JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 5, No. 3 June 2023 DOI: https://doi.org/10.34288/jri.v5i3.555
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

ANALYSIS OF INDONESIAN LANGUAGE DATASET FOR TAX COURT CASES: MULTICLASS CLASSIFICATION OF COURT VERDICTS

Ade Putera Kemala[-1*], Hafizh Ash Shiddiqi[-2]

Data Science[1*], Computer Science[2]
School of Computer Science, Bina Nusantara University
Jakarta, Indonesia
ade.kemala@binus.ac.id[1*)], hafizh.shiddiqi@binus.ac.id[2)]
(*)Coresponding Author

Abstrak

Pajak adalah kewajiban yang timbul akibat adanya undang-undang, menciptakan kewajiban bagi warga negara untuk memberikan sebagian pendapatan mereka kepada negara. Pengadilan Pajak berperan sebagai otoritas peradilan bagi wajib pajak yang mencari keadilan dalam sengketa pajak. Penelitian ini menyajikan analisis dataset pengadilan pajak dalam bahasa Indonesia dengan tujuan melakukan klasifikasi multiclass *untuk memprediksi putusan pengadilan. Sebelum digunakan dataset melalui tahap pra-pemrosesan untuk membersihkan data, proses augmentasi data menggunakan metode oversampling dan label weighting untuk mengatasi ketidakseimbangan kelas. Dua model, yaitu bi-LSTM dan IndoBERT, digunakan untuk melaksanakan proses klasifikasi. Penelitian ini menghasilkan model akhir dengan akurasi 75,83% menggunakan model IndoBERT. Hasil penelitian menunjukkan efektivitas kedua model dalam memprediksi putusan pengadilan. Penelitian ini memiliki implikasi dalam memprediksi kesimpulan pengadilan dengan informasi kasus yang terbatas, dan memberikan wawasan berharga untuk proses pengambilan keputusan hukum. Temuan ini berkontribusi pada bidang analisis data hukum, menampilkan potensi teknik NLP dalam memahami dan memprediksi hasil pengadilan, sehingga meningkatkan efisiensi proses hukum.*

Kata kunci: NLP; Tax; BERT; Deep learning; Klasifikasi

Abstract

Tax is an obligation that arises due to the existence of laws, creating a duty for citizens to contribute a certain portion of their income to the state. The Tax Court serves as a judicial authority for taxpayers seeking justice in tax disputes, handling various types of taxes on a daily basis. This paper presents an analysis of an Indonesian language dataset of tax court cases, aiming to perform multiclass classification to predict court verdicts. The dataset undergoes preprocessing steps, while data augmentation using oversampling and label weighting techniques address class imbalance. Two models, bi-LSTM and IndoBERT, are utilized for classification. The research produced a final result of model with 75.83% using IndoBERT model. The results demonstrate the efficacy of both models in predicting court verdicts. This research has implications for predicting court conclusions with limited case details, providing valuable insights for legal decision-making processes. The findings contribute to the field of legal data analysis, showcasing the potential of NLP techniques in understanding and predicting court outcomes, thus enhancing the efficiency of legal proceedings.

Keywords: *NLP; Tax; BERT; Deep learning; Classification*

INTRODUCTION

Taxation, from an economic standpoint, refers to the transfer of resources from the private sector to the public sector. From a legal perspective, taxation is an obligation that arises due to the existence of laws, creating a duty for citizens to contribute a certain portion of their income to the state (Sutedi, 2022). Tax is a compulsory contribution imposed by the government on taxpayers, whether individuals or corporations, and it is enforced based on legal provisions. The government does not provide direct compensation to taxpayers; however, tax revenue should be utilized for the maximum prosperity of the people and the needs of the state (Halim et al., 2014).

Taxation in Indonesia is primarily regulated in the constitution through Article 23A of the *Undang Undang Dasar* (UUD) *Tahun* 1945 (Pracasya, 2021) "Taxes and other compulsory

levies for the purposes of the state are regulated by law". Based on the collecting institutions, taxes in Indonesia are levied by both central and regional institutions. The types of taxes managed by the central government include Income Tax (PPh), Value-Added Tax (PPN), Luxury Goods Sales Tax (PPnBM), Stamp Duty (*Bea Materai*), Land and Building Tax for Plantations, Forestry, and Mining. These taxes are mostly managed by the Directorate General of Taxes and the Ministry of Finance (Farouq, 2018).

Taxes play a crucial role in the life of a nation, particularly in the implementation of development, as taxes serve as a source of state revenue to finance expenditures, including development expenses. The general functions of taxes are budgetary, regulatory, stability, and income redistribution functions (Rohendi, 2014).

The Tax Court is a judicial body that exercises judicial authority over taxpayers or tax payers seeking justice in tax disputes they encounter. The Tax Court has the same status, rank, and independence as other equivalent courts. This Tax Court operates within the framework of state administration and has an organizational structure that ultimately reports to the Mahkamah Agung (Supreme Court) (Sandra, 2021).

NLP (Natural Language Processing) research has been flourishing in recent years due to the advancements in context-based models in NLP research, revolutionalized by the publication of the Transformer model (Vaswani et al., 2017). One of the developments stemming from the Transformer model is the BERT (Bidirectional Encoder Representations from Transformers) model (Devlin et al., 2018), which is an architecture created by stacking the Encoder component of the Transformer model. Previously, the popular method involved using word vectors like word2vec (Church, 2017) or glove(Pennington et al., 2014) combined with deep neural networks such as LSTM (Yu et al., 2019).

Using BERT architecture, it is possible to create a pre-trained model that is trained with a huge amount of unlabeled data to provide it with a general understanding of language. Subsequently, the pre-trained model can be further fine-tuned using a small amount of labeled data to adapt it to a specific task. The BERT model has achieved remarkable results in various NLP tasks, such as classification (Sun et al., 2019), question answering (Wang et al., 2019), and named entity recognition (Church et al., 2020).

One of the implementation of the BERT architecture, pre-trained on the Indonesian language, is referred to as IndoBERT (Wilie et al., 2020). This model is trained with dataset called Indo4B which is a 23GB collection of corpus dataset, including Wikipedia, twitter, newsletter data. The resulting model has achieved state-of-the-art results in several NLP tasks specific to Indonesian language.

The aim of this paper is to analyze the dataset of legal cases from the Indonesian Tax Court and attempt to predict the court verdict using available data by leveraging the capabilities of the BERT model for multiclass classification tasks.

Due to the limited availability of textual dataset in the Indonesian language (Ferdiana et al., 2019), there is still a gap that needs to be addressed in Indonesian natural language processing research. Therefore, this research aims to explore and utilize pre-trained models to build a classifier model for analyzing and classifying court-based data.

RESEARCH METHODS

This quantitative research focuses on analyzing and performing a multiclass classification task to predict court verdicts based on provided Indonesian language text data.

Most of the research was conducted in the virtual space, utilizing cloud services provided by Google for the required computational tasks. The execution timeframe for this research was from June 1, 2023, to July 7, 2023.

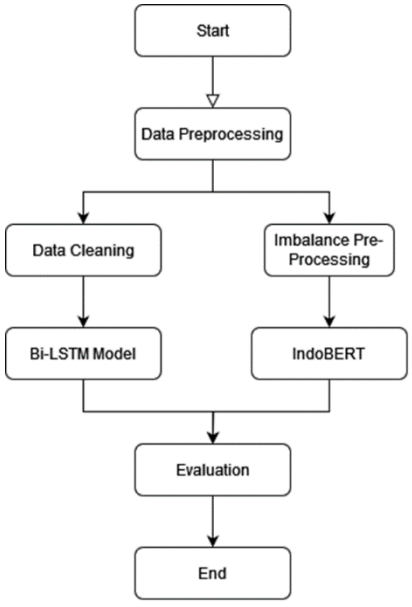

Figure 1. Flowchart research

Figure 1 shows the steps involved in this research, starting with the dataset processing and utilizing the data to develop two NLP models. Each

model has different mechanisms, resulting in different treatments for the training dataset. Once the training process is completed, the accuracy of the models is evaluated.

Dataset

The dataset used is the Indonesian Tax Court Verdict Summary, which is a secondary dataset and is open source, obtained through Kaggle (Christian, 2021). This dataset consists of 12,283 text data entries describing the content of tax disputes in the tax court along with the court decisions. Each data row represents a case in the tax court. The dataset consists of 7 columns. The detailed explanations for each column can be found in Table 1.

Table 1. Dataset Features

No	Column	Desc
1	Text	Textual data from court documents
2	*Nomor_putusan*	Court decision number
3	*Tahun_pajak*	Tax year
4	*Jenis_pajak*	Tax type
5	*Tahun_putusan*	Year of court decision
6	*Pokok_sengketa*	Main dispute
7	*Jenis_putusan*	Type of verdicts

The dataset contains tax disputes handled by the tax court from 2005 to 2020. Among these 7 columns, the *jenis_pajak* (tax type) and *pokok_sengketa* (main dispute) columns will be used as features for analysis, while the *jenis_putusan* (verdict type) column will be used as the label column. In this stage, we will attempt a multiclass classification task using the dataset. We will analyze the dataset, clean it, and determine which parameters to use. Based on preliminary inspection, we are particularly interested in using the *pokok_sengketa* (main dispute) as the primary feature for determining the verdict. This column contains the textual main object of discussion regarding the dispute related to this legal case.

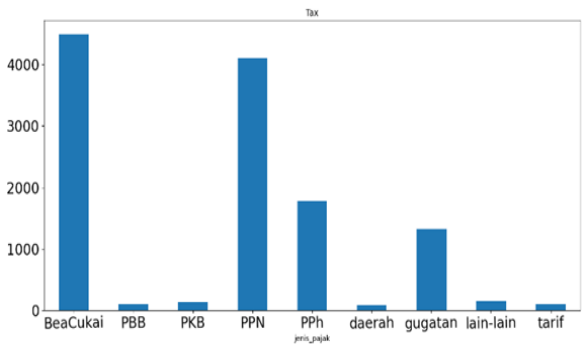

Figure 2. Tax Type Distribution

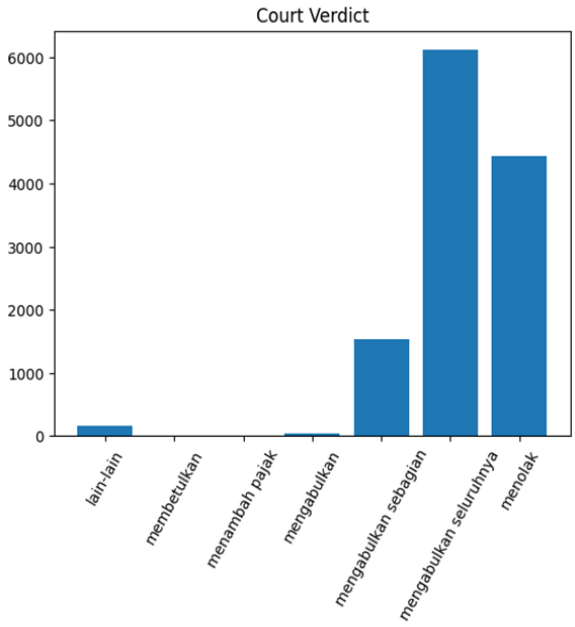

Figure 3. Court Verdict Distribution

Figures 2 and 3 display the statistical data distribution based on the *jenis_putusan* (decision type) and *jenis_pajak* (tax type) columns. It is evident that based on the *jenis_putusan* column, there are three types of verdicts with higher frequencies which is : *mengabulkan seluruhnya* (approved completely), *mengabulkan sebagian* (partially approved), and *menolak* (rejected). For convenience, *menolak* will be referred to as label '0', *mengabulkan sebagian* as label '1', and *mengabulkan seluruhnya* as label '2'.

In Figure 2, it is shown that there are four major types of tax cases in the dataset: BeaCukai (customs), PPN (Value-added Tax), PPh (Income Tax), and gugatan (Tax Lawsuit). Based on this preliminary examination, we will only utilize these four types of cases that have verdicts falling into the three most common categories in the dataset.

Table 2 displays the final data to be used for training and testing. There is a significant imbalance in the dataset labels, and various approaches will be employed to address this issue later.

Table 2. Data Distribution

Data	Total Count	Class Distribution (0 : 1 : 2)
All Data	11380	4011 : 1464 : 5905
BeaCukai	4371	1340 : 265 : 2766
PPN	3992	1201 : 773 : 2018
PPh	1730	573 : 404 : 753
Gugatan	1287	897 : 22 : 368

Models

There are two models used in this research. The first model is the bi-LSTM model. For this algorithm, the dataset is further pre-processed by removing stopwords and punctuation. Afterward, the dataset is split into an 80:20 distribution as training data and testing data. The resulting training data is fed into a bi-LSTM network to train the model.

To ensure the best result, we employ a Gridsearch method (Akiba et al., 2019) to find the optimal hyperparameters for the model. The search space consists of the embedding dimension ranging from 50 to 300, LSTM units ranging from 64 to 256, and dropout rates ranging from 0 to 0.5. The best hyperparameters generated by this method and the default initial values to train the model can be observed in Table 3. The resulting models then evaluated with testing data to find the accuracy of the model.

Table 3. Hyperparameter Searching

Hyperparameter	Range	Best Value	Default Value
Embedding dimension	50 – 300	271	100
LSTM unit	64 – 256	75	128
Dropout	0 – 0.5	0.11	0.5

The second model used in this research is called IndoBERT, which is an implementation of the BERT architecture specifically for the Indonesian language. For this model, pre-processing is not required since the model focuses on the context of the sentences as a whole. Deleting or modifying the sentences may remove or alter the context.

The data split used remains the same, with an 80:20 ratio for the train data and test data. The model is trained with the following hyperparameter settings: maximum input length of 512, batch size of 16, epoch of 10, and a learning rate of 5e-06. Several scenarios were tested with this model, and further details of these scenarios will be explained in the following section.

Imbalance Dataset

To address data imbalance, two methods will be employed namely : label weighting and oversampling. Label weighting is a technique used to address the issue of imbalanced datasets in machine learning. It involves assigning different weights to the labels or classes in the dataset based on their frequency or importance (Madabushi et al., 2020). With label weighting, each label will be assigned a different weight based on its frequency distribution. Label weighting helps in giving more importance to the minority classes

On the other hand, for the oversampling method, the EDA (Easy Data Augmentation) technique (Wei & Zou, 2019) combined with WordNet will be used to generate new synthetic data using similar words from the WordNet corpus. It is worth noted that only the train data will be oversampled to prevent data testing leakage.

There are several steps implemented by the EDA method to perform data augmentation, including replacing words with their synonyms, deleting a percentage of words, rearranging the positions of words in a sentence, and randomly inserting words into sentences. These steps are randomly applied to each data row. This augmentation will result in a more balanced dataset, where all classes with fewer instances will be augmented to match the size of the majority class. The results of the oversampling technique are presented in Table 4.

Table 4. Data Distribution after Oversampling

Data	Class Distribution (0 : 1 : 2)
BeaCukai	2166 : 2180 : 2195
PPN	1950 : 1833 : 1607
PPh	924 : 630 : 607
Gugatan	731 : 630 : 846

RESULTS AND DISCUSSION

The overall results of the research are presented in Table 5. Several insights can be derived from these results.

Table 5. Research Result

Model	Data		Acc
IndoBERT	All data	Normal	**75.83%**
		Weighted label	75.04%
	BeaCukai Data	Normal	**79.66%**
		Weighted label	78.29%
		Oversampling	65.49%
	PPN Data	Normal	**80.48%**
		Weighted label	79.50%
		Oversampling	51.81%
	PPh Data	Normal	**66.48%**
		Weighted label	65.61%
		Oversampling	43.06%
	Gugatan Data	Normal	**86.05%**
		Weighted label	81.78%
		Oversampling	64.34%
Bi-LSTM	All data	Normal Parameter	67.36%
		Best Parameter	**67.57%**

The first insight observed from the results is that context-based models like BERT outperformed word vector-based models with deep

neural networks such as bi-LSTM. Even with the best parameters applied to the bi-LSTM model, there was no significant increase in performance compared by using the initial hyperparameter in bi-LSTM model.

For the IndoBERT model, several scenarios were tested. The first scenario involved fine-tuning the model using all available data. There was no significant difference between using normal distribution data or weighted distribution data; both scenarios achieved an accuracy of 75%.

After observing the performance using all the data, it was decided to attempt dividing the data by tax types. The rationale behind this is that different types of taxes may involve distinct wording and contextual elements in each legal case. The results indicated that the fine-tuning process using BeaCukai and PPN data yielded better results compared to the previous approach. It is worth noting that among all tax types, PPh (Income Tax) proved to be the most challenging for the model to classify accurately.

Regarding the gugatan (Tax Lawsuit) data, its comparatively higher performance may be attributed to significant label imbalances within the testing data. There were only sufficient data for two labels, while the last label had very little representation. As a result, the model primarily focused on classifying the two main labels, which was relatively easier than classifying all three labels, leading to a more significant improvement in performance.

It is also observed that using the weighted label and oversampling methods does not contribute to an improvement in the models' performance. In the case of oversampling, where several words in a sentence may be altered, removed, or added, it can potentially change the contextual meaning of the text, making it more challenging for the BERT model to accurately analyze the true context of the data.

CONCLUSIONS AND SUGGESTIONS

After analyzing the data, performing the necessary preprocessing data for each model it is concluded that context-based model like BERT performed best on a multiclass classification task. Interestingly, neither label weighting and oversampling method yielded a better result in this particular case. This research has implications for predicting court conclusions with limited case details, providing valuable insights for legal decision-making processes. The findings contribute to the field of legal data analysis, showcasing the potential of NLP techniques in understanding and predicting court outcomes, thus enhancing the efficiency of legal proceedings. Suggestion for future research is to use 'text' column in dataset as a new features and extract relevance information from it in order to perform multiclass classification with a better result.

REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631.

Christian, W. (2021). *Indonesian Tax Court Verdict Summary*. https://www.kaggle.com/datasets/christianwbsn/indonesia-tax-court-verdict

Church, K. W. (2017). Emerging Trends: Word2Vec. *Natural Language Engineering*, *23*(1), 155–162. https://doi.org/10.1017/S1351324916000034

Church, K. W., Luoma, J., & Pyysalo, S. (2020). Exploring cross-sentence contexts for named entity recognition with BERT. *ArXiv Preprint ArXiv:2006.01563*, *23*(1), 155–162.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*.

Farouq, M. (2018). *Hukum pajak di Indonesia*. Prenada Media.

Ferdiana, R., Jatmiko, F., Purwanti, D. D., Ayu, A. S. T., & Dicka, W. F. (2019). Dataset Indonesia untuk Analisis Sentimen. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *8*(4), 334–339.

Halim, A., Bawono, I. R., & Dara, A. (2014). Perpajakan: Konsep, Aplikasi, Contoh, dan Studi Kasus. *Jakarta: Salemba Empat*.

Madabushi, H. T., Kochkina, E., & Castelle, M. (2020). Cost-sensitive BERT for generalisable sentence classification with imbalanced data. *ArXiv Preprint ArXiv:2003.11563*.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Pracasya, D. P. (2021). Penerapan Peraturan Perundang-Undangan Pajak Daerah Atas

Perubahan Pasal Mengenai Perpajakan Dalam Undang-Undang Dasar Republik Indonesia Tahun 1945. *" Dharmasisya" Jurnal Program Magister Hukum FHUI*, *1*(2), 13.

Rohendi, A. (2014). Fungsi budgeter dan fungsi regulasi dalam ketentuan perpajakan indonesia. *Jurnal Ecodemica: Jurnal Ekonomi, Manajemen, Dan Bisnis*, *2*(1), 119–126.

Sandra. (2021). *Mengenal Tugas dan Wewenang Pengadilan Pajak*. https://www.pajakku.com/read/60cc494e58d6727b1651ab0f/Mengenal-Tugas-dan-Wewenang-Pengadilan-Pajak

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification? *China National Conference on Chinese Computational Linguistics*, 194–206.

Sutedi, A. (2022). *Hukum pajak*. Sinar Grafika.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *ArXiv Preprint ArXiv:1706.03762*. https://arxiv.org/abs/1706.03762

Wang, Z., Ng, P., Ma, X., Nallapati, R., & Xiang, B. (2019). Multi-passage bert: A globally normalized bert model for open-domain question answering. *ArXiv Preprint ArXiv:1908.08167*.

Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *ArXiv Preprint ArXiv:1901.11196*.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., & Bahar, S. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. *ArXiv Preprint ArXiv:2009.05387*.

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation*, *31*(7), 1235–1270.