Forecasting Future Economic Uncertainty with Sentiments Embedded in Social Media

Jinhang Jiang[1[0000-0002-3941-5161]], Mei Feng[2[0000-0002-8928-8228]], Kiran Kumar Bandeli[1[0000-0002-2517-4467]] and Karthik Srinivasan[3[0000-0002-1608-6190]]

[1] Walmart Inc., 702 SW 8th St. Bentonville, AR 72716, USA
[2] Biopharmaceutical Innovation & Optimization Center, University of Kansas, 2097 Constant Ave., Lawrence, KS 66047, USA
[3] School of Business, University of Kansas, 1654 Naismith Dr, Lawrence, KS 66045, USA
{Jinhang.Jiang, KiranKumar.Bandeli}@walmart.com
{meifeng, karthiks}@ku.edu

Abstract. This paper suggests a method for predicting the Economic Uncertainty Index in the Equity Market by analyzing social media data from Reddit. The proposed framework includes various custom pre-processing and analytics techniques, such as BERTopic for identifying latent topics and regularized linear models. Using this framework, this study conducts a sophisticated descriptive and explanatory analysis of a vast collection of Reddit posts about personal finance. The research provides valuable insights into Reddit's discussion topics and their ability to forecast economic uncertainty accurately. Overall, this study highlights the potential of explainable deep learning and social media data to enhance economic decision-making and forecasting.

Keywords: Topic Modeling, Unstructured Information Extraction, Social Media Analytics, Computational Economics, Explainable Deep Learning, Economic Uncertainty Analysis.

1 Introduction

Social media platforms have become an invaluable source of information for studying various social, economic, and political phenomena in modern society [1–3]. 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 investors' sentiments toward public stocks and market indices [4, 5]. Social media data has also been studied to characterize individual and organizational outcomes such as consumer behavior and firm equity value [6, 7]. 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 [8] as a regression problem. This study aims to investigate the relationship between public sentiment and the EUI. While previous studies relied solely on parsing news articles or social media posts, this study uses Reddit discussions as a richer source of information that captures the shift of the public's concerns. Understanding the relationship between the social media discourse and EUI variations can contribute to research on predicting economic trends, benefiting macroeconomists, policymakers, firms, and investors.

Our study used Reddit conversations discussing personal finance from late 2021 to mid-2022 to predict future economic uncertainty. We present three contributions: (1) a new dataset containing more than a million Reddit posts focusing on personal finance issues, (2) a unique deep learning method to extract topics from Reddit for modeling economic uncertainty, and (3) identification of topics whose daily volume on social media platforms may predict future economic uncertainty.

2 Background

The EUI index is based on the idea that newspaper text searches can provide valid proxies for economic and policy conditions, particularly in countries with limited data sources or earlier periods. The authors showed that the EUI could effectively capture economic uncertainty stretching back several decades [8]. In particular, Equity Market-related Economic Uncertainty Index is a type of EUI that explicitly measures the level of uncertainty in the market. Increased levels of economic uncertainty can result in heightened volatility in the market, as investors may be more hesitant to make decisions in an uncertain environment. Therefore, tracking the EUI can provide valuable insights for economists and financial analysts, helping them make more informed decisions when it comes to investing and trading.

Social media platforms are rich with information about individual, organizational, and social behavior across time and geographies. Studies have examined causal and associative links between social media and stock market uncertainty [9–11]. 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 [12, 13]. 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 [14]. Reddit is a discussion forum with users focusing on detailed conversations based on recent events and news articles [15]. Previous research has either focused on microblogging platforms like Twitter or analyzed the macroeconomic indicators with news articles. Our study is the first to explore and model the possible link between social media conversations and EUI using interpretable deep learning methods.

In our study, we explore the use of social media data to predict economic uncertainty. Selecting an appropriate data source is crucial, and after reviewing publicly

available data repositories, we found that Facebook and LinkedIn were unsuitable due to their emphasis on personal and business networking rather than public broadcasts of socio-political opinions. While Twitter is commonly used for opinion-sharing, its conversational nature makes it challenging to collect large amounts of data on a specific topic. In contrast, Reddit posts are typically longer and more descriptive, making them a more suitable data source for our study.

Literature in personal finance indicates a relationship between economic conditions and individual financial success [16], and reciprocally, it is hypothesized that personal financial decisions and concerns may be reflected in the overall economic conditions, and this subreddit's large and active community can provide us with valuable insights into this relationship. Therefore, we selected the "Personal Finance" subreddit as our data source, given its focus on topics related to budgeting, saving, investing, and retirement planning, which we believe may be linked to broader economic issues. This subreddit has 17.1 million users and is the twenty-second largest subreddit on Reddit, making it a valuable data source for our study.

3 Data processing

From November 23rd, 2021, to June 25th, 2022, we leveraged the Python Reddit API Wrapper, 'PRAW' [17], to scrape the posts and replies within the "Personal Finance" subreddit at one-hour intervals. This resulted in a data set spanning 215 days. 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.

Table 1. Data Summary

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

We performed multiple steps of data cleaning to prepare the dataset for modeling. These steps included removing documents with fewer than ten 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.

We first identified topics across reddits (i.e., Reddit posts), using BERTopic [18], 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 user activity level 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 215 days, resulting in 215 observations used for training and validation.

4 Methods

In this section, we provide details of our method framework.

4.1 Topic Extraction

BERTopic is a topic modeling method based on word embedding [19] and transformer technologies [20]. 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 [21]. 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 [18]. 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 compete with other classical and more recent topic modeling approaches across various benchmarks. Identifying the optimal number of topics using metrics such as perplexity or coherence [22] can be computationally intensive. Therefore, we propose comparing model configurations' performance with different topic counts with our framework.

4.2 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. The Least Absolute Shrinkage and Selection Operator, i.e., Lasso, is a popular regularization-based feature selection method [23]. Lasso accomplishes variable selection by shrinking insignificant predictor variables' coefficients to zero, particularly in high-dimensional data.

4.3 Cross Validation

While the Lasso helps eliminate 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 following sub-section or used as the final predictive model.

Fig. 1 summarizes our proposed framework for using topics to measure and predict future economic uncertainties.

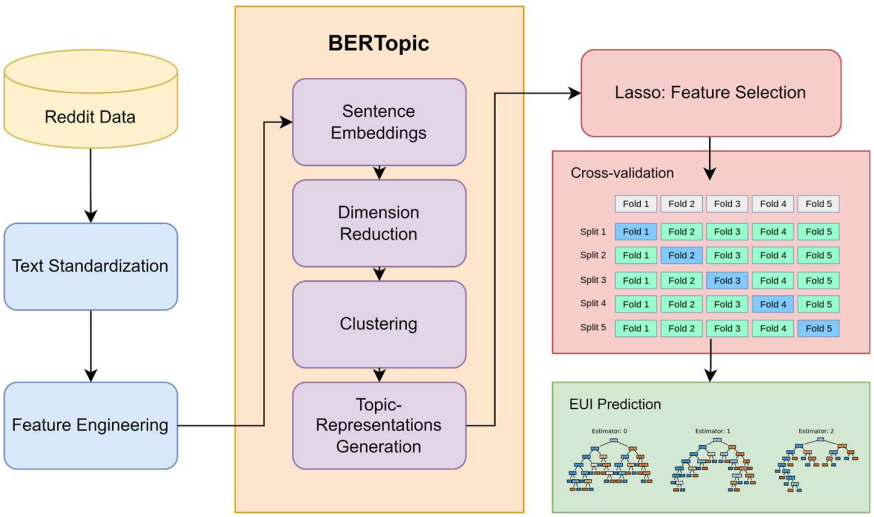

Fig. 1. Method Framework for Determining Economic Uncertainty Using Reddit Text

5 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 [24], reduced the embedding dimensions using the UMAP procedure [25], clustered the documents using the HDBSCAN method [26, 27] and finally derived the topic representations using the c-TF-IDF method [18]. Note that the base method of BERTopic [18] offers the flexibility to change the components of our framework to any widely used software (i.e., K-means can be used 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.

We aggregated topic groups for each date to analyze the daily topic frequency. However, computing optimal topic numbers based on perplexity was computationally challenging using BERTopic. We compared model performance using fixed topic numbers of 10, 30, 50, 100, and 200 to address this. Additionally, we used the default clustering algorithm to generate 2504 topics automatically, resulting in six datasets for analysis.

We conducted experiments with six sets of data and three future periods of EUI, leading to 18 model combinations for comparison. Due to the large number of features (up to 2500+), we used the Lasso method to select contributing predictors for the input feature set. Next, we trained a Random Forest Regressor with default parameters in Python's *scikit-learn* package using 10-fold cross-validation for the feature ranking step in our framework. We used Symmetric Mean Absolute Percentage Error (SMAPE) to evaluate the accuracy of our predictions. SMAPE is a suitable metric because it treats both positive and negative errors symmetrically and favors over-forecasting, which aligns with our objective of presenting conservative projections. The percentage scale of SMAPE is easy to understand and resistant to the influence of extreme values.

The results of topic modeling are shown in Table 2. The estimators with the most topics yield the best in predicting the EUI one day after the posts were made, with an average SMAPE of 45.65% and a standard deviation of 1.19%. We performed feature selection using Lasso and found that combining the best performance resulted in a final count of 194 selected features, representing approximately 92.27% of the total number of features being filtered out. We also do not see a linearity between the number of topics and the model performance, and thus we moved forward with the topics generated from the default method in the remaining experiments.

Table 2. Cross Validation Results – SMAPE

Num. of 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**

We implemented permutation importance to measure the importance of selected features concerning EUI one day after posts are made with the help of a Python library called Eli5 [28]. Permutation feature importance is a model evaluation method that can be applied to any fitted estimator with tabular data, which is particularly useful for non-linear or complex models. It is defined as the decrease in model performance, measured by evaluation metrics like F1 or R^2 when a single feature value is randomly shuffled [29]. 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" [28]. Fig. 2 displays the permutation importance for the top 10 topics after 100 iterations with 10-fold cross-validation.

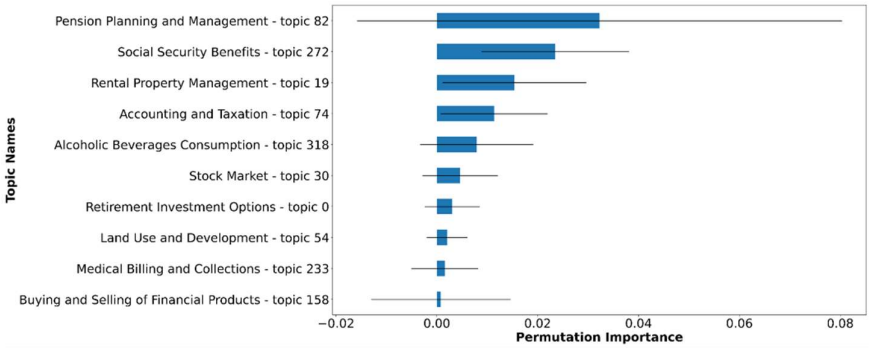

Fig. 2. Permutation results for the top ten features

6 Discussion and Conclusions

Our framework shows the best performance in predicting EUI scores one day after the posts or replies on the "Personal Finance" subreddit, based on SMAPE metrics. The data set with the most topics automatically optimized by the BERTopic algorithm outperforms other datasets in five out of six instances. The Lasso algorithm successfully filters out 92.27% of the total features, and we identified ten topics that have the most significant impact on the next day's EUI scores after conducting permutation analysis.

Following the findings from Fig. 2, we dive into the details and analyze the topics contributing 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 financial planning and management assistance. 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. Conversations over buying and selling financial products and stock markets are also the core topics as people's behavior is influenced by their perception of current and future economic conditions. The topic of alcoholic beverages 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, an individual's perception of the current and future economy is captured.

The proposed framework presents a practical approach for collecting data, identifying key topics, and predicting future EUI scores. However, the current implementation has limitations with respect to its forecasting capabilities. Due to the inherent time series nature of EUI, using a Random Forest Regressor with N-fold cross-validation may potentially result in data leakage issues. To address this limitation, future work will focus on incorporating traditional time series models and deep neural networks as alternative forecasting methods. By doing so, we hope to better account for the temporal nature of the EUI scores and potentially enhance the accuracy and robustness of our predictions. In addition, the proposed framework will be enhanced by integrating the selected topics as additional regressors to improve its overall design. These developments will pave the way for more accurate and reliable predictions of EUI scores and inform policymakers with more explainable and interpretable insights derived from the sentiments embedded in social media.

It is essential to clarify that our study does not claim any causal relationship between the topics and the EUI. Instead, our focus lies in the predictive aspect. We observe that individuals share financial and economic-specific issues on the forum, and the sentiments expressed in their posts often result from various social or economic events. While individual cases may not carry significant predictive value, aggregating similar topics and sentiments can be indicative and predictive of a potential shift in future economic trends. In summary, the posts on the forum do not cause changes in the EUI, but they can reflect in advance the possibility of such changes. Our study emphasizes prediction rather than establishing causal links.

Our study proposes a framework to predict the Equity Market-related Economic Uncertainty Index using data from the "Personal Finance" subreddit. We conducted 18 modeling experiments with varying numbers of topics and future EUI scores. The best estimator was found to be the one with 2504 original topics generated by BERTopic for predicting EUI one day after posts and replies were made. Retirement investment, social security benefits, renting, pension, medical bills, etc., were found to be the top topics influencing prediction performance. Our study suggests that the volume of posts on these topics can indicate economic fluctuations and online special-interest groups may have insights into future economic conditions.

References

1. Kurumathur SK, Bhatt P, Hariharan G, et al (2022) Examining the Public Response to Vigilantism: A Multi-dimensional Model of Social Media Discourse. ICIS 2022 Proceedings
2. Gunaratne C, De D, Thakur G, et al (2022) Evolution of Intent and Social Influence Networks and their Significance in Detecting COVID-19 Disinformation Actors on Social Media
3. Moraffah R, Unni SJ, Raglin A, Liu H (2022) Causal Data Fusion for Multimodal Disaster Classification in Social Media

4. Maqsood H, Mehmood I, Maqsood M, et al (2020) A local and global event sentiment based efficient stock exchange forecasting using deep learning. Int J Inf Manage 50:432–451. https://doi.org/10.1016/j.ijinfomgt.2019.07.011
5. 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
6. Ashish KR, Bezawada, Kumar A, et al (2018) From Social to Sale: The Effects of Firm Generated Content in Social Media on Customer Behavior. J Mark 53:1689–1699
7. Velichety S, Shrivastava U (2022) Quantifying the impacts of online fake news on the equity value of social media platforms – Evidence from Twitter. Int J Inf Manage 64:102474. https://doi.org/10.1016/J.IJINFOMGT.2022.102474
8. Baker SR, Bloom N, Davis SJ (2016) Measuring economic policy uncertainty. Quarterly Journal of Economics 131:1593–1636. https://doi.org/10.1093/qje/qjw024
9. Ortiz DP (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
10. John K, Li J (2021) COVID-19, volatility dynamics, and sentiment trading. J Bank Financ 133:. https://doi.org/10.1016/j.jbankfin.2021.106162
11. 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:. https://doi.org/10.3390/jrfm13090208
12. Ryu PM (2018) Predicting the unemployment rate using social media analysis. Journal of Information Processing Systems 14:. https://doi.org/10.3745/JIPS.04.0079
13. Kalamara E, Turrell A, Redl C, et al (2022) Making text count: Economic forecasting using newspaper text. Journal of Applied Econometrics 37:896–919. https://doi.org/10.1002/jae.2907
14. 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. Front Phys 10:. https://doi.org/10.3389/fphy.2022.864207
15. Proferes N, Jones N, Gilbert S, et al (2021) Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. Social Media and Society 7:. https://doi.org/10.1177/20563051211019004
16. Garman ET, Forgue R (2014) Personal Finance. Cengage Learning
17. Boe B, Payne J, P. AD (2014) No Title. The Python Reddit API Wrapper, PRAW
18. Grootendorst M (2022) Neural Topic Modeling with a Class-Based TF-IDF Procedure. BERTopic
19. Mikolov T, Sutskever I, Chen K, et al (2013) Distributed representations ofwords and phrases and their compositionality. In: Advances in Neural Information Processing Systems

20. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding
21. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. The Journal of Machine Learning Research 3:993–1022
22. Newman D, Lau JH, Grieser K, Baldwin T (2010) Automatic evaluation of topic coherence. In: 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
23. Tibshirani R (1996) Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society, Series B (Methodogical), pp 267–288
24. Reimers N, Gurevych I (2019) Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Conference E (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
25. McInnes L, Healy J, Melville J (2020) Uniform Manifold Approximation and Projection for Dimension Reduction. UMAP
26. 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
27. McInnes L, Healy J, Astels S (2017) Hdbscan: Hierarchical Density Based Clustering. The Journal of Open Source Software 2:11
28. Korobov M, Lopuhin K (2016) Eli5. 5
29. Breiman L (2001) Random Forests. Mach Learn 45:5–32