Explainable Artificial Intelligence (XAI) towards Model Personality in NLP task

Nadhila Nurdin[1*], Dimas Adi[1]

***Abstract*— In recent years, the development of Deep Learning in the field of Natural Language Processing, especially in sentiment analysis, has achieved significant progress and success. It is because of the availability of large amounts of text data and the ability of deep learning techniques to produce sophisticated predictive results from various data features. However, the sophisticated predictions that are not accompanied by sufficient information on what is happening in the model will be a major setback. Therefore, the significant development of the Deep Learning model must be accompanied by the development of the XAI method, which helps provide information about what drives the model to get predictable results. Simple Bidirectional LSTM and complex Bi-GRU-LSTM-CNN model for Sentiment Analysis were proposed in the present research. Both models were analyzed further using three different XAI methods (LIME, SHAP, and Anchor) in which they were used and compared to two proposed models, proving that XAI is not limited to giving information about what happens in the model but can also help us to understand and distinguish models' personality and behaviour.**

***Keywords*—Deep learning, Explainable artificial intelligence, Natural language processing, Sentiment analysis**

I. INTRODUCTION

In recent years, Machine Learning and Deep Learning have yielded cutting edge execution in autonomous systems, computer visions, prediction, and classification tasks. Various prediction tasks and classification that include diverse data such as images, text, speech, and video have been able to achieve near-human or even better accuracy. In any case, it is conceivable that individuals stress over how to ensure a deep learning model settles on the right decision when the high accuracy obtained from complex models that even experts find hard to explain. Deep learning techniques help to diagnose in the healthcare area; for instance, in work [1] apply it to detect breast cancer. In order to check the results obtained, [1] needs to consult with the human doctor to make sure the diagnosis is rational.

Many researchers assign themselves to develop explainable machine learning methods In order to fulfil the needs for trust and improvements in deep learning models. For instance, a modular and extensible approach for explaining the prediction of any classifier in an interpretable manner which is named Local Interpretable Model agnostic Explanations (LIME) [2], approximate the classifier locally with an interpretable model and use a linearly weighted combination of input features to explain the prediction. However, this technique coverage (the region where explanation applies) is unclear since this explanation is somehow local [3]. Following a novel rule-based method, a model-agnostic explanation called anchors was introduced; the method is based on if-then rules. Moreover, there are also many frameworks created to make the explainability task easier. For example, the work in [4] presents a framework that uses a unified approach to interpreting model prediction called SHAP (SHapley Additive exPlanations). SHAP has the goal to explain the prediction of an instance x by computing the contribution of each feature to the prediction.

Natural Language Processing applications such as machine translation, speech recognition, and information retrieval have been widely used. The mutual relationships between language, society, and the individual are the key to the impact of NLP in societal aspects. As social media has recently become part of peoples everyday lives, NLP can have a direct influence on individual users' lives [5]. Many NLP tasks are now able to reach impressive performance by developing Deep Learning. Furthermore, in NLP tasks, explainable methods need to be produced to make the Deep Learning model more transparent and understandable in order to prevent the risk of harm to subjects, especially for medical application. The previous work focused on the explainability models on the NLP field [6]. Hu Jin in [6] implemented the explainable method with the combination of Bi-directional LSTM and CRF (Bi-LSTMCRF) used in Named Entity Recognition (NER).

In this paper, a comparison of three different methods of models explanation (LIME, SHAP, and Anchor) were performed with two different architecture models of Deep Learning on the sentiment analysis task. By utilizing three different Explainability methods to perform analysis on these two models, the distinct behaviour of the models that enables the process of disclosing the model's personalities were able to be investigated. Model's personality explains more about models rationale, models strengths and weaknesses, and thus provide an understanding of their future behaviour [7].

A. Sentiment Analysis

Sentiment Analysis (SA), also known as Opinion Mining (OM), was originally defined as aiming to infer positive or negative opinions/sentiments from the text. The purpose of this approach is to discover the other pieces of associated information, which is important for practical applications of the opinions [8]. Opinions could direct or affect people's choices in which they look for the view of others

[1] AGH University of Science and Technology, al. A. Mickiewicza 30, 30-059 Krakow, Poland. E-mail: nadhila@student.agh.edu.pl; nadhilanurdin@gmail.com

until they agree about their own acts. Opinion Mining could play an essential role in the daily lives such as for business, marketing, recommender systems, government, Web monitoring, etc. In the era of social Big Data, Sentiment Analysis is also used to describe, predict, and determine human behaviour in social media. Text analysis is one of the key elements to process this task since 80% of internet data is text [9]. Sentiment Analysis is the science of using text analysis to determine the polarity of the sentiment (positive, negative, or neutral). Sentiment classification is often used as the approaches to classify the sentiment polarity of a sentence. This approach involves several techniques such as Natural Language Processing (NLP) and Machine Learning (ML).

B. NLP and ML Methods

Most of the works in the field of NLP were focused on feature engineering and extraction, such as bag-of-words, TF-IDF, and N-grams techniques [10]. Formerly, those techniques have been used in combination with conventional machine learning. Word embedding and neural network-based models such as recurrent neural networks (RNN) and long short term memory (LSTM), which are methods that are able to extract individual features, automatically made those methods recently become popular for executing sentiment classification. In recent times, Deep Learning techniques for solving various Natural Language Processing tasks (e.g. Sentiment Analysis, word embedding, machine translation, and named entity recognition) have become popular methods. The end-to-end model is used by these approaches to learn and extract the feature, which thereafter executes the classification. This technique has been able to outperform conventional Machine Learning and achieve state-of-the-art result performance [11].

In work [12], four deep recurrent architectures dealing with the task of offensive tweet detection were reported. Cambray et al. [12] built Neural Network architectures that are based on LSTMs and GRUs with a simple bidirectional LSTM as a baseline system and then further increased the complexity of the models by adding convolutional layers. The result of the work showed that the simple architecture performed slightly better than the complex one. LSTM (Long Short Term Memory) method has the ability to extract individual features automatically since it is designed to capture long term dependencies [13]. GRU (Gated Recurrent Unit) is similar to an LSTMs unit but without an output gate [14]. This has made those of two methods recently become popular for executing sentiment classification and implementing a split- process-merge architecture with LSTM and GRU as processors. Moreover, work [15] build a framework to improve the accuracy of sentiment analysis using an ensemble of CNN and bidirectional LSTM (Bi-LSTM) networks and test them on popular sentiment analysis databases such as the IMDB review and SST2 datasets. Despite the performances of the above studies, these models are usually applied in a black-box manner which is lack transparency. Therefore, a simpler explanation to understand how the model works is required in such a complex model. The combination of Bi-LSTM, GRU, and CNN and the single Bi-LSTM provided by [12] with some additional layer on the architecture to perform the sentiment analysis task were performed in this paper. Additionally, the work with explainability techniques in order to see how exactly the model works were conducted. By choosing those of two models, the difference of models personalities in the simple one (Bi-LSTM) and the complex one (Bi-GRU-LSTM- CNN) was able to be observed.

C. XAI (Explainability in Artificial Intelligence)

The aspects of life, life efficiency, and human capacities are gradually affected by the current development of AI. AI's inability to explain the decisions and actions of humans has been tackled by the advancement of Deep Learning systems. However, Deep Learning could not completely provide trusted information about what happened in the system. By the recent literature survey [16], it is proven that the black-box system imposes vulnerability on our society [17]. Therefore, XAI is needed to support the output of a Deep Learning model. for the prescription system, a simple percentage number of binary predictions or not enough for experts that work in this system for supporting their diagnosis [18].

The need of the user to understand, properly trust and be able to effectively manage the output of the Machine Learning or Deep Learning model afterwards makes Defense Advanced Research Projects Agency (DARPA) create the Explainable Artificial Intelligence (XAI) project [19]. Furthermore, in an effort to create more explanation techniques that can be interpreted, easily traced and can be trusted, because basically, humans will prefer systems that have these criteria [20]. Google, the one that initiated People +AI Research (PAIR) also working on this method.

1) LIME: LIME (Local Interpretable Model-agnostic Explanations) is a model agnostic technique that produces linear approximations by taking random samples in the local neighbourhood from a complex model, which is then adjusted a simpler linear model for a new synthetic data set. LIME is not generating an explanation for the whole model but rather explains the model only at that locality [3].
2) SHAP: SHAP (SHapley Additive exPlanations) [5] works by computing the shapley value. This value, originating from a coalitional theory of games, explains the contribution of every feature of an instance and a method to fairly distribute the payout among the features [21].
3) ANCHORS: ANCHORS [4] is similar to LIME; local explanations for predictions of the black-box Deep Learning model are generated by deploying a perturbation-based strategy. However, the resulting explanations in Anchors are conveyed as easy-to-understand IF-THEN rules.

The remaining of this paper is structured as follows. In section II, we present the details of the whole sentiment Analysis model and Explainability method. The experimental results are discussed in Section III, while section IV concludes this paper.

II. METHOD

In this paper, two different Artificial Neural Network models for processing US Airline Sentiment data [22] were proposed. Dataset contains Tweet sentiment about six US Airlines which have been labelled with a positive, negative, and neutral. The label data with numbers 0, 1, and 2 were changed as negative, positive, and neutral, respectively. Furthermore, common data cleaning step for tweet data such as the lowering case, punctuation removing, URL removing, and mention removing were performed. In addition, for the pre-processing step, sequence padding with sequence pads() function from Keras [23] was utilized, and the word embeddings with a randomly initialized embedding layer were trained during the execution of the whole model. As described in section II, Bi-LSTM and Bi- GRU-LSTM-CNN from the previous architecture model [12] was chosen and improved by adding the dropout layer and dense layer. The additional layer functioned to prevent and minimize the overfitting of the result. The diagrams of the proposed architecture models with additional layers are shown in Figure 1 and 2, respectively. LIME, SHAP, and Anchor methods were implemented to explain the models then all of the obtained results were compared.

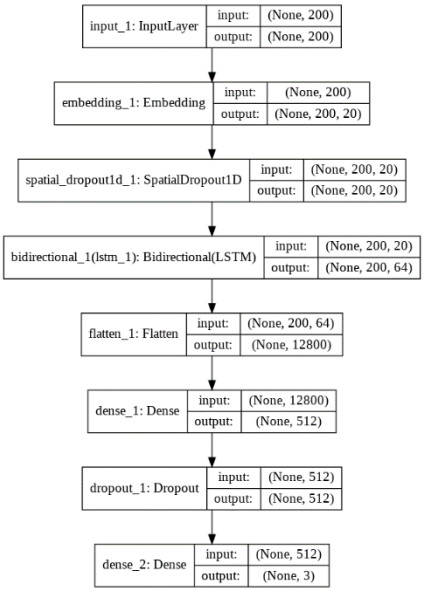


Figure 1. Architecture of Proposed Bi-LSTM

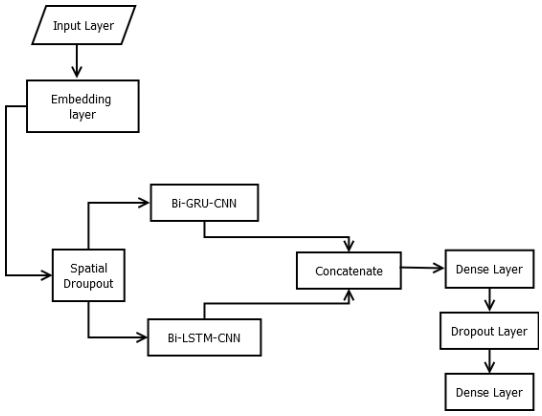
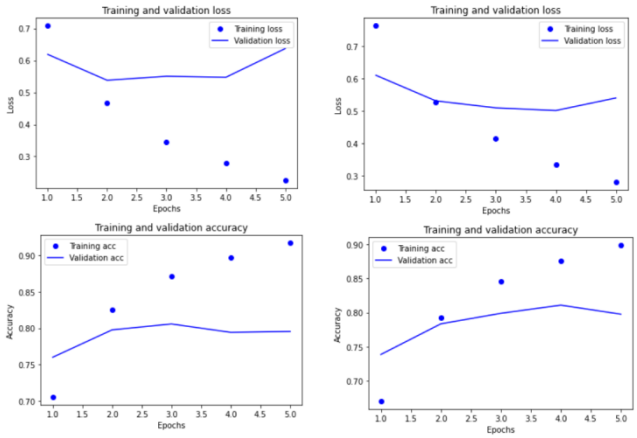


Figure 2. Architecture of Proposed Bi-GRU-LSTM-CNN

III. RESULTS AND DISCUSSION

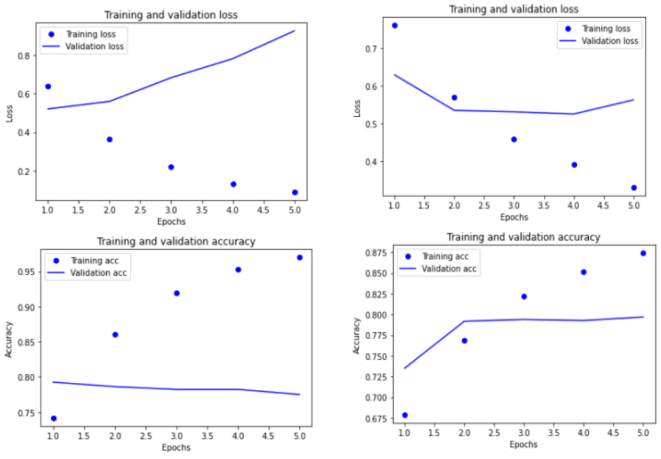
A. *Model Performance and Comparison*

Bi-LSTM and Bi-GRU-LSTM-CNN reported in [12] (without additional layer) were compared with the model we developed. Based on Figure 3 and Figure 4, respectively, it could be observed that the model with additional layers gives better learning in the training process.



a) Bi-LSTM **b) Proposed i-LSTM**

Figure 3. Learning curve of a) Proposed Bi-LSTM and b) Bi-LSTM.



a) Bi-GRU-LSTM-CNN **b) Proposed Bi-GRU-LSTM-CNN**

Figure 4. Learning curve of a) Bi-GRU-LSTM-CNN b) Proposed Bi-GRU-LSTM-CNN

The model with additional layers also provides better accuracy. The detailed model performance and metrics of the model are shown in table 1 and 2, respectively.

TABLE 1.
MODEL METRICS.

Class	Model	Precision	recall	F-1
Negative	Bi-LSTM	0.87	0.86	0.86
	Improved Bi-LSTM	0.85	0.91	0.87
	Bi-GRU-LSTM-CNN	0.83	0.89	0.86
	Improved Bi-GRU-LSTM-CNN	0.84	0.90	0.87
Positive	Bi-LSTM	0.69	0.67	0.68
	Improved Bi-LSTM	0.76	0.66	0.71
	Bi-GRU-LSTM-CNN	0.72	0.59	0.65
	Improved Bi-GRU-LSTM-CNN	0.76	0.66	0.70
Neutral	Bi-LSTM	0.60	0.64	0.62
	Improved Bi-LSTM	0.66	0.59	0.62
	Bi-GRU-LSTM-CNN	0.61	0.58	0.59
	Improved Bi-GRU-LSTM-CNN	0.63	0.57	0.60

TABLE 2.
MODEL PERFORMANCE ACCURACY.

Model	Accuracy
Bi-LSTM	0.78
Improved Bi-LSTM	0.80
Bi-GRUN-LSTM-CNN	0.77
Improved Bi-GRU-LSTM-CNN	0.79

B. *Explainability on LIME*

The result of LIME Explainer on Model I in Figure 5 showed that word "not" not only had the highest value on negative class but also lowered the prediction on the positive class words in the sentence that came after it. Since the word, "not" indicated such a strong negative signal.

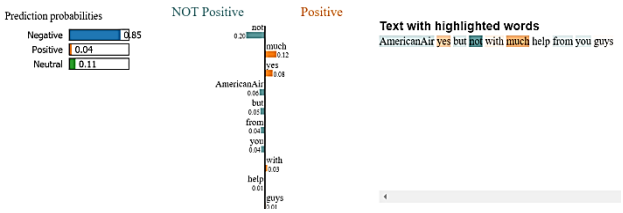


Figure 5. Result of LIME Explainer on Bi-LSTM Model(Model I).

In Figure 6, the same phenomenon could not be discovered as we saw on Model I. Model II could not be focused better on the important feature that will affect the sentiment result. From both examples in Figure 5 and Figure 6, distinct behaviour on both models could be pointed out. The Model I was able to predict the class with the variations of a probability value on all features, and it could focus better on the important feature. On the other hand, Model II distributed almost the same value in each feature.

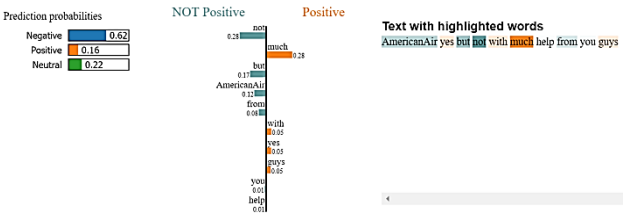


Figure 6. Result of LIME Explainer on Bi-LSTM Model(Model II).

C. *Explainability on SHAP*

The result of Explainability on Model I in Figure 7 showed the word "not" strongly pushed the output prediction value. The Explanation model showed that Model I gave a higher prediction probability to the feature that has a significant impact on the final prediction results, such as the word "not" in this case.

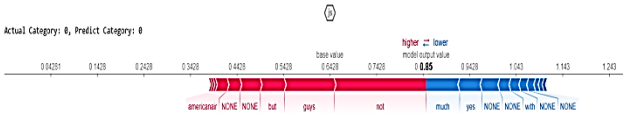


Figure 7. Result of SHAP Explainer on Bi-LSTM Model(Model I),

On the other hand, the Explanation of Model II in Figure 8 showed that the model gave almost balance value to every feature. Model II took many considerations on every feature that had a contribution to the final prediction result. The personalities of both models could be analogized as the child who does not take many considerations to make a decision (Model I) and the elderly who act wisely before making decisions (Model II).

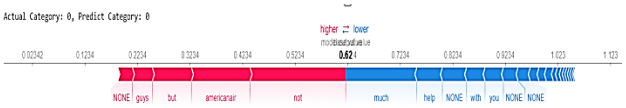


Figure 8. Result of SHAP Explainer on Bi-GRU-LSTM-CNN Model(Model II),

D. *Explainability on Anchor*

The explanation of both models in Figure 9 and 10 showed Model I produced a higher probability than Model II, which was only 83.2 %. However, on Model II, the word ”not” was identified by Anchor as the most influential word feature for the final prediction.

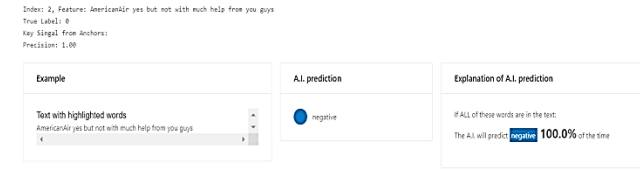


Figure 9. Result of Anchor Explainer on Bi-LSTM Model(Model I).

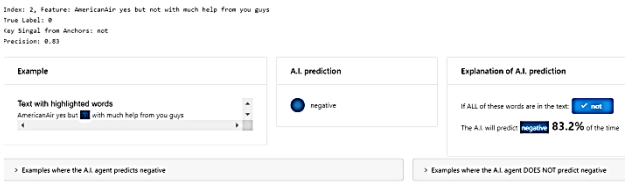


Figure 10. Result of Anchor Explainer on Bi-GRU-LSTM-CNN Model(Model II).

IV. CONCLUSION AND FUTURE WORK

Three different methods of XAI (LIME, SHAP, ANCHOR) on two similar models (one is rather simple, the other one is complex) of Natural Language Processing on Sentiment Analysis were implemented. Our contribution has been revealed that by improving the models' architecture for the Sentiment Analysis and analyzing the proposed models by XAI methods, the distinct behaviour of the models could be discovered.

The points that we can conclude and suggest from this research to further research are as follows. LIME explainer helps to understand what features contribute to the prediction and for further understanding are deepened by deploying the SHAP explainer on the model. A more detailed examination of features that drive predictions higher or lower has been made possible by SHAP explainer. Extending the model sentiment is required to obtain better performance accuracy, such as using the pre-trained word embedding. More research on the comparison of the different explainers to similar models is required to understand and better predict the behaviour of the models. Our research has shown that XAI is capable of explanation personalization via user interaction. It is possible to improve the prediction of models not only motivated by higher accuracies but also by a deeper understanding of what Deep Neural Network models actually do.

REFERENCES

[1] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," arXiv preprint arXiv:1606.05718, 2016.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD Int. Conf. on knowledge discovery and data mining, 2016, pp. 1135–1144.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[4] S. Lundberg, and S. I. Lee, "A unified approach to interpreting model predictions," in Advances in neural information processing systems, pp. 4765–4774, 2017.

[5] D. Hovy, and S. L. Spruit , "The Social Impact of Natural Language Processing," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Berlin, Germany, pp. 591–598, Aug. 2016. https://www.aclweb.org/anthology/P16-2096

[6] J. Hu, Explainable Deep Learning for Natural Language Processing, 2018.

[7] Kzuzuo, Personality seen in natural language deep learning model, and its interpretation (SHAP etc.) and prospect, 2020, https://https://medium.com/@kzuzuo/personality-seen-in-natural- language-deep-learning-model-and-its-interpretation-shap-etc-29465ddee25 [accessed June. 14 , 2020].

[8] B. Liu, "Handbook Chapter: Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing," Handbook of Natural Language Processing. Marcel Dekker, Inc. New York, NY, USA, 2009.

[9] I. E. Alaoui, Y. Gahi, R. Messoussi, Y. Chaabi, A. Todoskoff, and A. Kobi "A novel adaptable approach for sentiment analysis on big social data," J. Big Data, vol. 5, pp. 12, 2018.

[10] A. Dey, M. Jenamani, and J. J. Thakkar, "Lexical TF-IDF: An n-gram feature space for cross-domain classification of sentiment reviews," in Int. Conf. on Pattern Recognition and Machine Intelligence, 2017, pp. 380–386.

[11] A. Sboev, I. Moloshnikov, D. Gudovskikh, A. Selivanov, R. Rybka, and T. Litvinova, "Deep Learning neural nets versus traditional machine learning in gender identification of authors of RusProfiling texts," Procedia computer science, vol. 123, pp. 424–431, 2018.

[12] A. Cambray, and N. Podsadowski, "Bidirectional Recurrent Models for Offensive Tweet Classification," arXiv preprint arXiv:1903.08808, 2019.

[13] J. Schmidhuber, and S. Hochreiter, "Long short-term memory," Neural Comput, vol. 9, pp. 1735–1780, 1997.

[14] K. Cho, B. V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.

[15] S. Minaee, E. Azimi, and A. A. Abdolrashidi, "Deep-Sentiment: Sentiment Analysis Using Ensemble of CNN and Bi-LSTM Models," arXiv preprint arXiv:1904.04206, 2019.

[16] A. Adadi, and M. Berrada , "Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018.

[17] I. Saftić, "Max Tegmark, Life 3.0: Being Human in the Age of Artificial Intelligence," Croatian Journal of Philosophy, vol. 18, pp. 512–516, 2018.

[18] E. Tjoa, and C. Guan, "A survey on explainable artificial intelligence (xai): Towards medical xai," arXiv preprint arXiv:1907.07374, 2019.

[19] DARPA, Explainable Artificial Intelligence (XAI), 2020, https://www.darpa.mil/program/explainable-artificial-intelligence [accessed June. 14 , 2020].

[20] J. Zhu, A. Liapis, S. Risi, R. Bidarra, and G. M. Youngblood, "Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation," in 2018 IEEE Conference on Computational Intelligence and Games (CIG), 2018, pp. 1–8.

[21] S Lipovetsky, and M Conklin, "Analysis of regression in game theory approach," Applied Stochastic Models in Business and Industry, vol. 17, pp. 319–330, 2001.

[22] Kaggle, US AIR tweets Dataset, 2020, https://www.kaggle.com/crowdflower/twitter-airline-sentiment [accessed June. 14 , 2020].

[23] François Chollet and others, Keras, 2015.