in economic uncertainty forecasting models. Although the volume of topics alone cannot result in changes to the EUI, it can indicate a correlation between peoples’ major concerns and the future uncertainty embedded in economics. Recurring and prominent topics across discussions serve as a reflection of public feelings and perceptions of the economy, which is dynamic, descriptive, and non-trial to quantify.

Limitations and Future Work

Given computational limitations, we were unable to conduct an exhausting exploratory analysis and process all sentences in the data set. Additionally, as the sentence transformers cannot handle long texts, we had to cut the text into fixed-length chunks. This slicing process results in broken semantics of each chunk. Moreover, each chunk will only be labeled with one topic, thus some of the labels possibly only captured the meaning of local syntax yet ignored the full context and authors' original intentions.

As a next step, we plan to utilize the partial_fit feature in BERTopic to fit the model with batch-like data, enabling incremental training of the topic model. Another future direction will be to standardize the process of merging topics and eliminating irrelevant topics, similar to dimension reduction but with the goal of retaining more useful information. To improve the predictive accuracy of the model developed using our proposed framework, features from additional data sources can be explored such as government announcements, corporate disclosures, and urban mobility indices. Further, text descriptive attributes such as emotions, readability, and sentiment can be augmented to topic-based features to improve predictive performance. Different topic modeling methods and machine learning methods can be benchmarked as a future extension of our study.

Conclusions

In this study, we model the Economic Policy Uncertainty Index (EUI) invented by Scott Baker, Nicholas Bloom, and Steven Davis using past information contained in the Personal Finance sub-group of users of Reddit.com, a popular social media platform dedicated to discussion of vart variety of subjects. Our research stands as the pioneering effort to derive content characteristics from online discussion platforms for the purpose of forecasting EUI and uncovering connections between particular themes in personal finance forums and forthcoming economic uncertainty. Our study presents a novel interpretable machine learning framework for filtering content from the online discussion platform that may be predictive of the EUI index. The framework consists of text standardization, feature engineering, topic extraction, dimension reduction, cross-validation, and EUI calculation modules. We conducted 18 modeling experiments with various combinations of the number of topics and future EUI scores. The best estimator was found to be the one with 2504 original topics automatically generated by BERTopic for EUI prediction one day after posts and replies were made. Some of the top topics influencing prediction performance include retirement investment, social security benefits, renting, pension, medical bills, etc. A change in the volume of posts containing these topics can serve as an indicator of economic fluctuations. Our study hints that online special-interest groups often get an intuition about dramatic changes occurring in the macro-environment and therefore express their sentiments that may be predictive of future economic conditions.

References

Ashish, K. R., Bezawada, Kumar, A., Bezawada, R., Rishika, R., Ramkumar, J., & Kannan, P. K. (2018). From Social to Sale: The Effects of Firm Generated Content in Social Media on Customer Behavior. *Journal of Marketing*, *53*(9), 1689–1699.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics*, *131*(4), 1593–1636. https://doi.org/10.1093/qje/qjw024

Benjamin, V., & Raghu, T. S. (2023). Augmenting Social Bot Detection with Crowd-Generated Labels. *Information Systems Research*, *34*(2). https://doi.org/10.1287/isre.2022.1136

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937

Boe, B., Payne, J., & P., A. D. (2014). *No Title*. The Python Reddit API Wrapper. https://praw.readthedocs.io/en/stable/)

Breiman, L. (2001a). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324)

Breiman, L. (2001b). Statistical Modeling: The Two Cultures. *Statistical Science*.

Chaudhary, R., Bakhshi, P., & Gupta, H. (2020). Volatility in International Stock Markets: An Empirical Study during COVID-19. *Journal of Risk and Financial Management*, *13*(9). https://doi.org/10.3390/jrfm13090208

Deng, S., Huang, Z., Sinha, A. P., & Zhao, H. (2018). The interaction between microblog sentiment and stock returns: An empirical examination. *MIS Quarterly*, *42*(3). https://doi.org/10.25300/MISQ/2018/14268

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* (Vol. 1). https://github.com/tensorflow/tensor2tensor)

Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning*. https://doi.org/10.48550/arxiv.1702.08608

Ensafi, Y., Amin, S. H., Zhang, G., & Shah, B. (2022). Time-series forecasting of seasonal items sales using machine learning – A comparative analysis. *International Journal of Information Management Data Insights*, *2*(1). https://doi.org/10.1016/j.jjimei.2022.100058

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *In KDD (Vol.*, *96*, 226–231.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The Elements of Statistical Learning. In *Springer series in statistics* (Vol. 1, Issue 1). https://doi.org/10.1007/b94608

Garman, E. T., & Forgue, R. (2014). *Personal Finance*. Cengage Learning.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). Bayesian Data Analysis. In *Bayesian Data Analysis* (3rd ed.). CRC press.

Grootendorst, M. (2022). *Neural Topic Modeling with a Class-Based TF-IDF Procedure*. http://arxiv.org/abs/2203.05794)

Hoffman, M. D., & Gelman, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*.

John, K., & Li, J. (2021). COVID-19, volatility dynamics, and sentiment trading. *Journal of Banking and Finance*, *133*. https://doi.org/10.1016/j.jbankfin.2021.106162

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., & Kapadia, S. (2022). Making text count: Economic forecasting using newspaper text. *Journal of Applied Econometrics*, *37*(5), 896–919. https://doi.org/10.1002/jae.2907

Kim, B. (Raymond), Srinivasan, K., Kong, S. H., Kim, J. H., Shin, C. S., & Ram, S. (2023). ROLEX: A Novel Method for Interpretable Machine Learning Using Robust Local Explanations. *MIS Quarterly*, *47*(3), 1303–1332. https://doi.org/10.25300/MISQ/2022/17141

Kim, B., Srinivasan, K., & Ram, S. (2019). Robust local explanations for healthcare predictive analytics: An application to fragility fracture risk modeling. *International Conference on Information Systems*.

Korobov, M., & Lopuhin, K. (2016). *Eli5* (Vol. 5). https://eli5.readthedocs.io/en/latest/

Kress, G. (2009). Multimodality: A social semiotic approach to contemporary communication. In *Multimodality: A Social Semiotic Approach to Contemporary Communication*. https://doi.org/10.4324/9780203970034

Kress, G. R. (2010). *Multimodality: A Social Semiotic Approach to Contemporary Communication*. Taylor & Francis.

Kurumathur, S. K., Bhatt, P., Hariharan, G., Valecha, R., & Rao, H. R. (2022). Examining the Public Response to Vigilantism: A Multi-dimensional Model of Social Media Discourse. *ICIS 2022 Proceedings*. https://aisel.aisnet.org/icis2022/social/social/19

Lundberg, S. M., Allen, P. G., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. https://github.com/slundberg/shap

Maqsood, H., Mehmood, I., Maqsood, M., Yasir, M., Afzal, S., Aadil, F., Selim, M. M., & Muhammad, K. (2020). A local and global event sentiment based efficient stock exchange forecasting using deep learning. *International Journal of Information Management*, *50*, 432–451. https://doi.org/10.1016/j.ijinfomgt.2019.07.011

McInnes, L., Healy, J., & Astels, S. (2017). Hdbscan: Hierarchical Density Based Clustering. *The Journal of Open Source Software*, *2*, 11. https://doi.org/10.21105/joss.00205)

McInnes, L., Healy, J., & Melville, J. (2020). *Uniform Manifold Approximation and Projection for Dimension Reduction*. http://arxiv.org/abs/1802.03426)

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations ofwords and phrases and their compositionality. *Advances in Neural Information Processing Systems*.

Molnar, C. (2020). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. *Book*.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. *NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*.

Ortiz, D. P. (2022). Economic policy statements, social media, and stock market uncertainty: An analysis of Donald Trump's tweets. *Journal of Economics and Finance*. https://doi.org/10.1007/s12197-022-09608-5

Padmanabhan, B., Fang, X., Sahoo, N., & Burton-Jones, A. (2022). Machine Learning in Information Systems Research. *MIS Quarterly*, *46*(1).

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media and Society*, *7*(2). https://doi.org/10.1177/20563051211019004

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In E. Conference (Ed.), *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* (pp. 3982–3992). https://doi.org/10.18653/v1/d19-1410

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. In *Nature Machine Intelligence* (Vol. 1, Issue 5). https://doi.org/10.1038/s42256-019-0048-x

Ryu, P. M. (2018). Predicting the unemployment rate using social media analysis. *Journal of Information Processing Systems*, *14*(4). https://doi.org/10.3745/JIPS.04.0079

Sachin, P. K., Schecter, A., & Li, W. (2022). Emotions in Microblogs and Information Diffusion: Evidence of a Curvilinear Relationship. *ICIS 2022 Proceedings*, 2471. https://aisel.aisnet.org/icis2022/social/social/17

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310.

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly: Management Information Systems*. https://doi.org/10.2307/23042796

Singh Chauhan, S., Srinivasan, K., & Sharma, T. (2022). A trans-national comparison of stock market movements and related social media chatter during the COVID-19 pandemic. *Journal of Business Analytics*. https://doi.org/10.1080/2573234X.2022.2155257

Taylor, S. J., & Letham, B. (2018). Forecasting at Scale. *The American Statistician*, *72*(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B (Methodogical), Pp*, 267–288. https://doi.org/https://doi.org/10.1111/j.2517-6161.1996.tb02080.x)

Velichety, S., & Shrivastava, U. (2022). Quantifying the impacts of online fake news on the equity value of social media platforms – Evidence from Twitter. *International Journal of Information Management*, *64*, 102474. https://doi.org/10.1016/J.IJINFOMGT.2022.102474

Wang, Y., Ramaprasad, J., & Gopal, A. (2022). Dancing to the #challenge: The Effect of TikTok on Closing the Artist Gender Gap. *ICIS 2022 Proceedings*. https://aisel.aisnet.org/icis2022/social/social/7

Xie, J., Chai, Y., & Liu, X. (2023). Unbox the Black-Box: Predict and Interpret YouTube Viewership Using Deep Learning. *Journal of Management Information Systems*, *40*(2), 541–579. https://doi.org/10.1080/07421222.2023.2196780

Yeşiltaş, S., Şen, A., Arslan, B., & Altuğ, S. (2022). A Twitter-Based Economic Policy Uncertainty Index: Expert Opinion and Financial Market Dynamics in an Emerging Market Economy. *Frontiers in Physics*, *10*. https://doi.org/10.3389/fphy.2022.864207