CSC696 Project Final Report

Sunghoon Kim
University of Arizona
poizzonkorea@email.arizona.edu

Abstract

For companies that trade with various countries, fluctuations in risk in those countries are of great concern. A company's financial loss can be reduced if the increase of risk can be predicted in advance and accurately. To quickly recognize changes in risk, information about countries is obtained through economic indicators and news, and a more accurate prediction is realized by implementing a machine learning model. In addition, since it can be related to decisive judgments with significant losses, explainability was implemented to provide a rationale for prediction to various stakeholders. As a result, countries' classification changes were partially predicted, and limited explainability was also provided.

1 Introduction

Forecasting financial time series aims to anticipate predictable patterns that will bring investors advantage in trading opportunities (Vukovic et al., 2022). In this project, I am going to implement a supervised machine learning model that predicts the classification of a country risk. When a situation such as a national default or social unrest occurs in a specific country, the economic damage to companies doing business with the country will be considerable. The purpose of this project is to minimize the economic damage by predicting these country risk in advance through this model. Fortunately, national defaults and social unrests do not occur suddenly, and predictive events occur. Our model analyzes the news including predictive events in the countries and uses it as dataset. In order for the results of our model to lead to the decision maker's judgment, our model must be able to explain to the users to understand the result of country risk. Namely, research for explainability of our model will also be conducted.

1.1 OECD Country risk classification

The aims of this project is to predict Organisation for Economic Co-operation and Development (OECD) Country risk classification. OECD has the country risk classification methodology that a group of country risk experts from Export Credit Agencies(ECA) meets several times a year to update the list of country risk classifications (OECD, 2022). Since the OECD's methodology uses a mixture of qualitative and quantitative models, it is difficult to predict the risk classification by traditional approaches such as regressions or decision trees. This classification is used by ECA, so it becomes the major feature to evaluate importer's country risk, financial supporting export companies. If the OECD classification is downgraded, exporting companies may be provided with unfavorable conditions or may not be provided with the conditions themselves when they apply to ECA for financing necessary for export.

1.2 GDELT

When estimating country risk, I wanted to refer the news with the geopolitical crisis or natural disaster. However, in order to designate and analyze specific media companies, it was difficult to determine whether the media companies would cover all the events of all countries, and which news has credibility from different viewpoints among media companies. So I decided to use an open source project called Global Data on Events, Location, and Tone (GDELTProject, 2022) as a dataset of our model. The GDELT Project monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, themes, sources, emotions, counts, quotes, images and events driving our global society every second of every day, creating a free open platform for computing on the entire world (Leetaru and Schrodt, 2013).

1.3 Approach of this project

This project will derive a machine learning model to predict OECD country classification by referring to the refined news dataset provided by GDELT and economic indicators provided by the International Monetary Fund (IMF, 2022). This model is also configured with Explainable Artificial Intelligence(XAI), to provide the user with a rationale for understanding the result of this model. As trials in the related work, I provide the explainability by using post-hoc explainability models such as LIME and ALE. In this project, it will be the first case of applying 1) XAI to country risk prediction and 2) country risk prediction for all OECD countries.

2 Related Work

The related work was analyzed by dividing it into 4 areas. The contents of the OECD country risk are described in 2.1, and the machine learning model predictions for the country risk indicators are described in 2.2. Also, the project using GDELT is described in 2.3, and the XAI work for credit risk is described in 2.4.

2.1 OECD Country Risk

As shown in the figure below, the OECD publishes country risk classification reports for 201 countries at least twice a year. It has been published 93 times from 1999 to March 2022 and the most recent publish was made on March 11, 2022.

Country Risk Classifications
of the Participants to the Arrangement on Officially Supported Export Credits
Valid as of: **11 March 2022**

			Classification		
nb	Country Code ISO Alpha 3	Country Name [1]	Previous	Current Prevailing	Notes
1	AFG	Afghanistan	7	7	
2	ALB	Albania	5	5	
3	DZA	Algeria	5	5	
4	AND	Andorra	-	-	(9)
5	AGO	Angola	6	6	
6	ATG	Antigua and Barbuda	7	7	(8)
7	ARG	Argentina	7	7	
8	ARM	Armenia	6	6	
9	ABW	Aruba	6	6	
10	AUS	Australia	-	-	(6)
11	AUT	Austria	-	-	(6) (7)
12	AZE	Azerbaijan	4	4	
13	BHS	Bahamas	4	4	
14	BHR	Bahrain	6	6	
15	BGD	Bangladesh	5	5	
16	BRB	Barbados	-	-	(5)
17	BLR	Belarus	6	7	
18	BEL	Belgium	-	-	(6) (7)
19	BLZ	Belize	-	-	(5)

Figure 1: Country Risk Classifications

Since both the GDELT dataset and the OECD country risk use the ISO country code, it is possible to link the two datasets. The national ratings are distributed from 1st to 7th grade, where 1st grade is a high-reliability country and 7th grade is a low-reliability country. In the most recent publish, two countries had their credit ratings downgraded, with Russia and Belarus both downgrading to 7th. Some countries are not evaluated for reasons such as “High Income OECD or Euro Area Country not reviewed or classified.” or “Currently not reviewed or classified.”, so I will exclude them from the prediction in this project as well.

2.2 ML Model for country risk indicators

"Multivariate CDS risk premium prediction with SOTA RNNs on MI[N]T countries" (Kutuk and Barokas, 2021) predicts CDS risk premiums of Mexico, Indonesia and Turkey by applying state-of-the-art forecasters in deep learning recurrent neural networks architectures. A credit default swap (CDS) is a financial swap agreement that the seller of the CDS will compensate the buyer in the event of a debt default (by the debtor) or other credit event (the Wharton School of the University of Pennsylvania., 2022). In short, an increase in the CDS premium indicates an increase in default risk, and a decrease in the CDS premium indicates a decrease in default risk. The architectures used in this article, RNN (ELMAN), NARX, GRU, and LSTM, will be applied as candidates to be the best learning architecture in this project as well. In order to check the learning performance, like this article, I consider a criterion of the performance such as MSE, MAE, and R^2(=Root Mean Square Error). Also, in this article, each country has different optimized machine learning architectures, and as in this project, it is expected that the dominant architecture will be easily distinguished as the number of country set increases.

"Are CDS spreads predictable during the Covid-19 pandemic? Forecasting based on SVM, GMDH, LSTM and Markov switching autoregression" (Vukovic et al., 2022) investigates the forecasting performance for CDS spreads by Support Vector Machines (SVM), Group Method of Data Handling (GMDH), Long Short-Term Memory (LSTM) and Markov switching autoregression (MSA) for daily CDS spreads of the 513 leading US companies, in the period 2009–2020. Although this article targeted companies for credit ratings instead of countries, it was found that even under special circumstances such as COVID-19, the performance of the rating prediction model was not significantly affected. This implies that if the resulting model of

this project is implemented well, it can be applied universally.

"Examination of Country Risk Determinants Using Artificial Neural Networks: The Case of Turkey" (Topak and Muzir, 2011) constructs a functional model of forecasting country risk changes in Turkey with the help of artificial neural networks. This article predicts the OECD country classification of Turkey, and the correct classification rate has increased up to 98.5% for the terms with improved conditions. This made it possible to realize that the predictive model for the OECD country risk classification is feasible. In addition, the influence of each economic indicator on the model is presented as shown in the table below, giving hints in the selection of economic indicators in this project.

INCREASE IN COUNTRY RISK	DECREASE IN COUNTRY RISK
Wholesale Price Index (WPI), Market Interest Rate, Open Market Operations (OMO)	GDP, Index of Employment, Consumer Confidence Index, Real FX Rate Index, External Debt, Domestic Assets, Issue Volume (Money Emission) Reserves, Wage Index

* Risk effects are determined with respect to ascending trend in the corresponding variables.

Figure 2: Effects of Economic Indicators in Risk Measurement of Model Variables

2.3 How to handle GDELT

The GDELT Project (Leetaru and Schrodt, 2013) provides a realtime network diagram and database of global human society for open research which monitors the world's events. GDELT dataset has 2 versions. GDELT 2.0 has 3 more fields and updates every 15 minutes, while GDELT 1.0 updates every single day, but GDELT 2.0 only manages events from February, 2015. So GDELT 2.0 doesn't fit the time-series in this project. Therefore, this project use GDELT 1.0 datasets.

GDELT 1.0 has 58 fields. In this project, I use the following five fields from a record: SQLDATE, EventBaseCode, AvgTone, NumMentions, Actor_CountryCode. SQLDATE is the date the event took place in YYYYMMDD format. EventBaseCode denotes level two leaf root node. For example, code 1452 (engaging in violent protest for policy change) has a EventBaseCode of 145 (Protest violently, riot, not specified below). Examples of EventBaseCode and its description is written in the table below. AvgTone is a numeric score from -100 to +100, this can be used as a method of filtering the "context" of events as a subtle measure of the importance of an event and as a proxy for the "impact" of that event. NumMentions is the total number of mentions of this event across all source documents. Actor_CountryCode is the country of the actors, which is a 3-character ISO country code for the location.

Table 1: Example of EventBaseCode and description

EventBaseCode	Description
101	Demand material cooperation
111	Criticize or denounce
121	Reject material cooperation
131	Threaten non-force
141	Demonstrate or rally
151	Increase police alert status
161	Reduce or break diplomatic relations

"Predicting Social Unrest Events with Hidden Markov Models Using GDELT" (Qiao et al., 2017) builds a Hidden Markov Models (HMMs) based framework to predict indicators associated with country instability using autocoded events dataset GDELT. Through this article, this project got hints on which fields to extract from GDELT. "Estimating countries' peace index through the lens of the world news as monitored by GDELT" (Voukelatou et al., 2020) predicts countries' peace index by Elastic Net, Decision Tree, and Random Forest, and the process of their dynamic training. This article describes in detail how to normalize the GDELT datasets to minimize differences by countries or timeslots.

2.4 XAI in credit risk

"Interpretable Machine Learning" (Molnar, 2022) introduced concepts and relevant libraries of various XAI papers. In this project, Accumulated Local Effects (ALE) (Apley and Zhu, 2020), which uses to explain a model itself, and Local interpretable model-agnostic explanations (LIME) (Mishra et al., 2017), which uses to explain a single prediction from the input, would take into account for the explainability of our model. "Towards Explainable Deep Learning for Credit Lending: A Case Study" (Modarres et al., 2018) explored the process of explaining credit lending decisions made by a neural network using three different attribution methods: LIME, DeepLIFT, and Integrated Gradients. "ENABLING MACHINE LEARNING ALGORITHMS FOR CREDIT SCORING" (Biecek et al., 2021) showed how to take credit scoring analytics in to the next level, namely they present comparison of various predictive models (logistic regression, logistic regression with weight of evidence transformations and modern artificial intelligence algorithms) and show that advanced tree based models give best results in prediction of client default. Both

documents apply the post-hoc explainability model to credit risk machine learning model, making a business case viable for now "not-so-black-box models". this project will also consider applying such a surrogate model as the primary goal, and expansion of the direct explainability model will also be considered. In particular, in the case of Experiment 1 of (Modarres et al., 2018), it seems that the idea can be applied to this project because an evaluation method that excludes humans' judgement is implemented in evaluating XAI.

3 Approach

This project implements a machine learning model predicting OECD country risk classifications with GDELT and economic indicators datasets. Since 97.39% of classifications are same as previous one, this project designates the model that predicts the replica of previous classification as the baseline model. In this project, there are four approaches, GDELT, Economic Indicators, Design of Model, and Explainability of this model, to surpassing baseline's accuracy with reasonable execution performance. The process of finding the optimal hyperparameter among the hyperparameters defined in the following subsections is explained in the Appendix A and B.

3.1 Trials for GDELT

The original csv file size of GDELT 1.0 from 1998 to March 2022, which corresponds to the classification period of the OECD, is more than 220GB. Direct access to these files as an input adversely affects the performance of the model, so summary files were created. The summary file is grouped by classification period and countries, and consists of sum of AvgTone and NumMentions for each EventBaseCode as Figure 3. When writing the summary file, consider hyperparameters as the following subsections to find a optimal way to utilize GDELT dataset.

Evaluation Period	Country code	Count_ 010	Sum_AvgT one_010	Sum_Numof Mentions_010	...	Count_ 203	Sum_Avg Tone_203	Sum_Numof Mentions_203
000	USA	3685	20853.81	15803	...	4	27.37	23
001	USA	630	3526.37	2837	...	1	5.55	2
...	...	...	...	...	...	...	...	...

Figure 3: Example for summary of GDELT

3.1.1 Hyperparameters for GDELT

(Gap in evaluation date vs reference date) It takes a certain amount of time to aggregate members' opinions for the OECD's country classification. Namely, there may be a gap between the reference date of the data required for classification and the date on which the classification is executed after collecting opinions. Assume that the gap will be one of `[0, 5, 10, 15, 20, or 30]` days.
(Extreme vs all news) It starts with the assumption that news with an significant impact may affect the change in OECD's country classification. Whether `only news's AvgTone <= -10 or >= 10` will be used for training or `all news` will be used for training dataset as a hyperparameter.
(Subject vs subject+object) Whether to use only the country corresponding to the `subject` of the news or to use both the countries corresponding to the `subject` and the `object` of the news as the training dataset is designated as a hyperparameter.

3.1.2 Choose EventBaseCodes in GDELT

Not all of the 190 EventBaseCodes provided by GDELT will affect the change of country credit ratings. By analyzing the GDELT corresponding to the only period in which the country credit rating change occurred, the ranking of EventBaseCodes with top or bottom mean AvgTone were derived. As a hyperparameter, only the GDELT in the top `[5, 10]` eventbasecodes are used as dataset.

3.1.3 Normalize scores in GDELT

The distribution of AvgTone and NumMentions values will be different in large countries such as the United States and China and small countries such as San Marino and Gambia because the frequency of news events is different. In order to correct for differences in the distribution between countries, AvgTone and NumMentions values were normalized as follows with reference to paper 'Predicting Social Unrest Events with Hidden Markov Models Using GDELT' (Qiao et al., 2017). The formula below 1 is a case of calculating a normalized score based on AvgTone for the evaluation period `t` of country `c`.

$$Score_{c,t} = \frac{AvgTone_{c,t}}{1/10 * \Sigma_{j=t-0}^{t-9} AvgTone_{c,j}} \quad (1)$$

3.2 Trials for Economic Indicators

Referring to paper "Examination of country risk determinants using artificial neural networks"(Topak and Muzir, 2011), among the economic indicators that can affect the OECD country's classification

Name	Freq
Total balance of Payment	Qt
Residential Real Estate Prices, Percent	Mn
Reserve and Foreign Currency Assets	Mn
Net lending/borrowing rate	Yr
Net lending/borrowing rate in public	Yr
Net operating balance of government	Yr
Total Capital Account, Debt	Mn
Economic Activity, Industry Production	Mn
Economic Activity, Retail Sales	Mn
Nominal Effective Exchange Rate	Mn
Prices, Consumer Price Index	Mn
Interest Rates, Money Market	Mn
Labor Markets, Employment	Mn
Gross Domestic Product	Mn
Assets (w/ Fund Record)	Mn
Liabilities (w/ Fund Record)	Mn
Financial Market Prices, Equities	Mn
Imports of Goods and Services	Mn
Exports of Goods and Services	Mn

Table 2: Candidate of Economic Indicators to affect OECD country grades. Freq : Frequency of publication, Qt : Quarterly, Mn : Monthly, Yr : Yearly

and are provided by the IMF(IMF, 2022), 19 indicators were selected as the table 2. According to the statistic system for each country, the case where a value was not published for the relevant indicator was regarded as 0. There are 2 steps to calculate input for out model. `1)` Normalize the values for each indicator and evaluation period as mean = 0 and standard deviation = 1. `2)` Mean of indicators in each country and evaluation period is our input of the model. The process of finding a combination that induces optimal accuracy among these indicators is described in the Appendix B.

3.3 Trials for design of model

Our model is a multiclass classification supervised model because it predicts the OECD country classification (1-7). Using the OECD country classifications, GDELTs and economic indicators mentioned above, 70% of the data is training set and 30% of the data is test set. In order to explain the design process of the model, it will be divided by 3 subsections, input data, output label and architecture.

3.3.1 Input data

The shape of our model's input data is time series. Namely, it's 3D array, which has 3 axes, `countries, evaluation period, and dataset for each country and period`. Dataset for each country and period doesn't have country code or evaluation date, since our model aims for independent model about country and evaluation date. Also, there is a hyperparameter about length of evaluation period as `[60, 36, 12, 6, 3]`.

3.3.2 Output label

The initial output label was to predict an OECD classification, from 1 to 7. However, as the accuracy decreased due to the tendency to follow the t-2 grade without recognizing the grade change at t-1, other output label was considered. Therefore, this model's final output label is that predicts one of `upgrade(1), no change(0), and downgrade(-1)` in the classification.

3.3.3 Architecture

The prediction performance is compared among Recurrent Neural Network(RNN), Long Short-Term Memory(LSTM), and Gated Recrurent Unit(GRU), which are representative machine learning architectures that process time series data. To check the performance of each architecture itself, I compose a model only with a single layer of above architecture and an activation function(softmax) layer.

3.4 Trials for explainability

To overcome the problem that is not suitable in regulated financial services, our model needs to provide details of our model or reasons of single prediction. In order to explain our model to users, I focus on explaining the correlation between input data and output label rather than explaining the architecture that requires technical background. Since our model has more than three types of input features, I decided to use ALE, which expresses the degree of influence for all input features, rather than PDP, which expresses the interaction of only two input features. Also, in order to explain the our model's result prediction, I use LIME to explain the value of each input feature and each input feature's contribution to the result of the output label.

4 Results/Discussion

4.1 OECD Classification Prediction Results

By applying the combination of the hyperparameters selected in Appendix A and the remaining hyperparameters, the dataset was shuffled 7 times

Gap period	Sub? Obj?	No of EventCode	Indicators	Score	Accuracy	MSE
10	**Subj+Obj**	**10**	**3**	**NumOfMentions**	**98.07%**	**0.0127**
20	**Subj**	**5**	**All Indicators**	**AvgTone**	**97.97%**	**0.0133**
30	Subj	5	1	NumOfMentions	97.96%	0.0134
15	Subj+Obj	10	2	AvgTone	97.93%	0.0135
15	Subj+Obj	5	2	NumOfMentions	97.91%	0.0136
			Baseline		97.58%	0.0590

Table 3: Top 5 model's hyperparameter and performance

to derive the Top5 model as shown in the table 3. The accuracy is between 97.91% and 98.07%. For the value of Indicator in Table 3, refer to the ID in Appendix B. Since the combination of hyperparameters is evenly distributed, but there does not seem to be an impressive advantage of a specific hyperparameter. Out of 160 hyperparameter combinations, 133 showed higher accuracy than baseline, 97.58%, confirming that our model could predict some changes in OECD country classifications.

4.2 Interpretation of the model

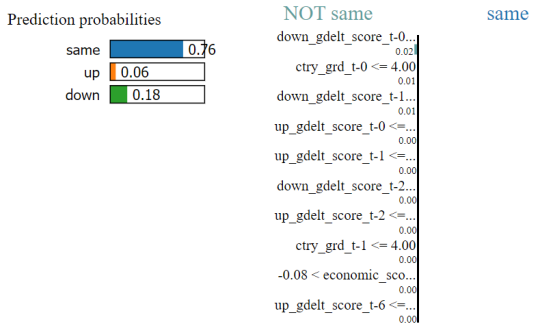

Figure 4: Probabilities and Contribution for each output label in LIME

Feature	Value
down_gdelt_score_t-0	-6.38
ctry_grd_t-0	4.00
down_gdelt_score_t-1	-6.39
up_gdelt_score_t-0	0.00
up_gdelt_score_t-1	0.00
down_gdelt_score_t-2	-6.76
up_gdelt_score_t-2	0.00
ctry_grd_t-1	4.00
economic_score_t-1	-0.03
up_gdelt_score_t-6	0.00

Figure 5: Input features and values in LIME

Above figures explain OECD classification prediction of our model for March 2022 in Russia by LIME. The left part of Figure 4 shows the predicted probability for each output label. The final prediction is that the credit rating is "same", but there is also an 18% probability of a credit rating downgrade. The right part of Figure 4 shows the degree of contribution to "same" for each input feature. Although the absolute value is rather small due to normalization, it can be seen that some contribution occurs even in not same. In LIME, the contribution of input features for each output label can be expressed as above. Figure 5 shows the list of input values applied to LIME.

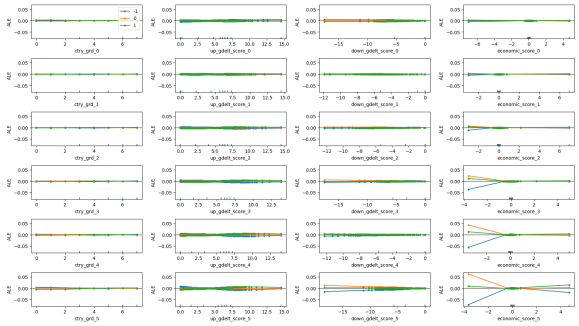

Figure 6: Effect of Input features in ALE

The figure 6 shows how much each input feature contributes to the prediction of our model through ALE. Since our final model uses 36 time series, it is too large for this paper to represent as a single image, so 6 time series ML model captured in the test phase was displayed. The lower part is the most recent time series and the upper part is the far past time series. It is inferred that the amplitude of the graph in the lower part is larger than that in the upper part, therefore the recent time series has more influence on the prediction of the model. Through such an image, targets to be improved for each input feature can be recognized.

5 Conclusions

The implementation process and results of "OECD Country Classification Prediction over XAI" have been described. This project contributed that the

prediction for OECD Country Classification could be derived more accurately with the help of GDELT and economic indicators. Although our model shows slightly higher accuracy than the baseline, I was able to identify improvement points such as low recall for better performance. Also, to improve the prediction performance, research on well-designed multi-layer models or refined methods of handling input features will be the future work. For explainability, future work such as improvement through feedback from users or implementation of built-in explainability rather than application of external packages can be considered.

References

Daniel W Apley and Jingyu Zhu. 2020. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1059–1086.

Przemysław Biecek, Marcin Chlebus, Janusz Gajda, Alicja Gosiewska, Anna Kozak, Dominik Ogonowski, Jakub Sztachelski, and Piotr Wojewnik. 2021. Enabling machine learning algorithms for credit scoring–explainable artificial intelligence (xai) methods for clear understanding complex predictive models. *arXiv preprint arXiv:2104.06735*.

GDELTProject. 2022. The gdelt story.

IMF. 2022. Imf data.

Yasin Kutuk and Lina Barokas. 2021. Multivariate cds risk premium prediction with sota rnns on mi [n] t countries. *Finance Research Letters*, page 102198.

Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pages 1–49. Citeseer.

Saumitra Mishra, Bob L Sturm, and Simon Dixon. 2017. Local interpretable model-agnostic explanations for music content analysis. In *ISMIR*, volume 53, pages 537–543.

Ceena Modarres, Mark Ibrahim, Melissa Louie, and John Paisley. 2018. Towards explainable deep learning for credit lending: A case study. *arXiv preprint arXiv:1811.06471*.

Christoph Molnar. 2022. *Interpretable Machine Learning*, 2 edition.

OECD. 2022. Oecd country risk classification. The Participants established a methodology for assessing country credit risk and classifying countries in connection with their agreement on minimum premium fees for official export credits.

Fengcai Qiao, Pei Li, Xin Zhang, Zhaoyun Ding, Jiajun Cheng, and Hui Wang. 2017. Predicting social unrest events with hidden markov models using gdelt. *Discrete Dynamics in Nature and Society*, 2017.

the Wharton School of the University of Pennsylvania. 2022. Cdos are back: Will they lead to another financial crisis?

Mehmet Sabri Topak and Erol Muzir. 2011. Examination of country risk determinants using artificial neural networks: The case of turkey. *International Research Journal of Finance and Economics*, 75(120).

Vasiliki Voukelatou, Luca Pappalardo, Ioanna Miliou, Lorenzo Gabrielli, and Fosca Giannotti. 2020. Estimating countries' peace index through the lens of the world news as monitored by gdelt. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 216–225. IEEE.

Darko B Vukovic, Kirill Romanyuk, Sergey Ivashchenko, and Elena M Grigorieva. 2022. Are cds spreads predictable during the covid-19 pandemic? forecasting based on svm, gmdh, lstm and markov switching autoregression. *Expert systems with applications*, 194:116553.

A Choice of GDELT's hyperparameter

A.1 Experiment Setup

Since it was undecided whether the score criterion of the final model should be AvgTone or NumOfMentions, both scores were used in the experiment to select the hyperparameter that has a common advantage. To reduce the random effect of the dataset, the OECD grade dataset is shuffled by 5 times. In order to reduce the preference due to hyperparameters other than test subject, other hyperparameters are randomly selected for each shuffled dataset by 5 times. I decided 4 hyperparameters, Gap period, Extreme news vs All news, History Length, and ML Architecture, by this experiment.

A.2 Result

In point of accuracy's view, it was rare to find significant differences between hyperparameters. Still, it is a waste of computing resources to use all the variables to select the final model, so I tried to reduce the variables. As a result of analyzing `Gap between evaluation date and reference date` first, the superiority of accuracy is revealed after 10 days as Table 4, so the two variables corresponding to before 10 days, `[0, 5]`, are removed.

In `Extreme news vs All news`, All news shows slightly better performance than

Gap period	AvgTone	NumMention
0	97.50	97.37
5	97.36	97.43
10	**97.55**	**97.34**
15	**97.57**	**97.57**
20	**97.52**	**97.51**
30	**97.58**	**97.50**

Table 4: Comparison Accuracy by Gap between evaluation date and reference date

News Type	AvgTone	NumMention
Extreme	97.49	97.42
All	**97.54**	**97.48**

Table 5: Comparison Accuracy by Extreme news vs All news

Extreme news in both score types as Table 5, `All news` is chosen.

In `History Length`, Length with 36 shows slightly better performance than other periods in both score types as Table 6, `36` is chosen.

In `Machine Learning Architecture`, LSTM shows slightly better performance than other architectures in both score types as Table 7, `LSTM` is chosen.

B Choice of economic indicators

Among the economic indicators in Table 2, this experiment was designed to select indicators having high correlation to OECD classification change as follows. Since the number of cases increases to test all combinations of 19 indicators, our optimal combination is assumed the combination with 10 indicators. To reduce the random effect of the dataset, the OECD grade dataset is shuffled by 5 times with normalized NumOfMentions as the score. In order to reduce the preference due to hyperparameters other than economic indicators, other hyperparameters are randomly selected for each shuffled dataset by 5 times. That is, the top three index combinations

History Length	AvgTone	NumMention
3	97.63	97.49
6	97.56	97.61
12	97.50	97.41
36	**97.67**	**97.68**
60	97.22	97.07

Table 6: Comparison Accuracy by History Length

Architecture Type	AvgTone	NumMention
RNN	97.51	97.44
LSTM	**97.52**	**97.46**
GRU	97.51	97.46

Table 7: Comparison Accuracy by ML Architectures

ID	List of indicators	Accuracy
1	Total balance of Payment, Net lending/borrowing rate, Net operating balance in public, Nominal Effective Exchange Rate, Economic Activity(Industry Product), Gross Domestic Product, Labor Markets(Employment), Financial Market Prices(Equities), Liabilities (w/ Fund Record), Prices(Consumer Price Index)	97.79%
2	Total balance of Payment, Residential Real Estate Prices, Total Capital Account(Debt), Exports of Goods and Services, Imports of Goods and Services, Economic Activity(Retail Sales), Financial Market Prices(Equities), Assets (w/ Fund Record), Prices(Consumer Price Index), Reserve and Foreign Currency Assets	97.77%
3	Residential Real Estate Prices, Net lending/borrowing rate in public, Net operating balance of government, Total Capital Account(Debt), Economic Activity(Industry Product), Nominal Effective Exchange Rate, Exports of Goods and Services, Gross Domestic Product, Assets (w/ Fund Record), Prices(Consumer Price Index)	97.76%
	All indicators	97.68%

Table 8: Our optimal combinations of economic indicators and those accuracy in this experiment

are selected based on the average of the accuracy of the 25-time execution results. As shown in the table 8, all of the top three combinations show higher accuracy than using all indicators.