International Journal of Internet, Broadcasting and Communication Vol.14 No.1 142-151 (2022)
http://dx.doi.org/10.7236/IJIBC.2022.1.142

IJIBC 22-1-18

A Study on Explainable Artificial Intelligence-based Sentimental Analysis System Model

Mi-Hwa Song

Assistant Professor, School of Smart IT, Semyung University, Jecheon, Korea
mhsong@semyung.ac.kr

Abstract

In this paper, a model combined with explanatory artificial intelligence (xAI) models was presented to secure the reliability of machine learning-based sentiment analysis and prediction. The applicability of the proposed model was tested and described using the IMDB dataset. This approach has an advantage in that it can explain how the data affects the prediction results of the model from various perspectives. In various applications of sentiment analysis such as recommendation system, emotion analysis through facial expression recognition, and opinion analysis, it is possible to gain trust from users of the system by presenting more specific and evidence-based analysis results to users.

Keywords: *Explainable Artificial Intelligence, LIME, Sentimental Analysis, Model Interpretability*

1. Introduction

The use of the term XAI (Explainable Artificial Intelligence) was first introduced in a paper by Buchanan and Shortliffe as “explainable decision-making systems” [1]. In 2004, researchers point out that computer systems and artificial intelligence systems are becoming more complex, but their self-explanatory capabilities remain at a standstill [2]. In this study, an architecture was proposed to explain the reasons for the actions of NPCs by changing the artificial intelligence of NPCs (Non-player Characters) included in the military simulation combat program. Command AI and control AI are separated, and the control AI expresses the situation in the game as a vector. Command AI analyzes the results of behavioral AI and issues commands. Officers who experienced simulated battles with NPCs analyzed the reasons for the actions of the AI after the battle based on the status recognition results of the command AI and the instructions of the control AI, and came to understand the reasons for the actions [2].

In 2016, DARPA, a leading institution for the realization of scientific ideals, is launching a project called Explainable artificial intelligence (XAI) [3]. XAI is a technique that adds explanatory possibilities so that it can be known on what basis an artificial intelligence model made a decision before reaching a specific conclusion. It is mainly used by people involved in machine learning to trust the system. Today's artificial intelligence is becoming more and more complex, and the amount of data that machines learn is too large for

Manuscript Received: December. 15, 2021 / Revised: December. 22, 2021 / Accepted: December. 27, 2021
Corresponding Author: mhsong@semyung.ac.kr
Tel: +82-43-649-1645, Fax: +82-43-649-1747
Professor, School of Smart IT, Semyung University, Korea

humans to interpret. This is why efforts and research are needed to give explanation capabilities to artificial intelligence through XAI and to secure rationality in artificial intelligence and human interaction [4, 5].

With the development of deep learning and CNN, research for interpretation of image data and model results is also increasing [6, 7]. In this study, an interpretable module is combined and applied to the sentiment analysis system model. The predictive and explanatory power of a model is evaluated by combining an interpretable method with a representative sentiment analysis model. By analyzing the results of submodules on observational cases for the best performing word vectorizers and classification models, you can improve the reliability of your models and predictions.

The structure and contents of this paper are as follows. Chapter 2 provides an overview of the sentiment analysis system, an overview of explainable artificial intelligence technologies, and the characteristics of related technologies. In Chapter 3, we propose a model of a sentiment analysis system that combines explainable AI. Chapter 4 describes the form and method of explanatory power through examples for XAI application of the proposed model. Finally, Chapter 5 concludes with conclusions and directions for future research.

2. Theory

2.1 XAI (Explainable AI)

LIME is an algorithm that makes it possible to explain classifier or regression prediction by approximating the data of a local range based on the individual data to be explained. LIME is an XAI technique that tells which areas the model focused on analyzing individual observations and which areas were used as the basis for prediction. In addition, LIME is a method that can be applied regardless of the model learning method (model-agnostic). Therefore, if there is a learning model in use, LIME can be applied to transform each AI model to be explanatory [9].

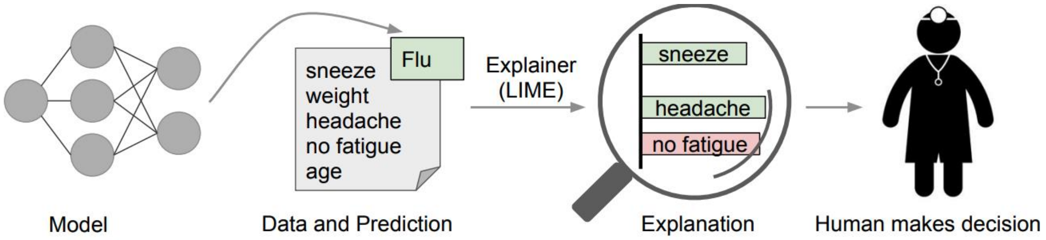


Figure 1. An Illustrative explanation of individual predictions

Figure 1 shows the concept of individual observations through the LIME explainer [9]. This model predicted that this patient would contract the flu, and the LIME explainer visualizes the patient's history by highlighting the cause of the flu prediction. Sneezing and headache contributed to the prediction of "flu", but provided evidence of "no fatigue". This allows health care providers to make informed decisions about whether the model's predictive power can be trusted [9]. The formula for model-agnostic LIME is as follows.

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Model g is an explainable model for instance x, which minimizes the loss function L (e.g. MSE). We measure the loss based on how close the predictions of the explanatory model are to the predictions of the original model f while keeping the model complexity Ω (g) (e.g. fewer variables) low. You can think of Ω(g) as the regularization (Lasso, Ridge, Elastic net, etc.) used in the linear regression model. More regulation means reducing variables to zero, that is, reducing the number of variables, and less regulation means maximizing variables.

SHAP (SHapley Additive exPlanations) is an explanatory method that increases the transparency of a model. Run it on individual instances to illustrate how it works, or show the extent to which SHAP values collected for global description positively or negatively influence the final prediction. When SHAP is used, various methods such as analysis of individual instances, analysis of relationships between specific features, comprehensive analysis of all instances, and analysis of model influence of all features, for the prediction result provided by the AI application system, etc. This has the advantage of being able to explain and analyze the prediction results in detail [10].

3. XAI-based sentiment analysis system model

A model for subject classification for general natural language texts, and polarity prediction for reviews, shows the results of predictions using accuracy metrics. This does not explain the reasons for which pathway led to the outcome of these predictions. The sentiment analysis system model combined with XAI is a structure that enables interpretation of model predictions by additionally applying XAI technology regardless of the type of model. Based on the existing sentiment analysis system model, it is a system configuration diagram that combines an explanatory framework that can be applied regardless of the type of model (Figure 2). It finds the best optimized model and delivers the predicted results of the model to analysts or decision makers. In the process, visual and textual descriptions of how the model behaves can be provided.

3.1 Experimental dataset

Train a model using newsgroup data provided through the scikit learn package and evaluate the model with test data. The dataset consists of 20 news categories. Additionally, the case of applying the description module to the sentiment analysis system using IMDB Movie Reivew data provided by Kaggle [11] will be explained. The IMDB is 25,000 movie review data, labeled as ‘positive’ or ‘negative’.

3.2 Model Training

A machine learning model vectorizes a string to process text data. The vectorizer uses the TF-IDF method. After the preprocessing process, training proceeds with an appropriate machine learning model that can classify news data. The polynomial distribution naive Bayes calculates the probability that a document belongs to a specific category for a vector input. The naive Bayes model probabilistically compares the results of how close the input news is compared to the word usage frequency of the previously trained model. In this case, if the alpha parameter is too small, overfitting may occur, and if it is set too large, underfitting may occur.

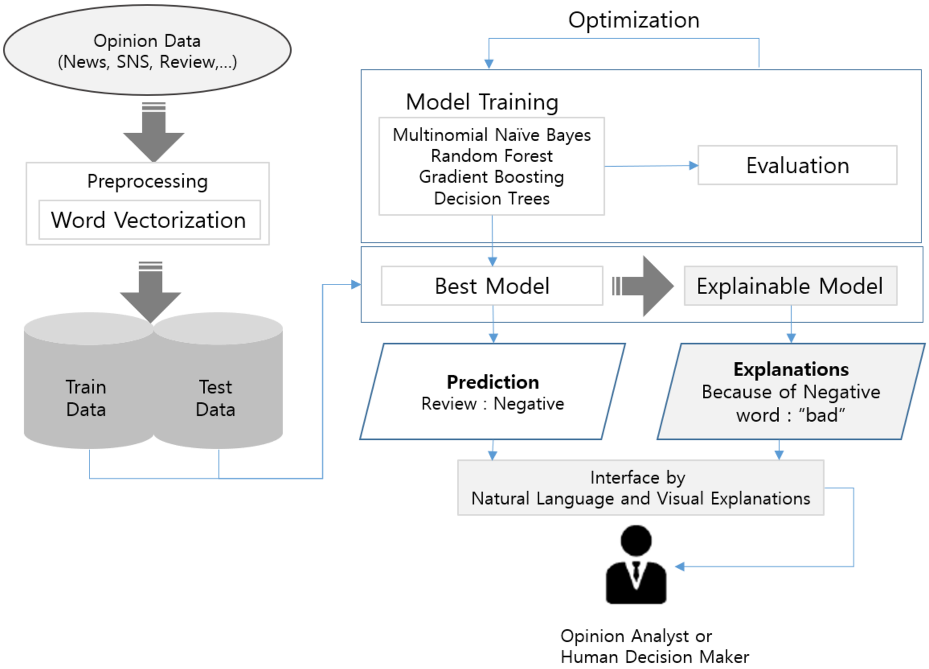


Figure 2. XAI-based Sentiment Analysis System Model

When the training of the model is finished, the performance of the model is measured with test data. The metric is f1-score. The model predicts the test data and compares it with the actual result to output the f1-score.

Table 1 shows the performance of several classifiers other than the naive Bayesian classifier. When TF-IDF vectorizer and NB are used together, the F1-score is 83.5%, indicating that it is the best model.

Table 1. Performance Evaluation of Several Classifier

Classifier	Parameter	F1-score
Multinomial Naïve Bayes	Alpha = 0.01	0.835
Random forest	n_estimators = 500	0.7687

Bagging	Base classifier = KNeighborsClassifier()
	n_estimators = 500
	max_samples = 0.5
	max_features = 0.5
		0.64426

Gradient boosting	n_estimators = 500	0.79093
Decision tree		0.72314
Extra trees	n_estimators = 500	0.81830

4. Interpretation of explanation based on the proposed model

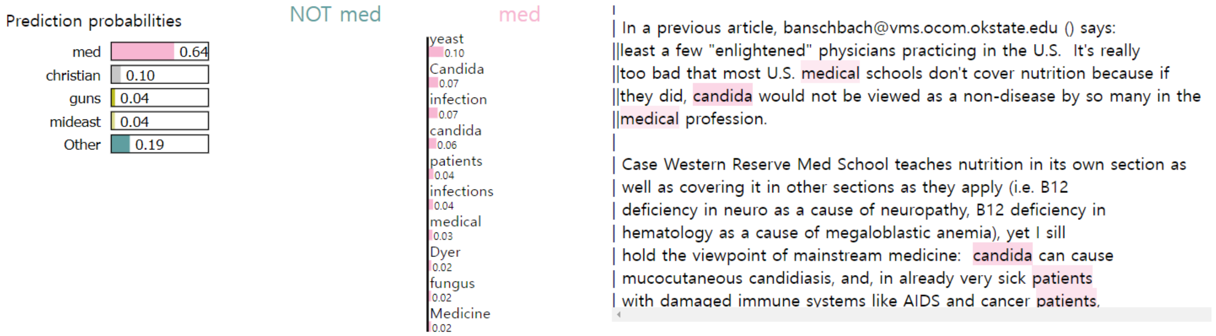
4.1 LIME Explanations of News group dataset

The implementation of LIME applies the text description module provided by Python. LIME explainer provides feature selection, BOW method, and kernel size as parameters. Only basic categories are applied here.

The visual interpretation applied with instance 0 and LIME is shown in Figures 3 and 4, respectively.

```
> In a previous article, banschbach@vms.ocom.okstate.edu () says:
>>least a few "enlightened" physicians practicing in the U.S.    It's really
>>too bad that most U.S. medical schools don't cover nutrition because if
>>they did, candida would not be viewed as a non-disease by so many in the
>>medical profession.
>
> Case Western Reserve Med School teaches nutrition in its own section as
> well as covering it in other sections as they apply (i.e. B12
> deficiency in neuro as a cause of neuropathy, B12 deficiency in
> hematology as a cause of megaloblastic anemia), yet I sill
> hold the viewpoint of mainstream medicine:    candida can cause
> mucocutaneous candidiasis, and, in already very sick patients
> with damaged immune systems like AIDS and cancer patients,
```

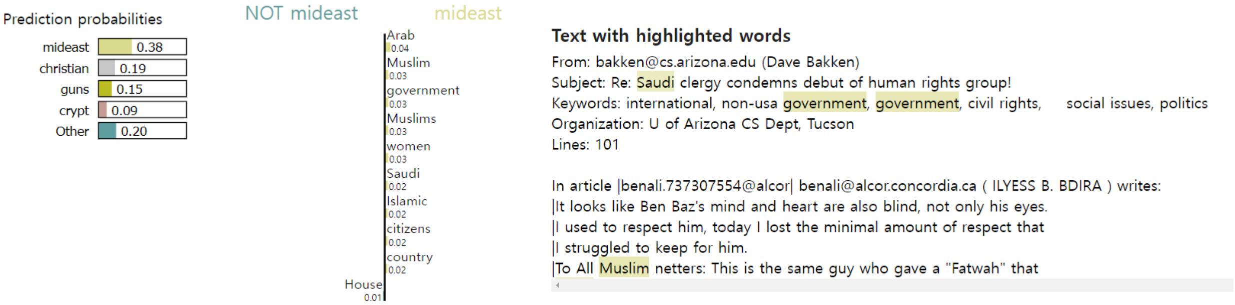
Figure 3. Instance 0 data in news dataset



```
|
| In a previous article, banschbach@vms.ocom.okstate.edu () says:
||least a few "enlightened" physicians practicing in the U.S.  It's really
||too bad that most U.S. medical schools don't cover nutrition because if
||they did, candida would not be viewed as a non-disease by so many in the
||medical profession.
|
| Case Western Reserve Med School teaches nutrition in its own section as
| well as covering it in other sections as they apply (i.e. B12
| deficiency in neuro as a cause of neuropathy, B12 deficiency in
| hematology as a cause of megaloblastic anemia), yet I sill
| hold the viewpoint of mainstream medicine:  candida can cause
| mucocutaneous candidiasis, and, in already very sick patients
| with damaged immune systems like AIDS and cancer patients,
```

Figure 4. Analysis result of LIME description implementation for instance 0

Figure 4 shows the result of the submodule output by the LIME descriptor for the test data instance 0. Submodules are marked with text highlighting. This news data shows that it has been classified into the 'med' category with a 64% probability. Even if you are not a medical professional, it clearly explains why words like 'yeast' and 'infection' fall into the 'med' category.



Text with highlighted words

```
From: bakken@cs.arizona.edu (Dave Bakken)
Subject: Re: Saudi clergy condemns debut of human rights group!
Keywords: international, non-usa government, government, civil rights,     social issues, politics
Organization: U of Arizona CS Dept, Tucson
Lines: 101

In article |benali.737307554@alcor| benali@alcor.concordia.ca ( ILYESS B. BDIRA ) writes:
|It looks like Ben Baz's mind and heart are also blind, not only his eyes.
|I used to respect him, today I lost the minimal amount of respect that
|I struggled to keep for him.
|To All Muslim netters: This is the same guy who gave a "Fatwah" that
```

Figure 5. Analysis result of LIME description implementation for instance 3

Figure 5 shows the analysis result of LIME for test data No.3. Words such as 'muslim', 'saudi', and 'government' can reasonably be taken to classify this article as 'mideast'.

It was confirmed that it works well by applying LIME to text data. LIME was able to confirm the basis of the model's judgment based on the highlighted part through submodule highlighting. LIME's text explainer can help identify the basis for Artificial Intelligence's decision and suggest a direction for model improvement.

4.1.2 LIME Explanation of IMDB Movie Review dataset

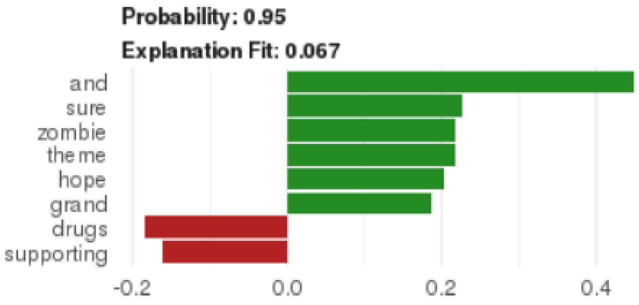


Figure 6. Feature importance plot for instance labeled "Negative"

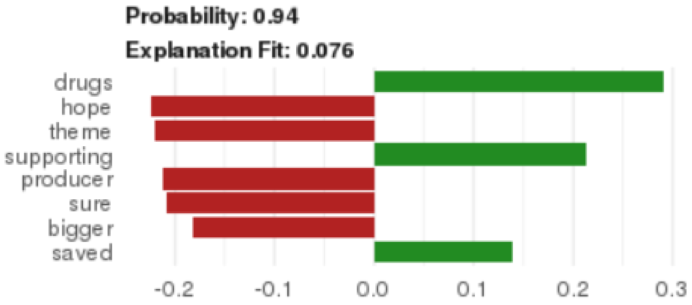


Figure 7. Feature importance plot for instance labeled "Positive"

Figure 6 shows a feature importance plot for individual observations classified as 'negative'. Green shows how much each feature supports the outcome of this prediction, and red shows how much each feature opposes this prediction. Words such as "sure", "hope", and "grand" describe features that support the result predicted as "negative". On the other hand, it can be seen that the word "drugs" is opposite to the prediction result. Figure 7 also shows feature importance plots for individual observations classified as 'positive'.

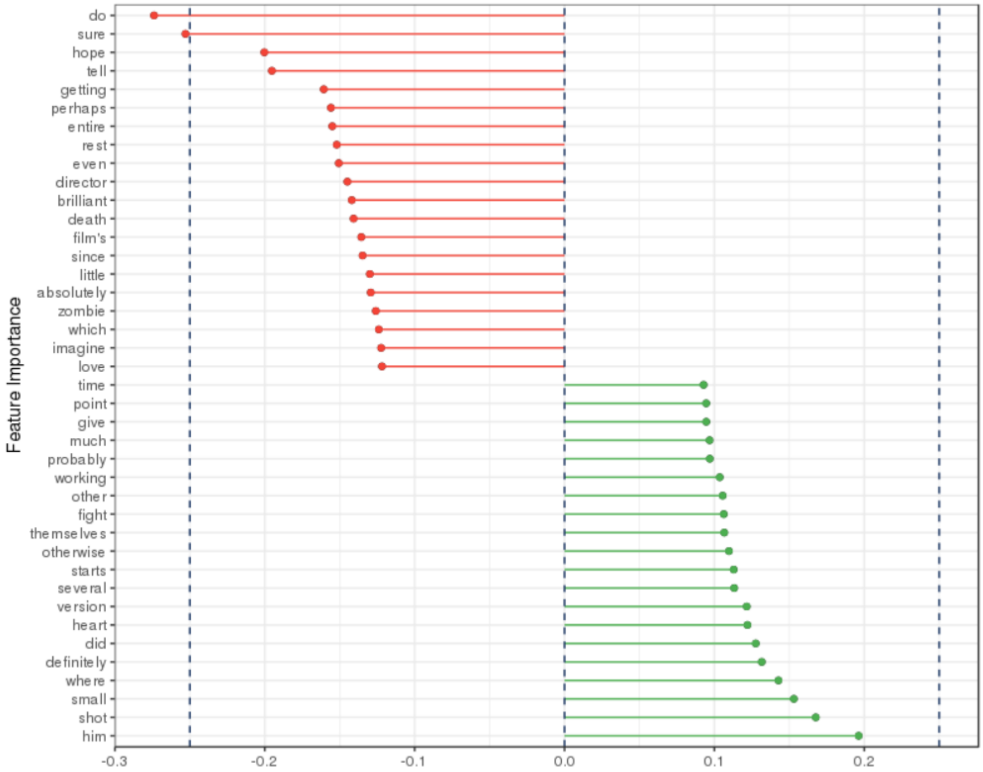
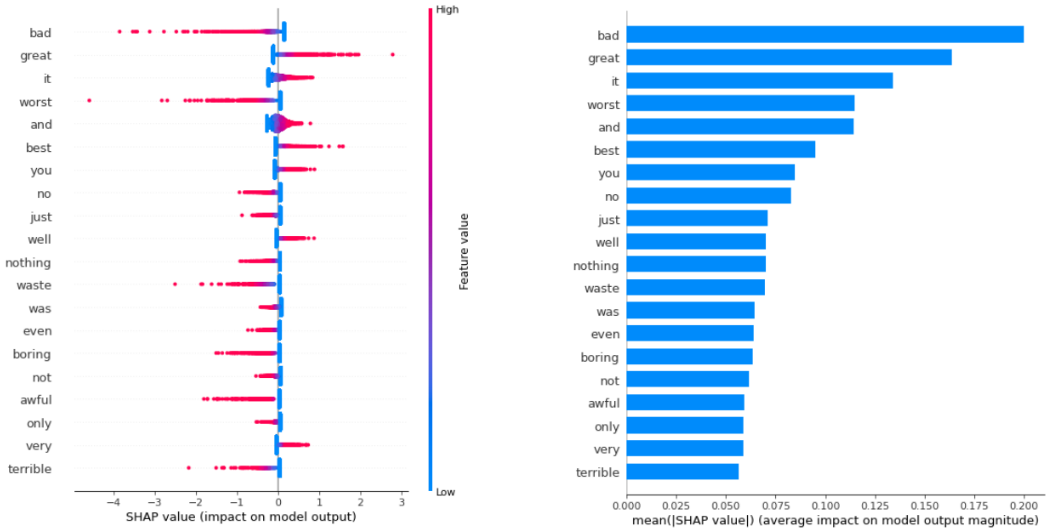


Figure 8. Feature importance rank : Positive and Negative correlations

In Figure 8, we try to visualize and explain the correlation by rank in order to figure out which features are more related to positive and negative reviews than other features. Features marked in red, such as “hope”, “sure”, and “getting”, are commonly used in review comments related to improvement. There might be a review like, "I hope this film takes a more subtle approach and improves the overall quality of the film." And the words appearing in negative movie reviews are expressions such as 'movie', 'all', 'rest', and 'director', indicating that there is a point in the quality of the movie, that is, improvement. Here we need to take a closer look at the words marked in green that lead to positive reviews. It is an effort to explain why rather ambiguous expressions, rather than words with clear patterns or meanings, are commonly used in positive reviews. When commenting praises a movie, one can hypothesize that it is highly dependent on that movie and that specific expressions will dominate. Therefore, a descriptive denoting a positive review is the same as a step word. It can be assumed that the expressions of a review praising a movie may appear in very specific words related to the movie itself.

4.2 SHAP based Explanations of IMDB dataset

4.2.1 Interpretation of influence for all features



(a) Summary plot (Scatter)

(b) Summary plot (bar)

Figure 9. Visual interpretation of how it affects the distribution of shapley values for all features (a) scatter plot (b) bar plot

Figure 9 (a) explains the effect of changes in the top 20 token values on the shapley value through a scatter plot graph. It is that features located at the top such as "bad", "great", it", and "worst" contribute more to the model than words located at the bottom such as "only", "very", "terrible". "bad" has the greatest variance. This is also confirmed in the bar graph of (b). The biggest word that determines the positive/negative of a review is 'bad', and the degree of influence is described in the order of 'great', 'it', 'worst', etc.

4.2.2 Interpretation of shape values for individual instances

In Figures 10 and 11, red is a feature that helps to increase the influence of positive reviews, and on the contrary, it helps to lower the influence of negative comments.

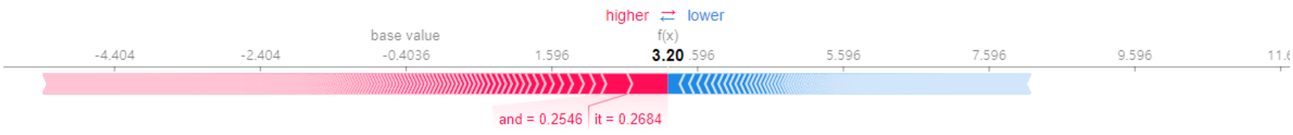


Figure 10. Instances identified as positive reviews (Instance 5) – force_plot (positive)

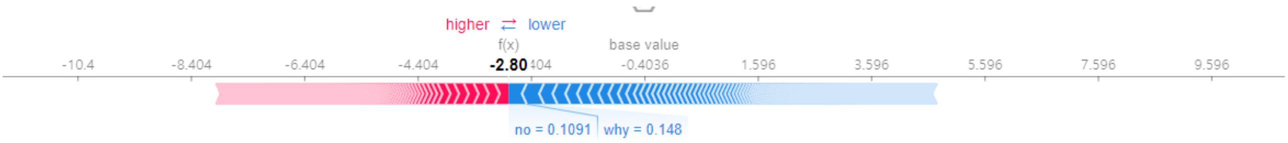


Figure 11. Instances identified as negative reviews (Instance 6) – force_plot (negative)

SHAP Explainer provides local explanations for individual instances. Figure 10 is a case predicted by a positive review. As shown in Figure 10, the execution result is provided based on the shapley value calculated according to the feature value. The darkest printed shapley value is 3.20, and this review predicted positive. It can be seen that the most important features that affected the prediction results appeared in the order of 'and' and 'it'. Notice that 'and' becomes the most important feature among the text corpus. It is interpreted to mean that it captures a high-level concept of text structure. Having a lot of "and" clearly indicates a more positive review. In Figure 11, the most distinctly output shapley value is -2.80. This means that this review was predicted to be negative. It can be seen that the features that most affected the results of this prediction appeared in the order of 'no' and 'why'.

5. Conclusions

In this paper, a model combined with explanatory artificial intelligence (xAI) models was presented to secure the reliability of machine learning-based sentiment analysis and prediction. The applicability of the proposed model was tested and described using the IMDB dataset. This approach has an advantage in that it can explain how the data affects the prediction results of the model from various perspectives. In various applications of sentiment analysis such as recommendation system, emotion analysis through facial expression recognition, and opinion analysis, it is possible to gain trust from users of the system by presenting more specific and evidence-based analysis results to users.

In this study, emotion analysis data and explanatory power of the model were analyzed using the results derived through LIME and SHAP, which are representative model agnostic xAI algorithms. This approach will not be easily understood by ordinary users. In addition, it leads to the need for research to develop an HCI-based explanatory interface module that interprets the model results.

Later stages of this study will need to verify interpretive power through the implementation of the proposed sentiment analysis system. To do this, we need to introduce a computable performance metric for the descriptor, and we need to consider the actual user satisfaction rating together.

Acknowledgement

This work was supported by the Semyung University Research Grant of 2021.

References

[1] E. H. Shortliffe and B. G. Buchanan, "A model of inexact reasoning in medicine," *Mathematical Biosciences*, vol. 23, no. 3, pp. 351-379, 1975/04/01/ 1975, doi: https://doi.org/10.1016/0025-5564(75)90047-4.

[2] M. Van Lent, W. Fisher, and M. Mancuso, "An explainable artificial intelligence system for small-unit tactical behavior," in *Proceedings of the national conference on artificial intelligence*, 2004: Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, pp. 900-907.

[3] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "XAI—Explainable artificial intelligence," *Science Robotics*, vol. 4, no. 37, p. eaay7120, 2019.

[4] L. Edwards and M. Veale, "Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for," *Duke L. & Tech. Rev.*, vol. 16, p. 18, 2017.

[5] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," presented at the The 5th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2018), 2018/05/31, 2018. [Online]. Available: https://arxiv.org/pdf/1806.00069.

[6] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, 2014: Springer, pp. 818-833.

[7] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann, "Devnet: A deep event network for multimedia event detection and evidence recounting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2568-2577.

[8] S. M. Mathews, "Explainable artificial intelligence applications in NLP, biomedical, and malware classification: a literature review," in *Intelligent computing-proceedings of the computing conference*, 2019: Springer, pp. 1269-1292.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, 2016. [Online]. Available: https://doi.org/10.1145/2939672.2939778.

[10] D. Dave, H. Naik, S. Singhal, and P. Patel, "Explainable ai meets healthcare: A study on heart disease dataset," *arXiv preprint arXiv:2011.03195*, 2020.

[11] "Sentiment Analysis of IMDB Movie Reviews." https://www.kaggle.com/lakshmi25npathi/sentiment-analysis-of-imdb-movie-reviews/data (accessed.