Predicting Unemployment Rates Using Textual Data

Bai Mian, Cai Mingyi, Li Yiqi, Wang Bi, Wu Yue

April 19th 2024

1 Introduction

Unemployment rate is a critical macroeconomic indicator with significant economic and psychological impacts on individuals actively seeking employment. High unemployment rates can lead to increased poverty, crime rates, and social discontent. Predicting unemployment allows economists to better understand economic cycles, anticipate recessions or recoveries, and assist both governments and the private sector in making sound economic decisions.

The significance of unemployment rate forecasts was recognized in the 1970s during the stagflation crisis, when high inflation coincided with high unemployment and slow economic growth Brunner et al. (1980); Grubb et al. (1982). Traditional time-series models were initially proposed for predicting unemployment rates. Floros (2005) used GARCH for the casestudy of UK. Jelena et al. (2017) predicted unemployment rates for the EU-28 countries using ARIMA. Recently, economists have turned to machine learning models that lack these drawbacks. Smalter Hall and Cook (2017) used four different neural network architectures to predict U.S. unemployment rates, with results surpassing those from the Survey of Professional Forecasters (SPF) benchmark. Kreiner and Duca (2020) also used Artificial Neural Networks (ANN) to predict U.S. unemployment rates, and their results similarly exceeded the benchmark. Katris (2020) employed a FARIMA model with GARCH errors and utilized Artificial Neural Network supported vector regression to predict unemployment rates in 22 European countries.

Our study aims to determine whether using news sentiment variables to predict unemployment rates can achieve prediction accuracy comparable to that of macroeconomic variables and potentially improve upon it. Following Barbaglia et al. (2021)'s approach in "Forecasting GDP in Europe with Textual Data," we use the original database from their study. We employ the FiGAS method, considering only words related to terms of interest in sentences to calculate sentiment related to economic themes, constructing sentiment indicators for different aspects of economic activity on a daily basis (Consoli et al., 2022). Focusing on the UK, we designed six sentiment indicators capturing attitudes in news about the economy, unemployment, inflation, manufacturing, finance, and monetary policy, using monthly data from January 1998 to December 2021. Our out-of-sample window spans from May 2010 to December 2021, forecasting unemployment rates for the next month, quarter, half-year, and year.

In this paper, we categorize variables into three groups for analysis: those involving only sentiment indices, those involving only macroeconomic variables, and a fusion of both, to evaluate the predictive accuracy of sentiment indices on unemployment rates. We employ advanced machine learning techniques such as LASSO and Random Forest (RF), and compare them with benchmark models such as Autoregressive (AR) and Random Walk (RW) models. Based on basic results, the Diebold-Mariano (DM) test is employed to explore the relative performance of AR, RF, as well as Elastic Net (ElNet) and RF models.

Regarding model performance, the AR model, serving as the benchmark, outperforms other machine learning models across multiple time dimensions. Among the machine learning approaches utilized in this study, Random Forest and bagging rolling window models demonstrate the best performance and are more stable across different forecasting horizons. Additionally, the Elastic Net model, using the Bayesian Information Criterion (BIC) for selection, slightly outperforms other models across multiple time dimensions. We also explore the differences between fixed window and sliding window approaches, finding that sliding window models generally provide superior time series forecasts regardless of the model employed. The final results indicate that real-time sentiment indices have an independent value in predicting unemployment rates in the UK and show an incremental effect in the predictive capabilities of combined sentiment and macroeconomic indices across most models. Furthermore, forecasts over shorter horizons typically outperform those over longer ones.

This paper proposes a novel unemployment rate forecasting model that integrates news sentiment variables with traditional macroeconomic indicators, enhancing its analytical dimension. Advanced machine learning techniques are employed to effectively process and analyze data. This integration improves forecast accuracy and enhances the model's resilience to economic fluctuations, resulting in a more robust and practical predictive framework.

2 Data

This study utilizes a dataset comprising two distinct segments: Sentiment variables and Macroeconomic indicators.

The sentiment variables are derived from a paper titled "Forecasting GDP in Europe with textual data" (Barbaglia et al., 2021). This dataset is publicly accessible and can be downloaded from the designated website associated with the original publication[1]. Since it is daily data, we aggregate the sentiment indicators from the daily to the monthly frequency by averaging the values within the month.

The Macroeconomic variables are derived from the UK-MD database, which is a large monthly macroeconomic dataset designed for empirical analysis in data-rich environments. The dataset contains 109 monthly macroeconomic and financial indicators divided into nine categories: labor, production, retail and services, consumer and retail price indices, producer price indices, international trade, money, credit and interest rates, stock market, and finally sentiment and leading indicators. The dataset is updated in real-time every month through the UK database and is available from Michael McCraken's webpage[2].

After merging the two datasets, we obtained the set of variables needed for this study. We use the data up to December 2021. Our sample ranges from January 1998 to

[1] https://journaldata.zbw.eu/dataset/forecasting-gdp-in-europe-with-textual-data
[2] https://www.stevanovic.uqam.ca/DS_UKMD.html

December 2021, with a total of 288 observations. Since we are forecasting the unemployment rate, we removed the unemployment rate-related variables from the macroeconomic dataset [3] and only used the variables with all observations during the sample period (106 variables). Additionally, we also included four principal component factors calculated from this set of variables as potential predictors. We consider four lags of all variables, as well as four autoregressive terms. The out-of-sample window is from May 2010 to December 2021. All variables were seasonally adjusted to achieve stationarity.

3 Methodology & Models

3.1 Benchmark Models

3.1.1 Random Walk & Autoregressive(AR)

The first benchmark model is the Random Walk model. The second benchmark model is the Autoregressive (AR) Model. These models are applied to forecast unemployment rates over various future horizons of 1, 3, 6, and 12 months.

3.2 Shrinkage

In this study, since it is high dimensional forecasting, we implement regression models with penalty term, including LASSO, Post-LASSO, and Elastic Net. All of them use a fixed-length rolling window framework. Each modeling window contains 148 observations, while the prediction window comprises 140 observations. These models are applied to forecast unemployment rates over various future horizons of 1, 3, 6, and 12 months.

[3] Specifically, we delete variables 'EMP', 'EMP_PART', 'EMP_TEMP', 'UNEMP_DURA_6mth', 'UNEMP_DURA_6-12mth', 'UNEMP_DURA_12mth+', 'UNEMP_DURA_24mth+', 'EMP_ACT' and 'EMP_ACT_RATE'

3.2.1 LASSO

LASSO introduces a l_1-norm penalty term to the regression loss function, promoting sparsity of obtained coefficient estimators in the resulting model (Tibshirani, 1996). Additionally, due to the sparsity property of obtained estimators, we can achieve variable selection implicitly when applying this approach. The objective function for LASSO is:

$$\min_{\beta} \left\{ \frac{1}{2n} ||y - X\beta||^2 + \lambda ||\beta||_1 \right\}, \tag{1}$$

where λ is the regularization parameter that controls the amount of shrinkage. The coefficient estimators produced by this approach are biased towards zero, which reduces variance and helps in model interpretability (Tibshirani, 1996).

3.2.2 Post-Lasso

Post-LASSO employs a two-step approach where variables selected by LASSO are re-estimated using ordinary least squares regression. This approach is aimed to reduce the bias while maintaining model simplicity (Belloni et al., 2012). The objective function for post-LASSO after variable selection is:

$$\min_{\beta \in S} \|y - X_S \beta\|^2, \tag{2}$$

where S represents the set of variables chosen by the LASSO method.

3.2.3 Elastic Net

Elastic Net is a hybrid approach that blends LASSO and Ridge penalties, useful for dealing with highly correlated data. The Elastic Net regression is described as:

$$\min_{\beta} \left\{ \frac{1}{2n} \|y - X\beta\|^2 + \lambda(\alpha \|\beta\|_1 + \frac{1}{2}(1-\alpha)\|\beta\|_2^2) \right\}, \tag{3}$$

where λ and α are regularization parameters optimized via criteria such as BIC, AIC, and AICc to balance the contributions of LASSO and Ridge penalties.

The selection of the optimal λ (or α) parameters for each model is aided by BIC, AIC, and AICc criteria, optimizing the trade-off between under-fitting and over-fitting issue.

3.3 Bagging

Bagging, an acronym for Bootstrap Aggregating (Breiman (1996)), is an ensemble learning technique that improves accuracy and reduces variance by voting within various models. This method is particularly effective in reducing the variance of complicated models, making them less prone to data randomness.

In this paper, we main implement two approaches, fixed-window and rolling-window, for bagging model regression. The fix-window method involves the deployment of random forest regression approaches, tailored to handle time series data by incorporating lagged variables. This method utilizes a predefined static window of data to both train and validate the models, contrasting with rolling window approaches where the window of data moves over time. The rolling-window approach employs a random forest model that incorporates both past observations and principal components as new predictors to forecast future values. The process is carried out over a predefined number of previous observations where, for each iteration, a new random forest model is trained on a rolling window of the data, gradually shifting forward in time. This approach is advantageous for capturing complex nonlinear relationships and interactions among variables in the time series data.

3.4 Random Forest

The Random Forest (RF) model was introduced by Breiman (2001) as a strategy to lower the variance observed in regression trees through a method known as bootstrap aggregation, or bagging. This involves combining multiple regression trees that are independently constructed using random subsets of data.

In this paper, we also utilize the fix-window and rolling-window methods for time series forecasting. For the RF model, we specifically adjust *'mtry'* parameter, thereby reducing the number of features considered at each split. This typically helps in reducing the model's variance and can potentially improve generalization by focusing on more relevant features.

4 Results

4.1 Basic Results

In the subsequent section, we delve into the comparative analysis of the outcomes derived from various predictive models.

To facilitate this comparison, we employ the Root Mean Square Error (RMSE) as a metric of predictive performance:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\text{truth}_i - \text{pred}_i)^2} \tag{4}$$

The lower the RMSE, the closer the fit is to the observed data, and hence the better the model's predictive accuracy.

Table 1: RMSE Comparison of Forecasting Methods Across 1, 3, 6, and 12 Month

Methods		1m	3m	6m	12m
Benchmark	Random Walk	0.1189	0.1276	0.1354	0.1556
	AR(4)	0.0915	0.1004	0.1017	0.1083
	AR(BIC)	0.0911	0.1003	0.1036	0.1055
Sentiments	LASSO(AIC)	0.103	0.102	0.110	0.105
	LASSO(BIC)	0.100	0.101	0.103	0.103
	LASSO(AICc)	0.102	0.102	0.109	0.104
	POST-LASSO(BIC)	0.097	0.103	0.108	0.106
	ELNET(AIC)	0.103	0.102	0.107	0.105
	ELNET(BIC)	0.104	0.100	0.103	0.103
	ELNET(AICc)	0.103	0.102	0.108	0.105
	Bagging(fixed)	0.116			
	Random Forest(fixed)	0.111			
	Bagging(rolling)	0.100	0.101	0.107	0.108
	Random Forest(rolling)	0.099	0.100	0.106	0.108
Macro	LASSO(AIC)	0.491	0.669	0.751	0.603
	LASSO(BIC)	0.104	0.103	0.294	0.268
	LASSO(AICc)	0.148	0.292	0.560	0.574
	POST-LASSO(BIC)	0.120	0.129	1.322	1.977
	ELNET(AIC)	0.520	0.671	0.757	0.498
	ELNET(BIC)	0.105	0.101	0.106	0.108
	ELNET(AICc)	0.240	0.584	0.707	0.390
	Bagging(fixed)	0.110			
	Random Forest(fixed)	0.105			
	Bagging(rolling)	0.0924	0.0953	0.103	0.106
	Random Forest(rolling)	0.0932	0.0941	0.102	0.105
All (Sentiment+Macro)	LASSO(AIC)	0.457	0.621	0.731	0.532
	LASSO(BIC)	0.104	0.102	0.236	0.343
	LASSO(AICc)	0.169	0.289	0.542	0.505
	POST-LASSO(BIC)	0.106	0.113	1.511	1.440
	ELNET(AIC)	0.479	0.619	0.703	0.458
	ELNET(BIC)	0.104	0.100	0.106	0.110
	ELNET(AICc)	0.396	0.589	0.642	0.374
	Bagging(fixed)	0.110			
	Random Forest(fixed)	0.107			
	Bagging(rolling)	0.093	0.094	0.104	0.106
	Random Forest(rolling)	0.092	0.093	0.102	0.106

4.1.1 Benchmark

Consistent Performance of AR Models: Stable RMSE Across Horizons
The Autoregressive models, namely AR(4) and AR(BIC), maintain relatively consistent RMSE values across different forecasting periods. The AR(BIC) model, for example, records RMSEs of 0.0911, 0.1003, 0.1036, and 0.1055 for the 1-month, 3-month, 6-month, and 12-month forecasts, respectively. This consistency suggests the model's adaptability and reliability over various temporal spans, providing stable forecasting performance.
Superiority of Short-Term Forecasts Over Long-Term Predictions
Combining the quantitative data from the RMSE table with the visual trends observed in the 1-step and 12-step benchmark forecast graphs, a clear pattern emerges in the predictive accuracy of the models across different forecasting horizons. Short-term 1-month predictions exhibit closer adherence to actual unemployment rates, evidenced by lower RMSE values of 0.0911 for AR(BIC) and 0.0915 for AR(4) models compared to 0.1189 for the Random Walk model. Figure 1 shows that AR model predictions closely following actual unemployment rates over time.

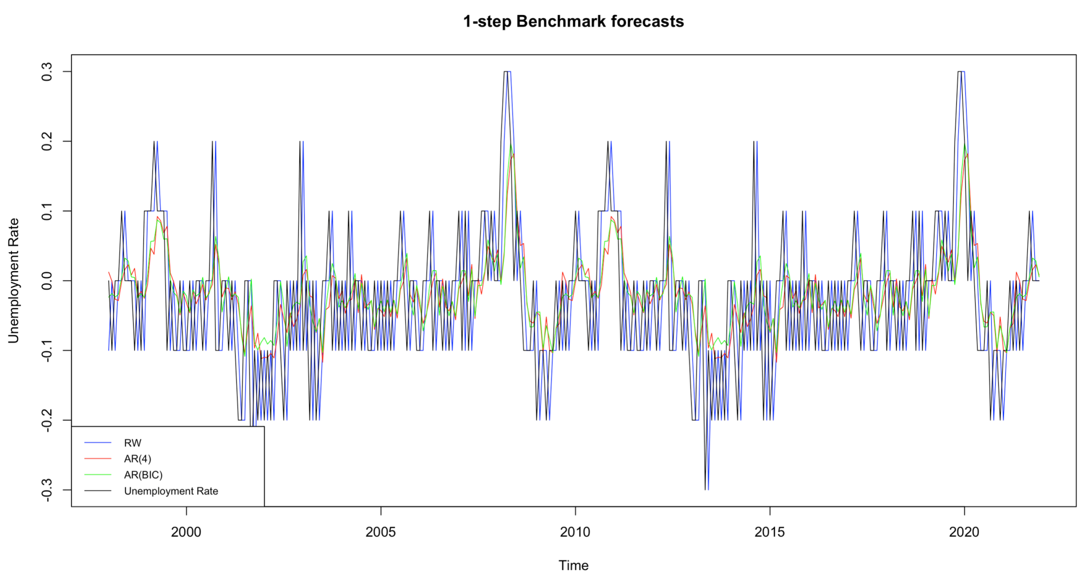

Figure 1: 1-step Benchmark Forecasts

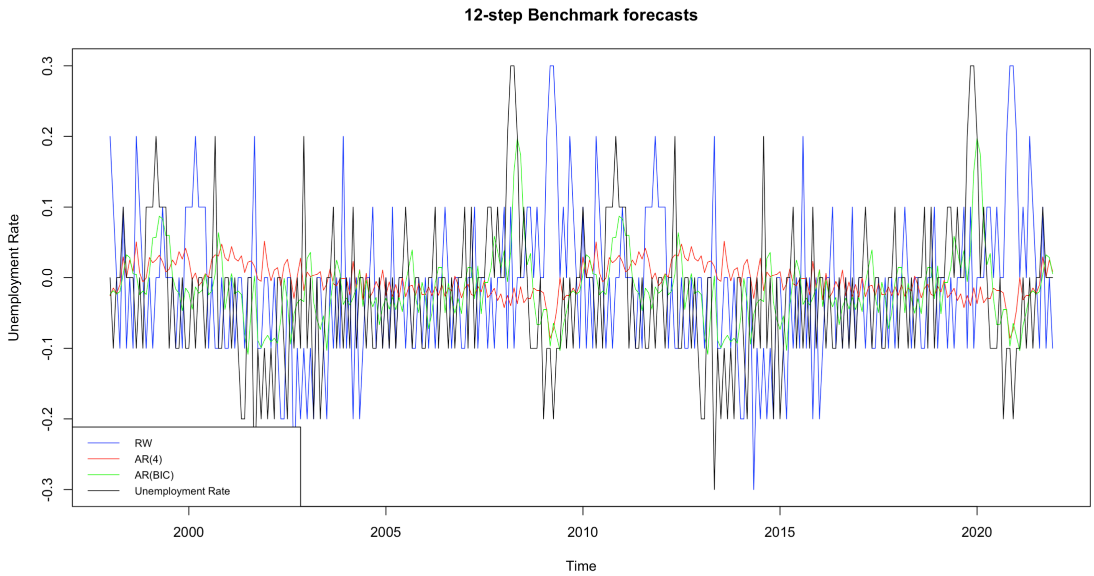

Figure 2: 12-step Benchmark Forecasts

Conversely, as the forecasting horizon extends to 12 months, all models exhibit an expected increase in RMSE values due to the inherent challenges associated with long-term economic forecasting. However, it is noteworthy that the AR models maintain a marked improvement over the Random Walk model, which suffers a notable rise in RMSE to 0.1556. In contrast, the AR(BIC) model shows considerable resilience with an RMSE of 0.1055, echoing the same advantage in the graphical trend **where the AR predictions appear less volatile and more aligned with the actual data.** The ability of AR models to integrate historical data trends and economic cycles allows them to provide more reliable forecasts.

4.1.2 Sentiments Variables

Consistent Performance of Sentiment Indices in Long-Term Forecasting. The analysis of sentiment variables in the context of predicting the UK's unemployment rates reveals a commendable level of accuracy, particularly in long-term forecasts. The dataset comprising the six sentiment indices demonstrates a strong predictive capability when utilized independently from other macroeconomic indicators.

The tight clustering of RMSE values across various methodologies, such as LASSO(AIC) with RMSEs from 0.102 to 0.104 and Elastic Net models with RMSEs ranging from 0.103 to 0.105 for all forecast horizons, suggests a small variance in predictive outcomes, indicating that different statistical learning techniques have similar efficiencies in exploiting these sentiment indicators.

These findings demonstrate that even relatively simple models can leverage sentiment indices to provide reliable forecasts. Therefore, these six sentiment variables may serve as potent tools for assessing economic trends in future macroeconomic modeling and policymaking processes.

4.1.3 Macroeconomic Variables

Predictive Performance Variance

A critical evaluation of macroeconomic models reveals stark performance disparities. Ensemble methods like bagging and Random Forests showcase enhanced effectiveness, especially for long-term forecasts. For instance, the RMSE for Random Forest (rolling) is notably lower at 0.092 at the 1m forecast, compared to more traditional models such as LASSO with an RMSE of 0.491 at the same horizon.

Comparative Analysis with Sentiment Indicators

When juxtaposed with sentiment-driven models, macroeconomic models show only a marginally better performance. This is evident when we consider the Random Forest (rolling) with an RMSE of 0.092 for the 1m horizon compared to 0.099 for the same model using sentiment indicators. This slender edge highlights the potency of sentiment-driven models, which, despite their simplicity and reliance on only six variables, demonstrate a commendable predictive validity.

Superiority of the Rolling Window Approach

The rolling window versions consistently outperform their fixed window counterparts.

For instance, the Random Forest (rolling) at a 6m forecast horizon achieves an RMSE of 0.102, in contrast to the fixed window approach's RMSE of 0.105. The rolling window's inherent flexibility, which allows it to continuously assimilate the latest data, confers upon it an advantage that is indispensable in the unpredictable domain of macroeconomic forecasting.

4.1.4 All Variables

In the examination of the predictive performance of combined sentiment and macroeconomic indicators ("All") against macroeconomic indicators alone, the data portrays a nuanced enhancement in forecasting capability across various models, although this improvement is modest. For example, the Random Forest (rolling) model's RMSE decreases from 0.0932 for macro alone to 0.092 with the combined indicators.

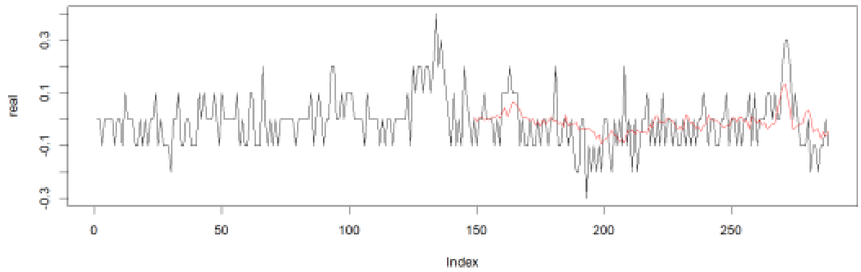

Figure 3: 1-step Senti Random forest-rolling window(RMSE=0.099)

Interpretation

The results indicate that while the inclusion of sentiment indicators does contribute to a modest improvement in forecasting unemployment rates, the extent of this enhancement is not profound. This suggests that while sentiment indicators do hold predictive value, especially in capturing the immediate perceptions that might precede economic shifts, they do not drastically alter the overall predictive

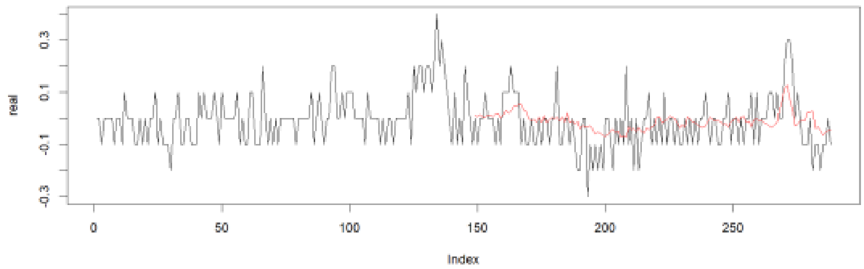

Figure 4: 1-step Macro Random forest-rolling window(RMSE=0.0932)

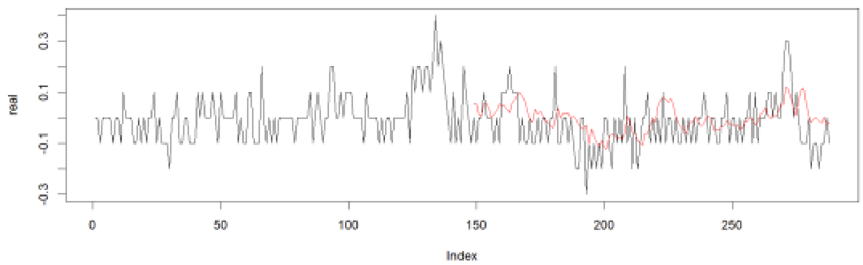

Figure 5: 1-step All Random forest-rolling window(RMSE=0.092)

landscape when combined with macroeconomic variables.

4.2 DM Tests

From the basic results we can find that benchmark models especially Autoregressive models(AR) outperform all the other models with the minimum RMSE. In addition, among non-benchmark models, Random Forest with rolling window and Elastic Net which λ is chosen by BIC produce more precise forecasts. Therefore, we want to explore two questions: 1) Does random forest have a different predictive ability from the AR benchmark? 2) Does Random Forest have equal predictive ability to Elastic Net(BIC)?

To test whether the forecasts from distinct models are different, we consider using

Diebold and Mariano (2002) Equal Predictive Ability Test[4] (DM Test). Since we can get more precise forecasts by combining sentiment variables and macro indicators, here we use the combined data to do DM Test.

4.2.1 AR(BIC) Benchmark Versus Random Forest

First to check DM stationarity Assumption. Figure 6 shows the loss differentials of AR-RF. The figure illustrates that the loss differential exhibits stationarity over the forecasting horizons of 1, 3, 6, and 12 months satisfying the assumption required for conducting the Diebold-Mariano (DM) test.

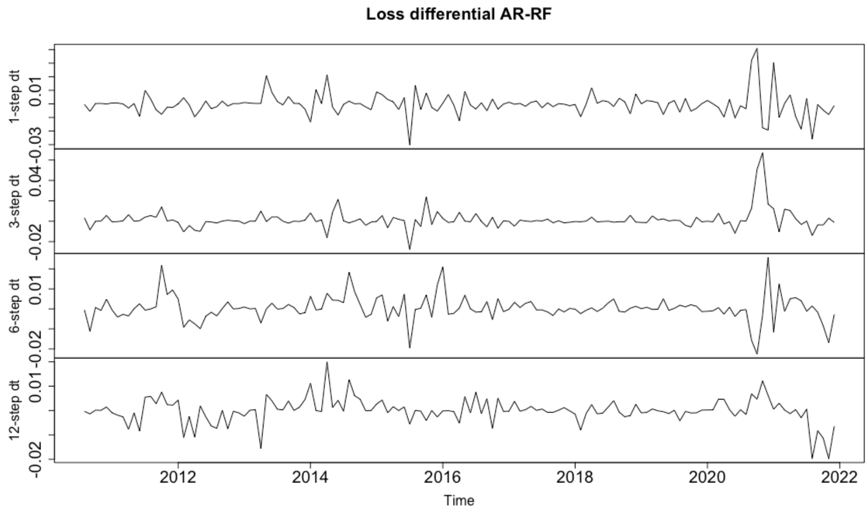

Figure 6: Loss Differential AR-RF

DM Tests are done through regressions on a constant with HAC standard errors. We use the Newey-West estimator with 5 lags according to the cube root P rule-of-thumb (140(1/3) = 5.192494 - choose 5). Table 2 shows the results of the comparison between AR(BIC) benchmark and Random Forest. We can conclude that the Random forest has equal predictive ability to the AR benchmark at all horizons since the T-statistics are

[4] Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 20(1), 134-144.

all below 1.96(5%).

	AR-RF	ELNET-RF
1-Step	-0.318	2.298
3-Step	1.156	1.978
6-Step	0.481	0.380
12-Step	-0.126	-0.372

Table 2: DM Test

4.2.2 Elastic Net(BIC) Versus Random Forest

Figure 7 shows the loss differentials of ELNET-RF. The figure indicates that the loss differential exhibits stationarity over all the forecasting horizons, satisfying DM Test assumption. We also use the Newey-West estimator with 5 lags.

Table 2 presents the results of the comparison between Elastic Net (BIC) and Random Forest. Random Forest shows a better predictive ability than Elastic Net at 1-month and 3-month forecasting horizons since the T-statistics are larger than 1.96. But Elastic Net performs equal to Random Forest in longer-term forecasting.

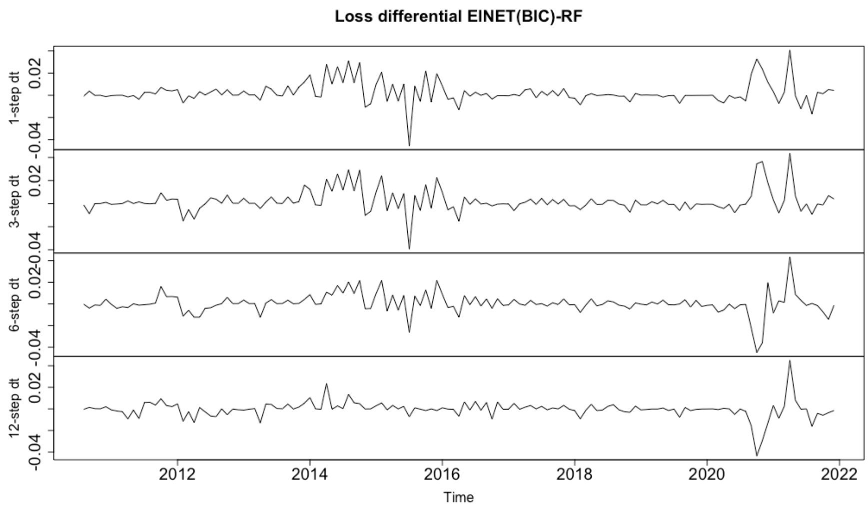

Figure 7: Loss Differential ElNET-RF

5 Conclusion

This paper focuses on the effectiveness of sentiment indicators constructed based on news articles in unemployment forecasting. We directly adopt the six sentiment indicators extracted and processed by Barbaglia et al. (2021) from major UK newspaper articles, combined with 99 macroeconomic indicators officially released in the UK, and advanced machine learning methods such as LASSO and Random Forests (RF), and compare them with benchmark models. The findings show that real-time sentiment indicators have independent value in forecasting UK unemployment alone and show incremental effects on the predictive power of sentiment indicators on macroeconomic indicators in most models.

Regarding model performance, AR outperforms other machine learning models as a benchmark model across multiple time dimensions. Indeed, it is not uncommon for traditional statistical models to outperform machine learning models in some cases when dealing with time series data. This may be attributed to the strong time dependence and predominantly linear relationship of time series data, features that may be more effectively captured by simple auto-regressive models. In addition, machine learning models are prone to overfitting in small samples and perform poorly on unknown data.

For different forecasting time horizons of the unemployment rate, short-term forecasts generally outperform long-term forecasts. This finding suggests that near-term sentiment indicators or macroeconomic indicators are more effective in reflecting current economic trends and cyclical changes. As the forecast horizon increases, the model must rely on more distant data, which increases future uncertainty and sensitivity to changes in the market or environment, leading to a decrease in forecast accuracy.

In addition, we have studied comparatively fixed and sliding windows. We can notice that for either model, the sliding window model is more advantageous in time series forecasting. By continuously including the latest data to capture recent trends and patterns in the time series, sliding windows enhance the model's adaptability to structural changes and data dynamics.

Future research will focus on optimizing model structures and parameter settings, particularly on how to enhance model performance across different time dimensions through ensemble methods and new machine learning technologies to perform better than benchmarks. Additionally, with the advancements in machine learning and big data technologies, developing advanced forecasting systems that can adapt to rapidly changing environments and real-time data flows will be crucial. We also plan to explore the forecasting of other macroeconomic indicators and how these advanced models can aid in devising more precise and effective economic policies and market analysis strategies.

References

Barbaglia, L., Consoli, S., and Manzan, S. (2021). Forecasting gdp in europe with textual data. *Journal of Applied Econometrics*.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24:123–140.

Brunner, K., Cukierman, A., and Meltzer, A. H. (1980). Stagflation, persistent unemployment and the permanence of economic shocks. *Journal of Monetary Economics*, 6(4):467–492.

Consoli, S., Barbaglia, L., and Manzan, S. (2022). Fine-grained, aspect-based sentiment analysis on economic and financial lexicon. *Knowledge-Based Systems*, 247:108781.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1):134–144.

Floros, C. (2005). Forecasting the uk unemployment rate: model comparisons. *International Journal of Applied Econometrics and Quantitative Studies*, 2(4):57–72.

Grubb, D., Jackman, R., and Layard, R. (1982). Causes of the current stagflation. *The Review of Economic Studies*, 49(5):707–730.

Jelena, M., Ivana, I., and Zorana, K. (2017). Modeling the unemployment rate at the eu level by using box-jenkins methodology. *KnE Social Sciences*, pages 1–13.

Katris, C. (2020). Prediction of unemployment rates with time series and machine learning techniques. *Computational Economics*, 55(2):673–706.

Kreiner, A. and Duca, J. (2020). Can machine learning on economic data better forecast the unemployment rate? *Applied Economics Letters*, 27(17):1434–1437.

Smalter Hall, A. and Cook, T. R. (2017). Macroeconomic indicator forecasting with deep neural networks. *Federal Reserve Bank of Kansas City Working Paper*, (17-11).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.