2022 5th International Conference on Networking, Information Systems and Security: Envisage Intelligent Systems in 5g//6G-based Interconnected Digital Worlds (NISS) | 978-1-6654-5363-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/NISS55057.2022.10085611

The Investigation into Deep Learning Classifiers Towards Imbalanced Text Data

Luthfia Nur Azizah*‡, Purnomo Husnul Khotimah†§, Andria Arisal†, Andri Fachrur Rozie†, Devi Munandar†,
Dianadewi Riswantini†, Ekasari Nugraheni†, Wiwin Suwarningsih†, Dian Kurniasari*

**Faculty of Mathematics and Natural Sciences*
University of Lampung
Bandar Lampung, Indonesia
‡luthfiaanurazizah@gmail.com

†*Research Center for Informatics*
National Research and Innovation Agency
Bandung, Indonesia
§purn005@brin.go.id

***Abstract*—Class imbalance is an important classification problem where failure to identify events can be hazardous due to failure of solution preparation or opportune handling. Minorities are mostly more consequential in such cases. It is necessary to know a reliable classifier for imbalanced classes. This study examines several conventional machine learning and deep learning methods to compare the performance of each method on dataset with imbalanced classes. We use COVID-19 online news titles to simulate different class imbalance ratios. The results of our study demonstrate the superiority of the CNN with embedding layer method on a news titles dataset of 16,844 data points towards imbalance ratios of 37%, 30%, 20%, 10%, and 1%. However, CNN with embedding layer showed a noticeable performance degradation at an imbalance ratio of 1%.**

***Index Terms*—deep learning, classifier, online news, performance**

I. Introduction

Text classification is a task in Natural Language Processing (NLP), which is the process of categorizing groups (classes) in a document based on previously known (supervised) classes [1], [2]. In binary classification problems, there is often a class imbalance where there are fewer positive classes than negative classes. This causes classifiers to tend to misclassify, i.e., data that should belong to the minority class are classified as the majority class [3]. This mistake is problematic because minority classes are often considered classes of interest.

The imbalanced class has been the subject of research for over a decade. According to L'heureux et al. [4], Japkowicz and Stephen showed that in an experiment, the severity of an imbalanced class problem depends on the complexity of the problem, the coefficient of the imbalanced class, and the total training sample size. As the class inequality coefficient increases, the problem of bias toward the majority becomes more serious [5].

Coronavirus Disease 2019 (COVID-19) is a global pandemic that first reported in Wuhan, China, in late December 2019, when patients contracted pneumonia of unknown cause. Since then, COVID-19 has rapidly spread to Thailand, Japan, South Korea, and virtually all countries in the world, with the primary case of COVID-19 declared in Indonesia on March 2, 2020 [6]. Led by this phenomenon, news about COVID-19 spread to virtually all media, including print, electronic, and online [7].

The speed, convenience, and lasting relevance of everything that happens in the community determine that online news portals have become one of the most important means of disseminating information and are becoming the medium that populations often use as their primary source of information [8]. Therefore, online news portals can become an unofficial yet dependable data source for monitoring the development of COVID-19 in near real-time [9].

According to Launa [10], the Pan American Health Organization (PAHO, 2020) has released COVID-19 news items, a dramatic increase since the COVID-19 pandemic was reported in late December 2019. In March 2020, approximately 19,200 people published papers on Google Scholar with common terms such as ”coronavirus”, ”covid19” or ”pandemic”. However, not all news articles extracted using generic terms contain reports of actual COVID-19 events, and some articles also contain reports of COVID-19 information. Extracting the content of the whole news would be a cumbersome calculation. According to Nugraheni et al. [9], the main topic of a news article can be identified by the news title, which can be used for classification based on which news is a COVID-19 event report (event news) and which news is COVID-19 information (non-event news).

Reliable classifiers for imbalanced data are an important research topic [3], as extreme imbalance classes are naturally participate in a lot of applications required in real-life, such as fraud detection and chemistry [11]. Currently, many studies have used classifiers of deep learning in imbalanced classes [1], [9], [12]–[14]. However, it is rare to find a discussion of which deep learning method is the most reliable as classifiers

Authorized licensed use limited to: Badan Riset Dan Inovasi Nasional. Downloaded on April 05,2023 at 04:41:22 UTC from IEEE Xplore. Restrictions apply.

against imbalanced classes. Our study is trying to fill this gap. We select several traditional machine learning methods and deep learning methods and compare the performance of each method. Additionally, we conduct sampling on COVID-19 online news titles to create different class imbalance ratios. The rest of this research consists of the following sections: Section 2 discuss our positioning towards the related studies, Section 3 presents our proposed method, Section 4 describes the results and analysism, and Section 5 concludes our study.

II. Related Studies

Imbalanced classification of data is a necessary research topic that arises in real-world situations where some data classes are often underrepresented and the inability of standard classifiers to opportunely distinguish underrepresented classes [15]. This inability resulted in misclassification that failed to detect COVID-19 events and resulted in stakeholders, such as the government, health institutions, communities, and others, being unable to prepare the responses or policies needed to deal with the COVID-19 pandemic.

Various methods have been described in the literature for solving the class imbalance problem, one of which is the ensemble technique. Boosting is the most popular and effective iterative ensemble technique. [16]. Freund and Schapire introduced AdaBoost as one of the earliest boosting techniques. AdaBoost needs a base classifier to create a more powerful and stable classifier [17]. In the literature, basic classifiers in the form of classical models such as SVM, Neural Network, and Naive Bayes are mostly chosen [16].

A previous study conducted by [17] in assessing the performance of three classifiers, namely, Logistic Regression, AdaBoost, and XGBoost, which were respectively applied to three datasets with three degrees of imbalance, concluded that the best method for highly imbalanced data is Logistic Regression. However, in a study conducted by [15], applying TF-IDF to SVM with a linear kernel and Logistic Regression to highly imbalanced datasets yielded more biased performance results than decision trees. With the development of deep learning, several recent studies [2], [9], [14], [18]–[23] show that deep learning methods are effective for text classification. However, its performance against imbalanced datasets is rarely discussed. Therefore, we used existing deep learning methods (MLP, CNN, and LSTM) as classifiers to investigate this issue. In addition, traditional machine learning methods (Naive Bayes, Logistic Regression, Decision Tree, SVM, AdaBoost, and Neural Network) were used as a comparison. We used the COVID-19 online news dataset to simulate an imbalanced dataset.

III. Methodology

We ran a query on Indonesia’s national online news portal to amass online news titles about COVID-19 that were posted predicated on designated keywords. Using simple random sampling, online news headlines labeled with events were sampled. We compare the performance of conventional machine learning and deep learning methods to find the most reliable model for classifying text data based on multiple class imbalance ratios.

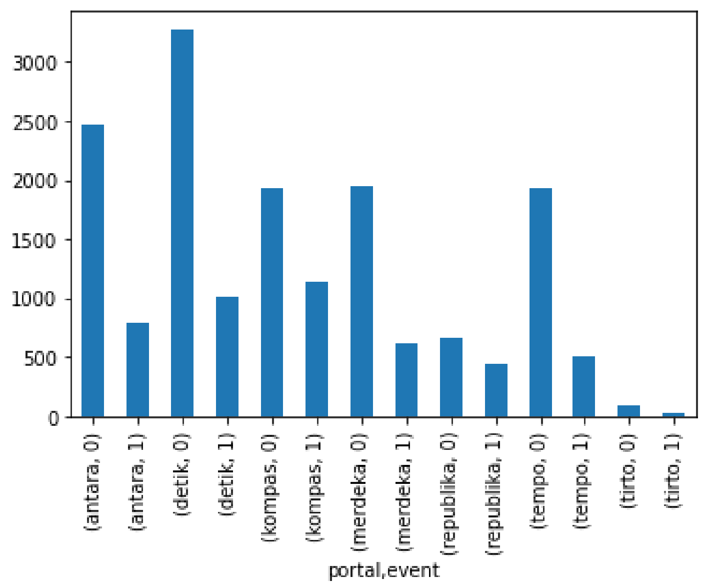

Fig. 1. Data distribution of online news titles of COVID-19.

A. Data Collection

Data were collected by web crawling several national online news medias (i.e., `antara.com`, `detik.com`, `kompas.com`, `merdeka.com`, `republika.co.id`, `tempo.co`, and `tirto.id`) and downloading information such as news titles, the news release date, and the name of the news portal found for further storage in the database. The COVID-19 dataset was taken from January 2020 to May 2020 using the keywords ”corona” and ”covid” [9]. There are 16,844 online news titles related to COVID-19 that will be used. The data have been labeled as event data, denoted by ’1’, and non-event data, denoted by ’0’. Event data is news titles that contain reports of actual COVID-19 events, while non-event data, is news titles that contain general information about COVID-19. The acquired data is an imbalanced dataset dominated by non-event data. The total data for COVID-19 related events and non-events are 4,549 and 12,295 (almost 1:3 in ratio). The composition details of the dataset used are shown in Fig. 1.

B. Preprocessing Data

The data that had been collected then enters the data preprocessing stage. Preprocessing data is very important to select the most suitable keywords and remove unimportant words that do not provide additional information to distinguish documents. Data preprocessing consists of several steps, namely case folding [24], deletion of punctuation marks and symbols [24], tokenization [25], stopword/filtering [25], and vectorization [14].

By applying Term Frequency-Inverse Document Frequency (TF-IDF), a word can be quantified in terms of how conse-

Authorized licensed use limited to: Badan Riset Dan Inovasi Nasional. Downloaded on April 05,2023 at 04:41:22 UTC from IEEE Xplore. Restrictions apply.

TABLE I
DATA COMPOSITION OF COVID-19 ONLINE NEWS TITLE

Sampling Size	Total Data	
	Event	Non-Event
37%	4.549	12.295
30%	3.688	12.295
20%	2.459	12.295
10%	1.229	12.295
1%	123	12.295

quential it is in a document by giving the most prevalent words a lower weight and increasing the weight of the rare ones [15].

We applied this TF-IDF to train Naive Bayes, Logistic Regression, Decision Tree, SVM, AdaBoost, Neural Network, MLP, CNN, and LSTM models. In addition, we utilize word embedding layers in our network architecture to vectorize text data. We utilize it to train and test MLP, CNN, and LSTM models.

C. Data Sampling

The imbalance ratio of event data and non-event data is not high. The ratio of event and non-event data, respectively, for data regarding COVID-19, is nearly 1:3. To simulate a higher ratio of imbalance data, undersampling is done towards the event data. We took samples from event data as much as 30%, 20%, 10%, and 1% in comparison to the non-event data. The composition details of the sample of each imbalance ratios is listed in Table I.

We sampled the data using a simple random sampling. Simple random sampling, the simplest sampling method, selects at most n different units from the N units in the population such that any combination of the n units has an equal probability of being selected as a sample [26].

The first step in simple random sampling is to assign a number from 1 to N to each member of the population. The second step is to take n samples of these numbers using a random number table, computer, or calculator. The next step is to verify that the numbers obtained are different. The final step is to assign population elements that match these numbers as samples [27].

D. Classifier Model

We detected events of COVID-19 using some classifier models from conventional machine learning and deep learning. There are Naive Bayes, Logistic Regression, Decision Tree, Support Vector Machine (SVM), AdaBoost, and Neural Network from traditional machine learning. We used Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), and Long-Short Term Memory (LSTM) from deep learning.

Naive Bayes, Logistic Regression, Decision Tree, SVM, AdaBoost, and NN models were created using Scikit-learn library. As for, MLP, CNN, and LSTM were created using Keras Tensor Flow library.

We use the same architecture as in [14] for MLP, CNN, and LSTM, which consists of `ReLU` activation function in a hidden layer or fully connected layer, the `sigmoid` activation function in the output layer, and the backpropagation technique with the `ADAM` optimizer.

IV. RESULTS AND DISCUSSION

In this study, we evaluated model performance using conventional machine learning and deep learning techniques. In general, each model was trained and tested using data sampling as shown in Table I. In particular,

we used TF-IDF in the network architecture in all the models we used in our study. Further, we used the word embedding layer in deep learning model network architectures, namely MLP, CNN, and LSTM. These model were evaluated using k-fold Cross-Validation so that the trained data and the tested data would have the same percentage for each iteration. We used k = 5 instead of 10 to evade small amounts of training and testing data. For example, the distribution of the data is 80:20. The ratio implies the division of training data (80%) and testing data (20%). In addition, we used early stopping based on the loss value.

We used four performance measures to indicate the performance of each classifier:

- Accuracy. Accuracy represents how accurate the model in classifying correctly.
  Accuracy = (TP + TN) / (TP + FP + FN + TN)
- Precision. Precision shows the accuracy level number between the information being seeked and the answer given by the model.
  Precision = (TP) / (TP+FP)
- Recall. Recall is the success rate of a model in retrieving information.
  Recall = (TP) / (TP + FN)
- F1 Score. F1 score is the harmonic average comparison of precision and recall.
  F1 Score = 2 * (Recall * Precision) / (Recall + Precision)

Based on Fig. 2, CNN + Emb. layer is the method with the highest accuracy at a sample size of 37%, with the number of event data being 4,549. Other methods accuracy increase as the sample size and the amount of event data get smaller. All of the methods show high accuracy, but there is a possibility that all of these methods are misclassified and therefore do not perform well. This phenomena happens because the class imbalance is very high, which results in all methods not studying the data distribution correctly. Thus, all of them only read non-event data in training and testing and considered very little event data as noise. Therefore, in higher level of imbalance ratio, accuracy may not be suitable for measuring the classifier's performance.

The next performance measure that we will look at is precision. Precision results depend on False Positive, which is non-event data classified as event data. The more non-event data is classified as event data, the smaller the precision value will be.

Authorized licensed use limited to: Badan Riset Dan Inovasi Nasional. Downloaded on April 05,2023 at 04:41:22 UTC from IEEE Xplore. Restrictions apply.

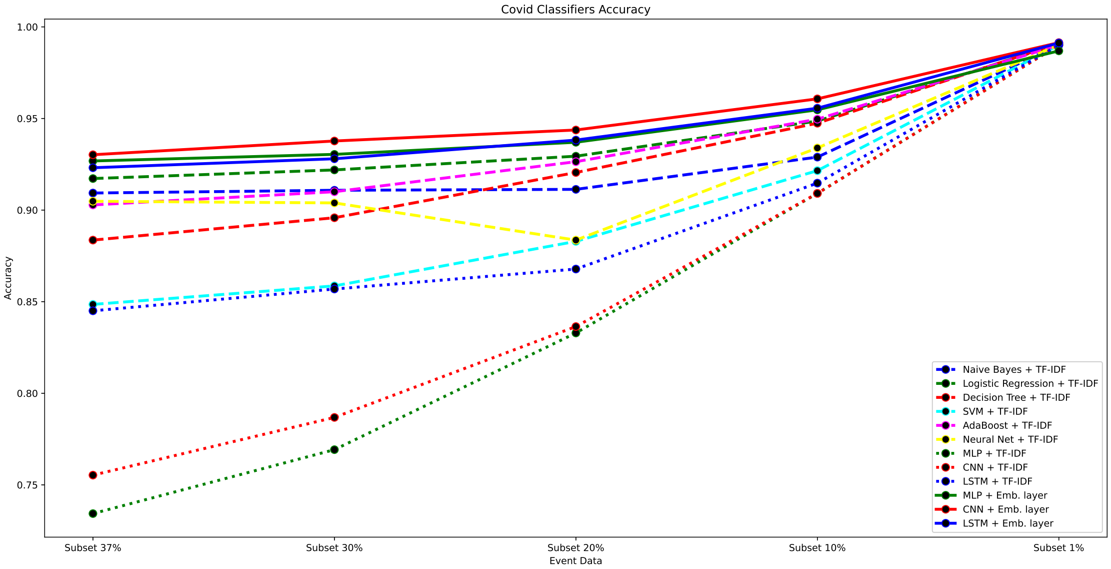

Fig. 2. Classifier Accuracy Comparison.

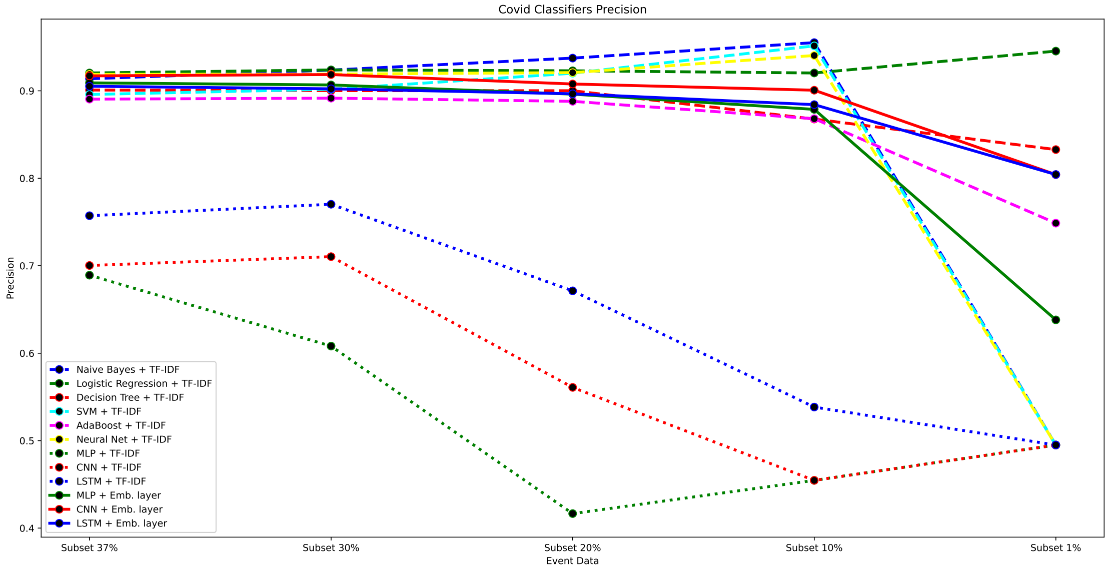

Fig. 3. Classifier Precision Comparison.

Based on Fig. 3, Naive Bayes + TF-IDF, SVM + TF-IDF, Neural Network + TF-IDF, Decision Tree + TF-IDF, AdaBoost + TF-IDF, MLP + Emb. layer, CNN + Emb. layer, and LSTM + Emb. layer have the lowest level of precision when the event data is very imbalanced at a sample size of 1%, with the number of event data being 123. These classifiers are starting to improve, showing good performance in predicting event data at a sample size of 10%, where the number of event data is 1,229, and so on. Logistic Regression + TF-IDF is the method with the highest precision on sample sizes of 37%, 30%, and 1%, with the number of event data being 4,549, 3,688, and 123.

The following performance measure that we examine is recall. The recall value depends on False Negative, which is

Authorized licensed use limited to: Badan Riset Dan Inovasi Nasional. Downloaded on April 05,2023 at 04:41:22 UTC from IEEE Xplore. Restrictions apply.

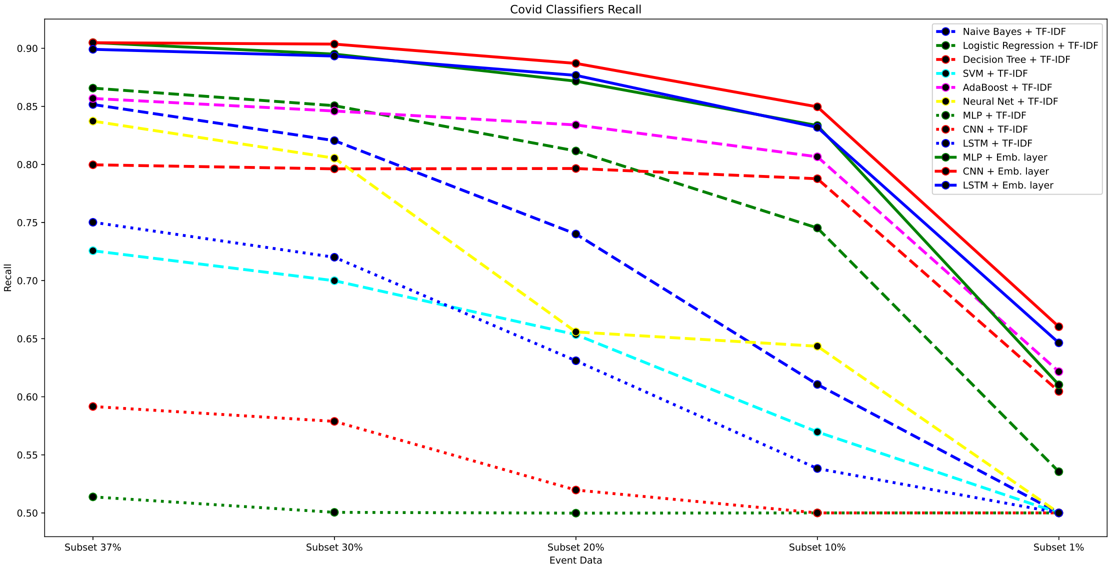

Fig. 4. Classifier Recall Comparison.

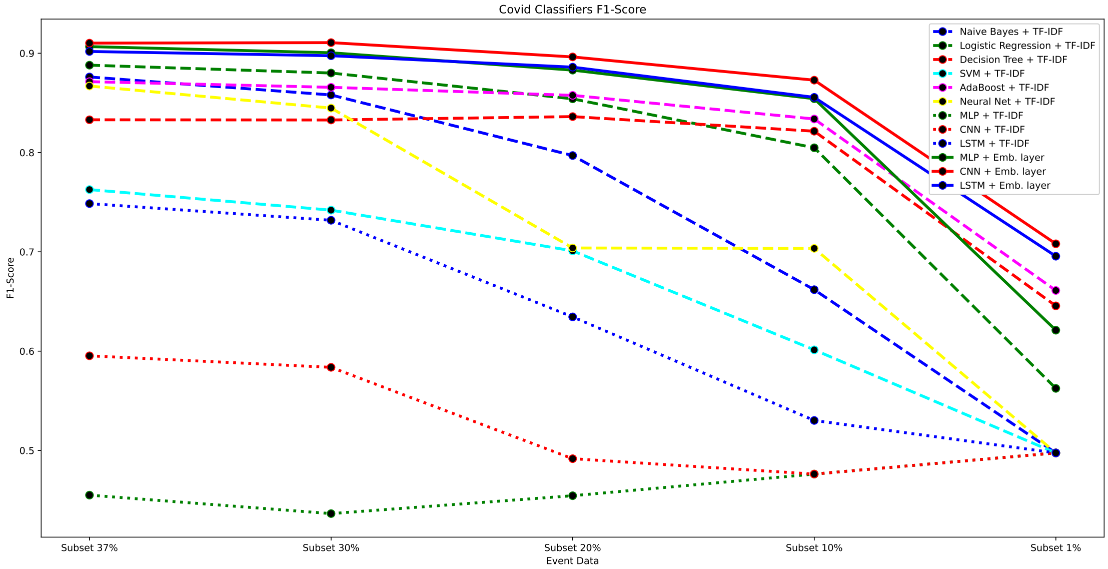

Fig. 5. Classifier F1 Score Comparison.

event data that is classified as non-event data. The more the event data is classified as non-event data, the smaller the recall value will be.

Based on Fig. 4, all classifiers have the lowest recall rate when the event data is very imbalanced at a sample size of 1%, with the number of event data being 123. These methods began to improve except Decision Tree + TF-IDf, showing good performance in predicting event data at a sample size of 10%, where the number of event data is 1,229, and so on. CNN + Emb. layer is the method with the highest recall in all sample sizes and the number of event data.

The succeeding performance measure is the F1 Score. Based on Fig. 5, all classifiers except MLP + TF-IDF and CNN + TF-IDF have the lowest F1 Score level when the event data

Authorized licensed use limited to: Badan Riset Dan Inovasi Nasional. Downloaded on April 05,2023 at 04:41:22 UTC from IEEE Xplore. Restrictions apply.

is very imbalanced at a sample size of 1%, with the number of event data being 123. These classifiers began to improve, except for Decision Tree + TF-IDf, MLP + TF-IDF, and CNN + TF-IDF, showing good performance in predicting event data at a sample size of 10%, where the number of event data is 1,229, and so on. CNN + Emb. layer is the method with the highest F1 Score across all sample sizes and the number of event data.

Compared to other methods, Logistic Regression + TF-IDF achieves higher scores for precision levels, and CNN + Emb. layer achieves higher scores for three performance measures, namely accuracy, recall, and F1 Score on average. Recall performance needs to be considered in imbalance dataset because the recall results depend on how much the minority class (event data) is predicted to be the majority class (non-event data). Because the imbalance ratio between event data and non-event data is very high, it will be a problem if very little of the event data is predicted to be non-event data. In addition to accuracy, the CNN + Emb. layer showed the best performance in recall. F1 Score is usually used to assess method performance on imbalanced data. We can see that the CNN + Emb. layer performs better than the other methods in the F1 Score.

The results of our study show the superiority of the CNN + Emb. layer method for our online news titles dataset when the imbalance ratios are 37%, 30%, 20%, and 10% compared to other models, even though the CNN + Emb. layer shows a marked decrease in performance at an imbalanced ratio of 1%.

V. CONCLUSION

This paper compares the performance of classifiers in detecting COVID-19 events in online news titles using conventional machine learning and deep learning models. The results of our study show that the CNN + Emb. layer method outperforms other models for our online news titles dataset when the imbalance ratios are 37%, 30%, 20%, and 10%, even though the CNN + Emb. layer shows a significant decrease in performance when the imbalance ratio is 1%.

As far as we know, slight research has been done to identify events in Indonesian online news titles. In future research, we devise to improve our research and apply the CNN model with parameter tuning and other techniques to improve performance across various datasets with 1% to extreme imbalance ratios, such as 0.1% and 0.01%.

REFERENCES

[1] M. Torki, M. Ibrahim, and N. El-Makky, "Imbalanced toxic comments classification using data augmentation and deep learning," 12 2018.

[2] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning–based text classification: a comprehensive review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–40, 2021.

[3] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[4] A. L'heureux, K. Grolinger, H. F. Elyamany, and M. A. Capretz, "Machine learning with big data: Challenges and approaches," *Ieee Access*, vol. 5, pp. 7776–7797, 2017.

[5] H. Liu, M. Zhou, and Q. Liu, "An embedded feature selection method for imbalanced data classification," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 3, pp. 703–715, 2019.

[6] K. RI, "Keputusan menteri kesehatan republik indonesia nomor hk. 01.07/menkes/413/2020 tentang pedoman pencegahan dan pengendalian corona virus disease 2019 (covid-19)," *Menteri Kesehatan Republik Indonesia*, 2020.

[7] K. Fadli, P. Novita *et al.*, "Analisis framing media online tentang pandemi covid-19 (studi kasus covid-19 pada media online tribun news. com dan kepri. co. id edisi bulan maret s/d juni 2020)," *JURNAL PURNAMA BERAZAM*, vol. 2, no. 2, pp. 172–200, 2021.

[8] M. D. R. Wahyudi, A. Fatwanto, U. Kiftiyani, and M. G. Wonoseto, "Topic modeling of online media news titles during covid-19 emergency response in indonesia using the latent dirichlet allocation (lda) algorithm," *Telematika*, vol. 14, no. 2, pp. 101–111, 2021.

[9] E. Nugraheni, P. H. Khotimah, A. Arisal, A. F. Rozie, D. Riswantini, and A. Purwarianti, "Classifying aggravation status of covid-19 event from short-text using cnn," in *2020 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*. IEEE, 2020, pp. 240–245.

[10] L. Launa, "Banjir infodemi: Viralitas akurasi berita virologi dalam fenomena coronavirus disease," *The Source: Jurnal Ilmu Komunikasi*, vol. 2, no. 2, pp. 1–21, 2020.

[11] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[12] M. Ramdhani, D. Maylawati, and T. Mantoro, "Indonesian news classification using convolutional neural network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, p. 1000, 08 2020.

[13] S. Shaikh, S. M. Daudpota, A. S. Imran, and Z. Kastrati, "Towards improved classification accuracy on highly imbalanced text dataset using deep neural language models," *Applied Sciences*, vol. 11, no. 2, p. 869, 2021.

[14] P. H. Khotimah, A. F. Rozie, E. Nugraheni, A. Arisal, W. Suwarningsih, and A. Purwarianti, "Deep learning for dengue fever event detection using online news," in *2020 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*. IEEE, 2020, pp. 261–266.

[15] C. Padurariu and M. E. Breaban, "Dealing with data imbalance in text classification," *Procedia Computer Science*, vol. 159, pp. 736–745, 2019.

[16] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017.

[17] S. B. S. Lai, N. H. N. B. M. Shahri, M. B. Mohamad, H. A. B. A. Rahman, and A. B. Rambli, "Comparing the performance of adaboost, xgboost, and logistic regression for imbalanced data," 2021.

[18] N. A. Hasanah, N. Suciati, and D. Purwitasari, "Identifying degree-of-concern on covid-19 topics with text classification of twitters," *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 7, no. 1, pp. 50–62, 2021.

[19] P. Kapil, A. Ekbal, and D. Das, "Investigating deep learning approaches for hate speech detection in social media," *arXiv preprint arXiv:2005.14690*, 2020.

[20] Q. Wang, W. Li, and Z. Jin, "Review of text classification in deep learning," *Open Access Library Journal*, vol. 8, no. 3, pp. 1–8, 2021.

[21] P. K. Yechuri and S. Ramadass, "Classification of image and text data using deep learning-based lstm model." *Traitement du Signal*, vol. 38, no. 6, 2021.

[22] N. Sun and C. Du, "News text classification method and simulation based on the hybrid deep learning model," *Complexity*, vol. 2021, 2021.

[23] M. A. Ramdhani, D. S. Maylawati, and T. Mantoro, "Indonesian news classification using convolutional neural network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 2, pp. 1000–1009, 2020.

[24] J. C. Young, A. Rusli *et al.*, "A comparison of supervised text classification and resampling techniques for user feedback in bahasa indonesia," in *2020 Fifth International Conference on Informatics and Computing (ICIC)*. IEEE, 2020, pp. 1–6.

[25] D. Parolo, "Deep learning for text classification: an application of generalized language models for italian texts," 2020.

[26] S. K. Thompson, "Sampling/by steven k. thompson." Tech. Rep.

[27] P. S. Levy and S. Lemeshow, "The population and the sample," *Sampling of Populations: Methods and applications. 4th ed. New York, USA: John Wiley and Sons*, pp. 11–42, 2008.

Authorized licensed use limited to: Badan Riset Dan Inovasi Nasional. Downloaded on April 05,2023 at 04:41:22 UTC from IEEE Xplore. Restrictions apply.