Integrating Social Media and Administrative Data for the Real-Time Prediction of the Consumer Confidence Indicator

Akvilė Vitkauskaitė[1,2], Andrius Čiginas[1,2]

[1]State Data Agency (Statistics Lithuania), Lithuania

[2]Vilnius University, Lithuania

Abstract

The Consumer Confidence Index (CCI) measures public sentiment about the economy through a survey of four questions regarding household finances, economic outlook, and spending plans over the past and coming year. It is calculated as the arithmetic mean of these responses.The main objective of this study is to nowcast and forecast the CCI. The aim is to estimate the current month's CCI values faster than those obtained using the traditional survey methodology, which usually provides results at the end of the month. For instance, while the official CCI for November would typically be available in the last few days of November, this research aims to provide an early estimate at the beginning of November, utilizing data collected at the start of the month. This is achieved by combining key economic indicators with historical CCI values. The research includes examining the relationship between traditional survey-based indicators and consumer sentiment expressed on social media platforms. Social media expressions, particularly from X (Twitter), are analyzed through its official API. The sentiment analysis of tweets has enabled us to create a Social Media Indicator (SMI) that offers a distinct advantage in our predictive models. To improve forecast accuracy, we include Google Trends data, which provides additional insights into consumer search behavior and related trends in economic confidence. In addition, the study explores the possibility of integrating key economic indicators from administrative data, such as inflation rate, income statistics, and unemployment. In general, obtaining data for research from popular social platforms such as Facebook and Instagram is not possible due to stringent privacy policies and data protection regulations. Nevertheless, data are easily and legally available from X, but this platform is not so popular in Lithuania. Therefore, the representativeness of X data raises special issues. Taking everything into account, by combining traditional economic indicators with advanced sentiment analysis from X and Google Trends data, the study seeks to deliver prompt CCI predictions ahead of standard survey timelines.

Keywords: Consumer confidence, X, Google Trends, Sentiment Analysis, Nowcasting

1. Introduction

Consumer confidence indicator (CCI) is a vital economic measurement that influences the decision-making processes of policymakers, businesses, and investors, providing valuable insights into individuals' sentiments and expectations regarding the state of the economy (Islam & Mumtaz, 2016). Traditionally, CCI has been derived from survey data, capturing the

opinions and perceptions of individuals through structured questionnaires (Mueller, 1963; The Conference Board, 2021).

With the rise of social media platforms and the abundance of user-generated content, there is a new opportunity to assess consumer sentiment in new ways. By analyzing sentiment from platforms like blogs, forums, and social media, new insights into public attitudes can be gained, which are useful for market research and business strategies (Aishwarya, Ashwatha, Deepthi, & Raja, 2019).

Further, integrating administrative data such as employment statistics and income records can enhance the accuracy and timeliness of CCIs. The study by Curtin (2007) emphasizes the significant relationship between changes in the unemployment rate and consumer sentiment, highlighting the powerful influence of employment conditions on consumer expectations.

Our study seeks to bridge the gap between traditional and emerging data sources by utilizing social media, Google Trends, and administrative data to measure and forecast CCI. We employed models like SARIMAX, VECM, Random Forest, and XGBoost, finding that XGBoost provided the highest accuracy, with SARIMAX also performing well.

2. Theoretical framework

2.1 Historical development of consumer confidence indicators

By the mid-twentieth century, consumer confidence measurements emerged as key predictors of economic trends, recognized by businesses and economists for their ability to forecast market behaviour and understand consumer choices (Logemann, 2020).

The development of the CCI involved collaboration across various organizations. George Katona and Rensis Likert pioneered the initial methods aimed at measuring consumer expectations to analyze spending and saving patterns in the late 1940s (Katona & Likert, 1946). Mueller (1963) played a significant role in advancing consumer confidence indices by introducing the concept of using them to forecast consumption patterns. Mueller (1963) further advanced CCIs by using them to forecast consumption, analyzing ten years of Michigan consumer survey data. Her research highlighted the crucial role of consumer confidence in spending habits through regression models considering previous consumption (Mueller, 1963).

The Consumer Confidence Board in the USA, crucial in developing the CCI since 1967, surveys individual perspectives on the economy and employment. This index is a key indicator of U.S. economic strength. In May 2021, the administration of the survey shifted from The Nielsen Company to Toluna, a tech company with a 36 million-member panel. Previously, until November 2010, TNS conducted the survey by mail (The Conference Board, 2021). Eurostat,

the EU's statistical office, oversees the monthly administration of the CCI across member countries (Eurostat, 2023). Since May 2001, Statistics Lithuania has conducted similar surveys, aiming to gather data on consumer purchase intentions, saving capabilities, and perceptions of the economic situation and its influence on their intentions (State Data Agency (Statistics Lithuania), 2023).

2.2 The role of social media data and Google Trends data in estimating consumer confidence

The integration of social media data has significantly enhanced consumer confidence indicator estimation, especially evident during the COVID-19 crisis when platforms like Twitter became crucial for real-time public sentiment analysis (Baldacci et al., 2022). For instance, Istat's Social Mood on Economy Index in Italy combines social media insights with traditional surveys to offer a nuanced view of consumer confidence (Catanese et al., 2022). Similarly, research by van den Brakel et al. (2017) introduced a multivariate time series model using social media for the Dutch Consumer Confidence Survey, improving estimate accuracy and timeliness [6]. Likewise, Austin et al. (2022) demonstrated the effectiveness of Google Trends data in predicting economic trends during critical periods.

2.3 Administrative data as auxiliary data for estimating CCI

Nowzohour and Stracca (2017) report that the CCI correlates strongly with general economic expectations, unemployment, and financial conditions, showing positive associations with future inflation and growth, and negative associations with unemployment and interest rates. Demirel and Artan (2017) use panel causality analysis to reveal causality from exchange rates and interest rates to economic confidence and from economic confidence to unemployment in the EU. Their findings underline confidence's crucial role in economic dynamics, influencing production, consumption, and inflation. This research supports using auxiliary data like unemployment figures to refine CCI forecasting accuracy, highlighting the correlations that enhance forecasting potential.

3. Methods

3.1 Data and dataset

This dataset combines public sentiment from social media, and Google trends data with key economic indicators. It contains monthly data from 2018 until November 2023, including the COVID-19 pandemic and geopolitical developments - Russia's invasion of Ukraine in February 2022. The observed declines in the CCI correspond to these periods of global and regional turbulence.

3.1.1 Social media data (SMI)

The Social Media Indicator (SMI) refines nowcasting for Lithuania's CCI using real-time X (Twitter) data. This index is derived from tweets, collected weekly via API using 101 economically focused keywords, including historical tweets to overcome the API's seven-day limit. Preprocessing involved cleaning texts, and removing special characters, spaces, and non-Lithuanian tweets. The sentiment was analyzed using NLP tools like TextBlob, Vader, Afinn, Transformers, and Flair, each providing unique interpretive strengths. Average sentiment scores are calculated monthly for active Twitter users who tweet more than once a month. Each tweet is weighted equally, ensuring a balanced view of economic sentiment. Sentiment categories are positive, neutral, or negative (for Flair: positive or negative). The final SMI for the month is determined by calculating the balance of these sentiments. The monthly SMI is computed by balancing positive and negative sentiments, converting this balance into a percentage, and adjusting to match the CCI's scale, as raw SMI values are typically ten times larger than the CCI's.

3.1.2 Google Trends

Google Trends data was manually collected using 'economics' as a control theme to standardize and compare search interest for 101 keywords, complying with platform limits and avoiding scraping issues. This method maintained dataset integrity and complied with Google Trends' terms of service by mimicking typical user behavior. Data for each keyword was manually downloaded and compiled into a monthly search interest dataset.

The data was analyzed using correlation analyses of specific three-word keyword combinations. To calculate GT+ and GT-, values from each of the three keyword columns were added and then averaged. The second-highest positively correlated set, 'purchases, work, costs' ('pirkiniai, darba, kainuoja' in Lithuanian), was selected as GT+ to ensure distinctiveness from GT-, which uses the top negatively correlated keywords 'buy, unemployed, inflation' ('pirkti, bedarbiams, infliacija'). This method maintains unique variables and prevents overlap in the keyword combinations.

3.1.3 Administrative data

The dataset integrates key economic indicators from Lithuanian official sources to provide a comprehensive economic context alongside social media insights. This includes monthly interpolated data on average wages and state pensions, an inequality indicator derived from the wage-pension ratio, and both seasonally adjusted and standard unemployment rates. The Consumer Price Index (CPI) offers a monthly view of inflation, reflecting cost-of-living changes. These metrics collectively enhance the analysis by offering

traditional economic signals that complement the sentiment captured through social media, providing a robust foundation for understanding consumer confidence in Lithuania.

3.2 Evaluation metrics

We assess model performance using key statistical metrics: Mean Absolute Error (MAE) for average prediction errors, Mean Squared Error (MSE), which averages squared differences to capture error variance and Root Mean Squared Error (RMSE), a standard deviation of errors. Additionally, the Akaike Information Criterion (AIC) is specifically used for comparing SARIMAX models, helping balance model complexity against the likelihood of the fit. We avoid using Mean Absolute Percentage Error (MAPE) due to its potential distortion when predicting values near zero, such as with our CCI which oscillates around zero. These metrics allow us to effectively evaluate and enhance our forecasting models.

3.3 Models development and implementation

This section covers four models used to forecast the CCI: SARIMAX, VECM, Random Forest, and XGBoost. It outlines the selection rationale for these models, key exogenous variables, and their influence on forecasting accuracy. Models are developed and evaluated by dividing the data into training, validation, and testing segments, each representing 75%, 15%, and 10% of the data. We evaluate model accuracy using a rolling forecast method that involves predicting the next step and then updating the model with the latest information before making another prediction to minimize the accumulation of forecast errors. Parameters for each model were optimized with training and validation data and then tested on unseen data to ensure forecast reliability and robustness.

3.3.1 SARIMAX

The SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables) model, an advancement of the ARIMA model, is highly effective in time series analysis by incorporating both seasonality and external factors. This model is structured with non-seasonal components denoted by (p,d,q) for the AR order, differencing order, and MA order respectively, and seasonal components (P,D,Q,s) representing the seasonal AR order, differencing order, MA order, and seasonal cycle length. A key feature of the SARIMAX model is the inclusion of 'X', signifying exogenous variables. These are external predictors or input variables that can influence the target variable being forecasted. Exogenous variables provide additional external information that can enhance the model's predictive capability (Peixeiro, 2022). Modelling starts with ensuring data stationarity using the Augmented Dickey-Fuller test and sequential differencing, supported by Seasonal and Trend Decomposition using Locally Estimated Scatterplot Smoothing (STL) for visualizing seasonal patterns and trends. The

decomposition shows some seasonal effects with irregularities, supporting the decision to use the SARIMAX modelling approach. Exogenous variables are chosen using Johansen's cointegration test to ensure long-term equilibrium with the CCI, ideal for I(1) time series (Johansen, 1995). These variables are further refined through Spearman's rank correlation test to identify optimal lags. Model parameters are selected to minimize AIC and MAE, enhancing accuracy and preventing overfitting. Residual analysis, including Q-Q plots and correlograms, confirms fit accuracy and normal, uncorrelated residuals. The rolling forecast implementation is fine-tuned using the training and validation data before the final evaluation of unseen test data.

3.3.2 VECM

The Vector Error Correction Model (VECM) analyzes CCI using cointegrated time series data, utilizing methods analogous to SARIMAX, such as stationarity checks and integration of exogenous variables. Essential to VECM is its focus on long-run equilibrium relationships among variables, highlighted by Engle and Granger (1987). According to them, cointegration indicates that certain sets of variables cannot significantly deviate from each other over time.

3.3.3 Random Forest and XGBoost

Random Forest (RF) and XGBoost, advanced machine learning techniques, excel in managing complex datasets and identifying key features. Breiman (2001) explained that the RF algorithm combines the outputs of numerous decision trees to get a single and more precise forecast. This helps avoid overfitting. XGBoost (eXtreme Gradient Boosting) is an advanced implementation of gradient boosting algorithms. It has become a very popular tool for machine learning competitions. XGBoost uncovers non-linear patterns in time series data and optimizes performance with techniques like early stopping (Chen & Guestrin, 2016; Woloszko, 2020). XGBoost and RF manage feature importance to assess each variable's impact on predictions. Initially, both models train on variables and their lags up to six times, refining feature selection based on their calculated importance during rolling forecasts. As the models process this data, they use their ability to estimate and rank the importance of each feature. Due to the implemented rolling forecasting method, the average importance is calculated in all stages. This approach allows for dynamically adjusting the model's focus on features that consistently show strong predictive power over different forecast periods. After determining the most influential features, the next phase involves refining these features further by applying different rolling window sizes to each. Finally, with this optimized feature set and their respective rolling averages, the models are tested on unseen test data.

4. Results and Conclusions

In this study, we aimed to predict the Consumer Confidence Index (CCI) using four models: SARIMAX, VECM, Random Forest, and XGBoost, integrating social media and administrative data. Model configurations and performance are illustrated in Figure 1, which compares actual CCI values against model predictions. Performance metrics, specifically Mean Absolute Error (MAE), are detailed in Table 1, which also lists each model's exogenous variables. In Table 1, 'periods' refer to the number of months over which rolling averages are calculated, smoothing out short-term fluctuations to highlight longer-term trends in the data. 'Lag' indicates the delay, in months, between the data point and its impact on the model, helping to account for time-dependent effects. XGBoost emerged as the top performer in accuracy, followed by SARIMAX. Random Forest and VECM were less accurate but still provided useful predictive insights.

Figure 1: Final models for CCI forecasting: SARIMAX(5,1,6,1,0,1,12), VECM, RF and XGBoost

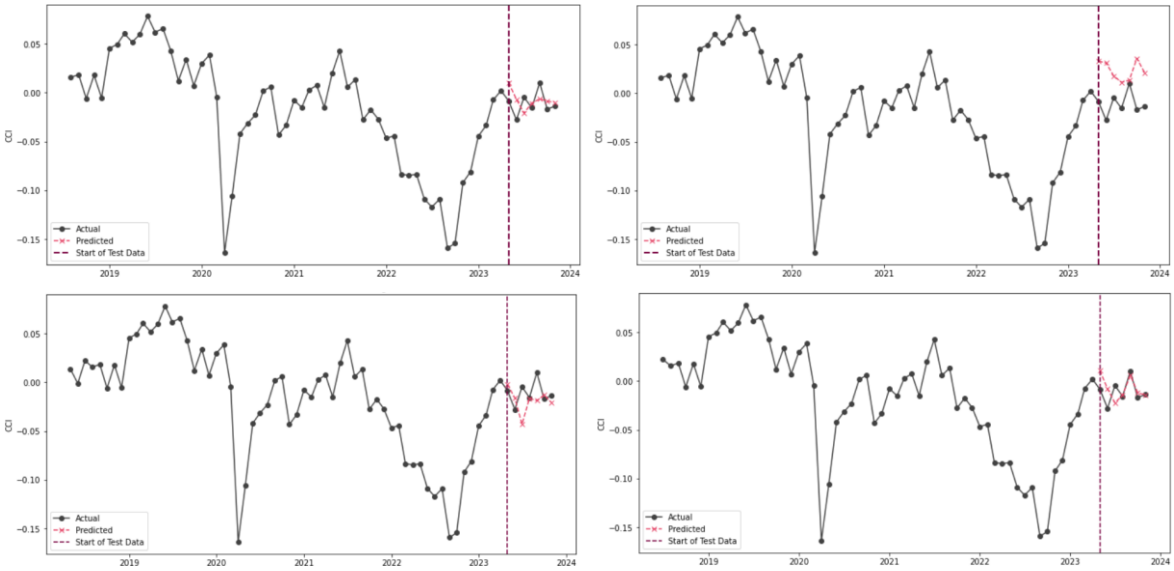

Table 1: Exogenous variables (features) and MAE in final forecasting models

Model name	Exogenous variables (features)	MAE
SARIMAX	Vader_SI_diff_1, CCI_diff_1, GT-_diff_1, Inflation_diff_1 (3 months lag), Unemployment_without_seasonality_diff_1 (current, 6 months lag)	0.0127
VECM	Vader_SI_diff_1, Inflation_diff_1, GT-_diff_1 (all up to 1 month lag)	0.0343
Random Forest	CCI (1 month lag; 2 periods), GT (current and 1 month lag; 2 periods), GT- (current and 1 month lag; 4 periods)	0.0139
XGBoost	GT- (current and 1 month lag; 2 periods), GT+ (1 month lag; 6 periods), Average wage (current; 4 periods), Inflation (3 months lag; 4 periods), Unemployment_without_seasonality (5 months lag; 6 periods)	0.0098

References

Aishwarya, R., Ashwatha, C., Deepthi, A., & Raja, J. B. (2019). A Novel Adaptable Approach for Sentiment Analysis. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *5*(2).

Austin, M. P. A., Austin, P., Marini, M. M., Sanchez, A., Simpson-Bell, C., & Tebrake, J. (2021). *Using the Google Places API and Google Trends data to develop high frequency indicators of economic activity*. International Monetary Fund.

Baldacci, E., Braaksma, B., Gálvez, A., Giannakouris, K., Olmos, B. G., Rivière, P., ... & Vertanen, V. (2022). Innovation during the COVID-19 crisis: Why it was more critical for official statistics than ever. *Statistical Journal of the IAOS*, *38*(2), 399-412.

Van den Brakel, J., Söhler, E., Daas, P., & Buelens, B. (2017). Social media as a data source for official statistics; the Dutch Consumer Confidence Index. *Survey Methodology*, *43*(2), 183-210.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Catanese, E., Scannapieco, M., Bruno, M., & Valentino, L. (2022). Natural language processing in official statistics: The social mood on economy index experience. *Statistical Journal of the IAOS*, *38*(4), 1451-1459.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Curtin, R. (2007). Consumer sentiment surveys: worldwide review and assessment. *Journal of business cycle measurement and analysis*, *2007*(1), 7-42.

Demirel, S. K., & Artan, S. (2017). The causality relationships between economic confidence and fundamental macroeconomic indicators: Empirical evidence from selected European Union countries. *International Journal of Economics and Financial Issues*, *7*(5), 417.

Engle, R. F., & Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, 251-276.

Eurostat. (2023). *Business and consumer surveys (source: DG ECFIN) (ei_bcs)*.

Islam, T. U., & Mumtaz, M. N. (2016). Consumer Confidence Index and Economic Growth: An Empirical Analysis of EU Countries. *EuroEconomica*, *35*(2).

Johansen, S. (1995). Likelihood-Based Inference in Cointegrated Vector Autoregressive Models Oxford University Press. *New York*.

Katona, G., & Likert, R. (1946). Relationship between consumer expenditures and savings: the contribution of survey research. *The Review of Economics and Statistics*, *28*(4), 197-199.

State Data Agency (Statistics Lithuania). (2023). *Consumer survey*. Retrieved from https://osp.stat.gov.lt

Logemann, J. (2020). Measuring and Managing Expectations: Consumer Confidence as an Economic Indicator, 1920s–1970s. *Futures Past. Economic Forecasting in the 20th and 21st Century*.

Mueller, E. (1963). Ten years of consumer attitude surveys: Their forecasting record. *Journal of the American Statistical Association*, *58*(304), 899-917.

Nowzohour, L., & Stracca, L. (2020). More than a feeling: Confidence, uncertainty, and macroeconomic fluctuations. *Journal of Economic Surveys*, *34*(4), 691-726.

Peixeiro, M. (2022). *Time series forecasting in python*. Simon and Schuster.

The Conference Board. (2021). *Consumer confidence survey technical note – May 2021*.

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010, May). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the international AAAI conference on web and social media* (Vol. 4, No. 1, pp. 178-185).

Woloszko, N. (2020). Adaptive Trees: a new approach to economic forecasting.