I. TITLE AND ABSTRACT

1.1 Title

BERTopic and RoBERTa as Models for Analyzing the Harmony of Legislation
About Investment in Indonesia

1.2 Abstract

The chaos of investment-related legal regulations in Indonesia is thought to be one of the reasons barriers to investor entry in Indonesia. Ego-sectoral between institutions and interests Diverse regions are also considered to be an obstacle to legal development in Indonesia. The imbalance of priorities and different perspectives between institutions and regions can makes it difficult to draft and implement laws and regulations that is consistent and effective. This condition is also influenced by the existence of ***hyper-regulation.*** This study attempts to examine the chaos of legal regulations. by using BERTopic and RoBERTa. Both models are used for answering the problem of alignment or harmony of laws and regulations related to investment in Indonesia by reviewing 110 related laws and regulations investment. The results of the study show that of the 110 laws and regulations that analyzed there were only 0.01% of verse pairs that were not in harmony. The findings of this study shows that by using BERTopic and RoBERTa rules Investment related legislation in Indonesia already supports the creation of an environment efficient investment. The results of this study can be an initial effort to analyze regulations other legislation in order to create legal consistency in Indonesia.

Keywords: BERTopic, Indonesia, Investment, Legislation, RoBERTa

II. INTRODUCTION

2.1 Background

The Indonesian government has an obligation to implement legal development nationally in a planned, integrated and sustainable manner, by maintaining the national legal system which guarantees the protection of the rights and obligations of all Indonesian people based on The 1945 Constitution of the Republic of Indonesia for the realization of Indonesia as a country based on law. The development of national law is carried out through formation of legislation and its instruments. Formation of legislation and its instruments legislation is the process of making legal regulations includes the stages of planning, preparation, compilation, formulation, discussion, ratification, promulgation, and dissemination of legal products. Meanwhile, Law Number 12 of 2011 and the update of Law Number 15 of 2019 states that Legislation is a written norm that contains legal norms that generally binding and formed or determined by state institutions or officials authorized through procedures set out in statutory regulations (Saputra, 2016).

The implementation of regional autonomy and the expansion of the authority of institutional agencies making the laws and regulations in Indonesia increasingly complex. Not a few there is overlap and inconsistency in legislation in Indonesia, especially when there is a transition of changes in the law. Ego-sectoral between institutions and regional interests which is diverse is also considered to be an obstacle to legal development in Indonesia. The imbalance of priorities and different perspectives between institutions and regions can makes it difficult to draft and implement laws and regulations consistent and effective.

Likewise with the laws and regulations regarding investment in Indonesia. Minister of Law and Human Rights, Yasonna H. Laoly, on The 2020 National Press Day commemoration states the need for legal and regulatory development which drives economic growth through investment (Public Relations, Legal, and Cooperation with the Ministry of Law and Human Rights, 2020). Regulatory reform in the field of business licensing, employment, development of MSMEs (Micro, Small and Medium Enterprises), which is supported with tax regulation reform it is very important to achieve the Work Program The President's priorities, which include human resource development and simplification of all forms of regulation. Regulatory reform is very important to do in order to resolve investment barriers, cut long bureaucratic chains, and to address overlapping and disharmonious regulations ***(hyper-regulation).***

In relation to the above problems, this study attempts to analyze harmonization of investment laws and regulations in Indonesia by implementing a model study of ***big data*** and ***artificial intelligence. Big data*** has the opportunity to analyze regulations legislation that is in a ***hyper-regulation position.*** In addition, ***artificial intelligence*** helps solve big data problems from a computational and analytical perspective. algorithm. One of the popular tools for solving big data problems is LLM ***(Large Language Model).***

2.2 Problem Formulation

Based on the background description, this study wants to see how the level of the alignment or harmony of laws and regulations in Indonesia regarding investment by raising the following issues.

1. How does the use of BERTopic affect effectiveness and efficiency? in the analysis of the alignment of investment-related laws and regulations?
2. How are the prediction results of the RoBERTa model regarding regulatory alignment? legislation related to investment?
3. How does the RoBERTa model perform and perform in analyzing alignment? laws and regulations related to investment?

2.3 Objectives

Study on the analysis of the harmonization of laws and regulations on investment in Indonesia using BERTopic and RoBERTa has the following objectives.

1. To find out how the use of BERTopic affects effectiveness and efficiency of the analysis of the alignment of related laws and regulations investment in Indonesia.
2. Analyze the prediction results of the RoBERTa model on regulatory alignment. legislation related to investment in Indonesia.
3. To determine the performance of the RoBERTa model in analyzing harmonization of investment-related laws and regulations in Indonesia.

2.4 Benefits

Study on the analysis of the harmonization of laws and regulations on investment in Indonesia using BERTopic and RoBERTa has the following benefits: following.

1. Prevent ambiguity regarding laws and regulations related to investment in Indonesia.
2. Provide input to the government in evaluating verses based on the results analysis of the harmony of each paragraph of the legislation so that it can strengthening policies that support sustainable investment growth and have a positive impact on the Indonesian economy and society.

3. Provide investment-related thought contributions by ensuring alignment legislation and minimizing legal uncertainty in Indonesia.

III. LITERATURE REVIEW

3.1 NLI *(Natural Language Inference)*

NLI ***(Natural Language Inference)*** is a method used to determine the relationship between the hypothesis and the given premises. NLI classifies hypotheses into three classes based on their premises, namely ***entailment, contradiction,*** and ***neutral.*** For To train and test an NLI model, a dataset containing premise and hypothesis pairs is required. hypothesis with the appropriate class label ***(entailment, contradiction,*** or ***neutral).*** Example The dataset generally used in the NLI method is SNLI ***(Stanford Natural Language Inference)*** (Stanford NLP Group, nd). Currently, there are several models that have been trained for NLI tasks such as EFL ***(Entailment as Few-Shot Learner)*** (Wang et al., 2021), RoBERTa, and ALBERT. With these capabilities, NLI can be utilized for identify the relationship between two statements.

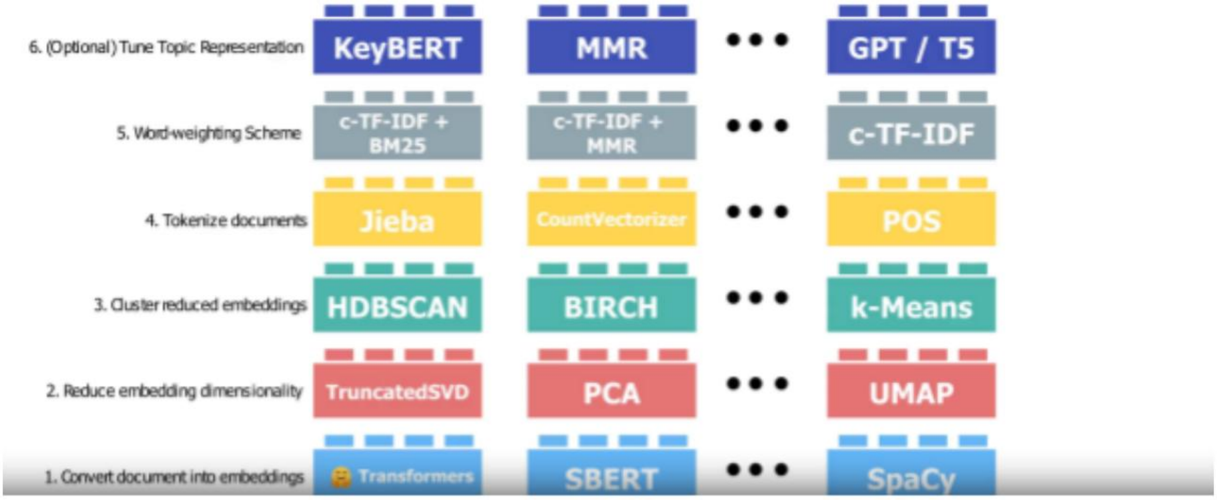
3.2 BERTopic *(Bidirectional Encoder Representations from Transformers Topic)*

BERTopic is a ***topic modeling*** method that uses a topic modeling approach. transformer -based , which primarily uses the BERT ***(Bidirectional)*** language model ***Encoder Representations from Transformers).*** This method was developed to find typical topics or patterns present in a collection of documents or texts without supervision. The BERTopic work process involves the following steps (Grootendorst, 2022).

1. The BERTopic process begins by generating a vector representation of the document. using pretrained ***transformer*** -based language models ;
2. After reducing the dimensions of the vector representation of the document, ***clustering*** to form groups of documents that have semantic similarities, each representing a different topic;
3. Finally, to overcome the view based on ***centroids,*** BERTopic developed a ***class-based*** version of the TF-IDF model to extract representations topics of each document group.

To carry out all these processes, BERTopic uses several submodels can be combined. Each submodel used has six uses, namely

embedding, dimensionality reduction, ***clustering,*** tokenization, ***word weighting,*** and optional models to perform ***tuning*** on the topic representation.



Gambar 1. Submodel BERTopic
(Sumber: https://maartengr.github.io/BERTopic/index.html)

3.3 RoBERTa ***(Robustly Optimized BERT Pre-training Approach)***

RoBERTa or ***Robustly Optimized BERT Approach*** is a language model that developed by FAIR (Facebook AI Research). This model is a variant of BERT ***(Bidirectional Encoder Representations from Transformers)*** which has undergone optimization through adjustments to techniques and procedures at the ***pre-training stage.*** According to Y. Liu et al. (2019), BERT is still not fully trained and can match or even outperforms the models released after it. The BERT model is a model with a ***Transformers*** -based architecture, the ***pre-training*** process of which is carried out by using two tasks, namely MLM ***(Masking Language Modeling)*** and NSP ***(Next Sentence) Prediction).*** The differences in the ***pre-training*** process and techniques carried out by the BERT model and RoBERTa is as follows (Liu et al., 2019).

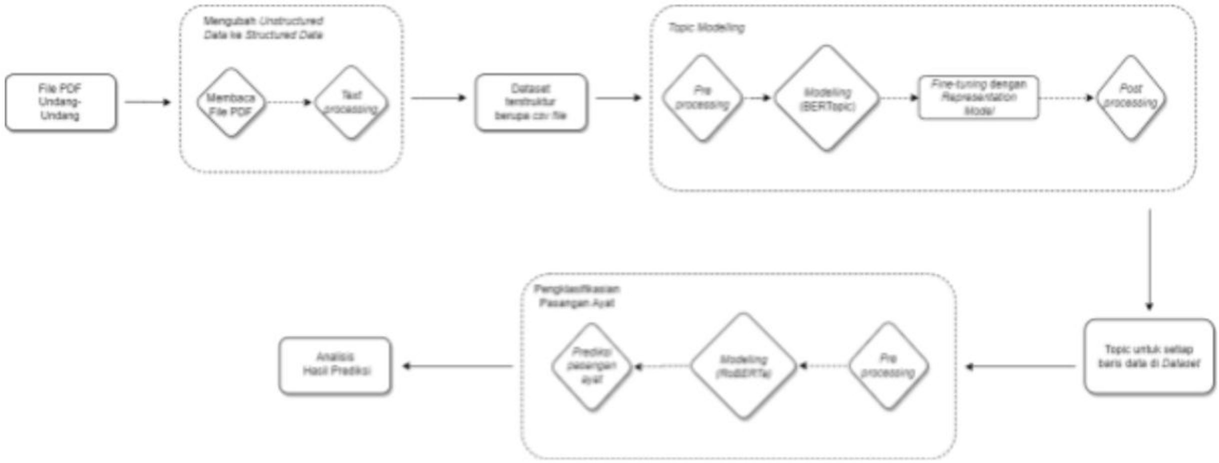
1. In the BERT model, the ***masking*** process is only carried out on ***the preprocessing*** data so that produces a static ***masking .*** This results in the same ***input*** will be repeatedly given to the model for each epoch. Meanwhile on RoBERTa model, the ***masking*** process is more dynamic where the ***training*** data will duplicated 10 times and ***masked*** in 10 different ways so that reduce the same ***input*** to the model.

2. In the BERT model, ***the input*** from the model is SEGMENT-PAIR + NSP (Next Sentence) Prediction) Loss. While in the RoBERTa model, ***the input*** of the model is changed become FULL-SENTENCES and do not use NSP loss.
3. The RoBERTa model is trained on more ***training*** data than BERT and trained on larger ***mini-batches*** than the BERT model.
4. Regarding ***the text encoding*** model BERT uses the ***character-level BPE*** implementation ***(Byte-Pair Encoding) vocabulary.*** Meanwhile, the RoBERTa model uses byte-level implementation of ***the BPE vocabulary.*** This makes the ***vocabulary*** size of the RoBERTa model becomes larger than the BERT model.

This optimization makes the RoBERTa model have better performance (approx. 2-20%) compared to the BERT model (Singh, 2021) and successfully obtained results ***state-of-the-art*** on three different benchmarks, namely GLUE, SQuaD, and RACE(Liu et al., 2019).

IV. METHODOLOGY

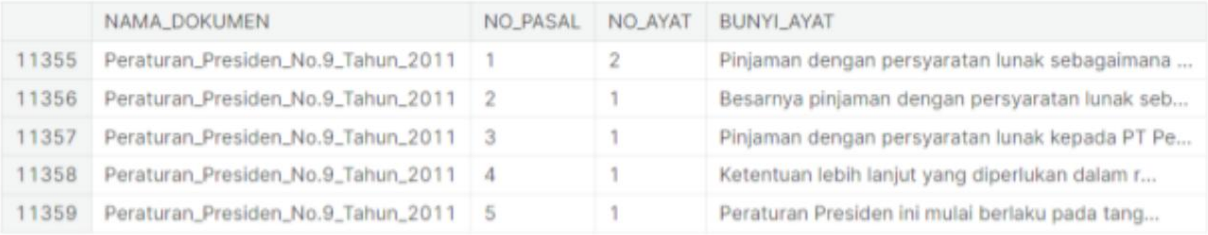
The following is a description of the problem solving flow/process carried out.



Gambar 2. Diagram alir

4.1 Converting *Unstructured Data* to *Structured Data*

First, each pdf file will be read and extracted using python ***library .*** PyMuPDF. The resulting ***string*** will go through a ***processing*** stage to extract each articles in the legislation and delete the parts that are not required. This ***text processing*** process is carried out by paying attention to the writing pattern on pdf file and assisted with ***regex*** in python. In addition, the deletion of characters is carried out ***newlines*** and excess spaces in each verse.



	NAMA_DOKUMEN	NO_PASAL	NO_AYAT	BUNYI_AYAT
11355	Peraturan_Presiden_No.9_Tahun_2011	1	2	Pinjaman dengan persyaratan lunak sebagaimana ...
11356	Peraturan_Presiden_No.9_Tahun_2011	2	1	Besarnya pinjaman dengan persyaratan lunak seb...
11357	Peraturan_Presiden_No.9_Tahun_2011	3	1	Pinjaman dengan persyaratan lunak kepada PT Pe...
11358	Peraturan_Presiden_No.9_Tahun_2011	4	1	Ketentuan lebih lanjut yang diperlukan dalam r...
11359	Peraturan_Presiden_No.9_Tahun_2011	5	1	Peraturan Presiden ini mulai berlaku pada tang...

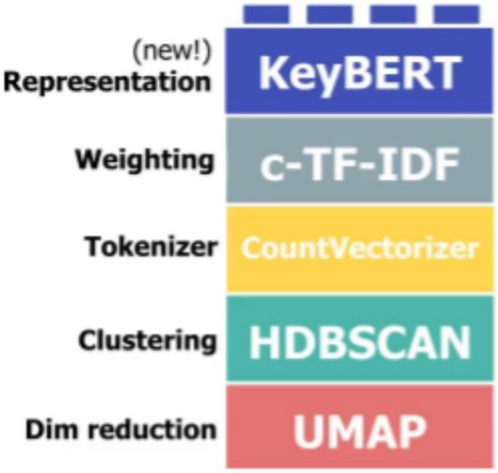
Gambar 3. *Structured Data*

4.2 *Topic Modeling*

At this stage, each verse will be grouped based on the topic discussed in the the verse. This stage aims to ensure the process of identifying the harmony between each put verses to be efficient. By doing ***topic modeling,*** pair the necessary verses identified the harmony of the verses only verses that discuss the same topic. While pairs of verses that have different topics do not need to be identified because both pairs These verses must be mutually neutral so that they are included in the harmonious relationship of verses.

The following is the process carried out at ***the topic modeling stage.***

1. Each verse in ***the dataset*** will undergo an ***embedding process.*** The ***embedding*** process is useful to convert the input into a numeric representation. This process is carried out using ***Sentence Transformers*** with the model used is paraphrase-multilingual-MiniLM-L12-v2 which supports multiple languages including Indonesian and has a smaller size than other models.
2. Modeling the model used to perform topic ***clustering*** in each verse. At this stage, the BERTopic ***topic modeling*** technique is used , where ***the sub*** -model... used is UMAP as ***dimensionality reduction,*** HDBSCAN as ***clustering model,*** CountVectorizer as ***tokenizer,*** and c-TF-IDF as scheme ***word weighting.*** In the model, the ***min_topic_size*** parameter is also set to 2 so that a topic can be formed if there are 2 verses that have the same ***cluster*** same. This is done so that there are fewer ***outliers ,*** thus reducing the data missing.



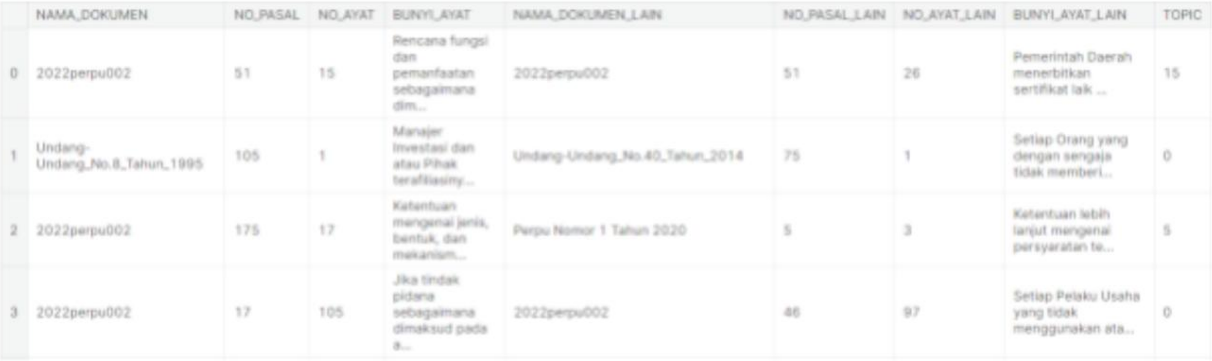
Gambar 4. Struktur Model

3. ***Fine-tuning keywords*** generated by c-TF-IDF using the KeyBERT model. By implementing this, the topic representation has a more coherent and able to reduce the presence of stop words ***in*** topic representations produced.

4. The results of the previous process will be ***post-processed. Post-processing*** which carried out with the aim of reducing ***mini clusters*** while reducing ***outliers*** on the dataset so that it can reduce the amount of missing data. This process is carried out by combining similar topics from all previous topics generated model.

4.3 Classification of Verse Pairs

At this stage, a classification will be carried out on each pair of verses that have the same topic. Previously, a ***pre-processing*** stage was required on the dataset to prepare the data before running it in the classification algorithm. ***Pre-processing*** stage done by grouping data rows based on topic categories. owned. In each group, each row of data will be paired with another row of data. using the ***itertools*** library in python. Finally, each pair of verses will be combined on one ***dataframe***



	NAMA_DOKUMEN	NO_PASAL	NO_AYAT	BUNYI_AYAT	NAMA_DOKUMEN_LAIN	NO_PASAL_LAIN	NO_AYAT_LAIN	BUNYI_AYAT_LAIN	TOPIC
0	2022perpu002	51	15	Rencana fungsi dan pemanfaatan sebagaimana dim...	2022perpu002	51	26	Pemerintah Daerah menerbitkan sertifikat laik ...	15
1	Undang-Undang_No.8_Tahun_1995	105	1	Manajer Investasi dan atau Pihak terafiliasiny...	Undang-Undang_No.40_Tahun_2014	75	1	Setiap Orang yang dengan sengaja tidak memberi...	0
2	2022perpu002	175	17	Ketentuan mengenai jenis, bentuk, dan mekanism...	Perpu Nomor 1 Tahun 2020	5	3	Ketentuan lebih lanjut mengenai persyaratan te...	5
3	2022perpu002	17	105	Jika tindak pidana sebagaimana dimaksud pada a...	2022perpu002	46	97	Setiap Pelaku Usaha yang tidak menggunakan ata...	0

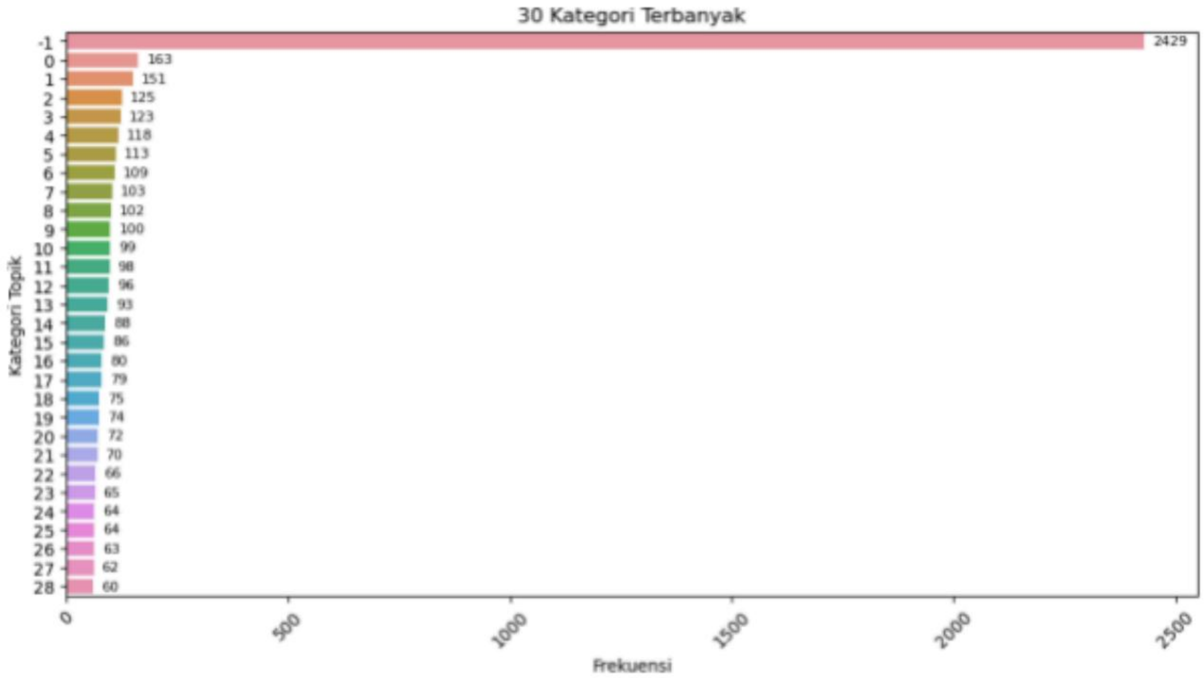
Gambar 5. Hasil *pre-processing*

To perform classification on each pair of verses in each row of data used ***pre-trained*** model Indo-roberta-indonli (Limcorn, 2022). Indo-roberta-indonli itself is a ***natural language inference (NLI)*** classification model which is based on the model Indonesian RoBERTa (Wongso et al., 2021) which is the Indonesian version of RoBERTa model. The Indo-roberta-indonli model has been trained on the IndoNLI dataset (Mahendra et al., 2021) thereby allowing the model to classify two statements/arguments in the categories of ***entailment, neutral,*** and ***contradiction.*** This model can used to analyze the relationship between two verses in legislation where for pairs of verses that have an ***entailment*** or ***neutral*** relationship then This pair of verses is classified as a pair of harmonious verses. While the verse pair which has a ***contradictory*** relationship is classified as a pair of verses that are not in harmony.

V. DISCUSSION

Initially there were 110 regulatory documents consisting of from Government Regulations, Presidential Regulations, Ministerial Regulations, Laws, and Government Regulation in Lieu of Law. Changing ***unstructured data*** into ***structured data*** produces a total of 11360 verses. This means it will be generated as many as 64,519,120 pairs of verses that need to be identified for their harmony. With the existence of ***topic modeling,*** the number of verse pairs that need to be identified for their harmony is 178,845 install verses so that the analysis process becomes more efficient.

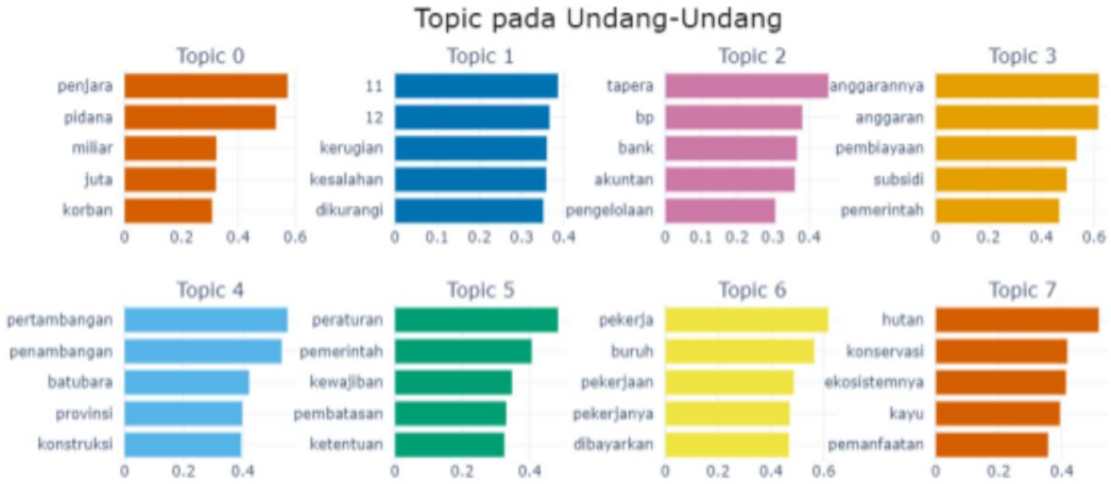
At the ***topic modeling*** stage , 949 topic categories were generated with frequencies of the 30 most categories can be seen in the following graph.



Gambar 6. Grafik frekuensi topik

In the graph above, the topic category numbered “-1” is ***an outlier*** where there is no other verses that have the same category as the verse. While the topic category numbered “0” is the topic with the highest frequency of 163 verses.

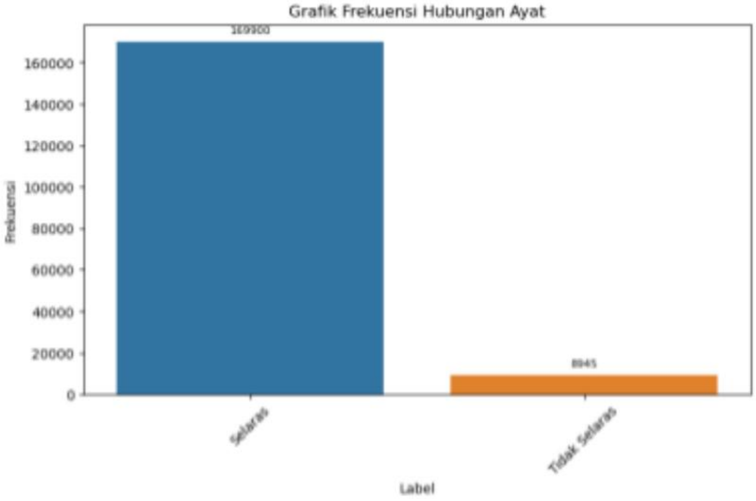
For ***keyword*** representation on several topics generated by the model can be seen in the graph below.



Gambar 7. *Topic word scores*

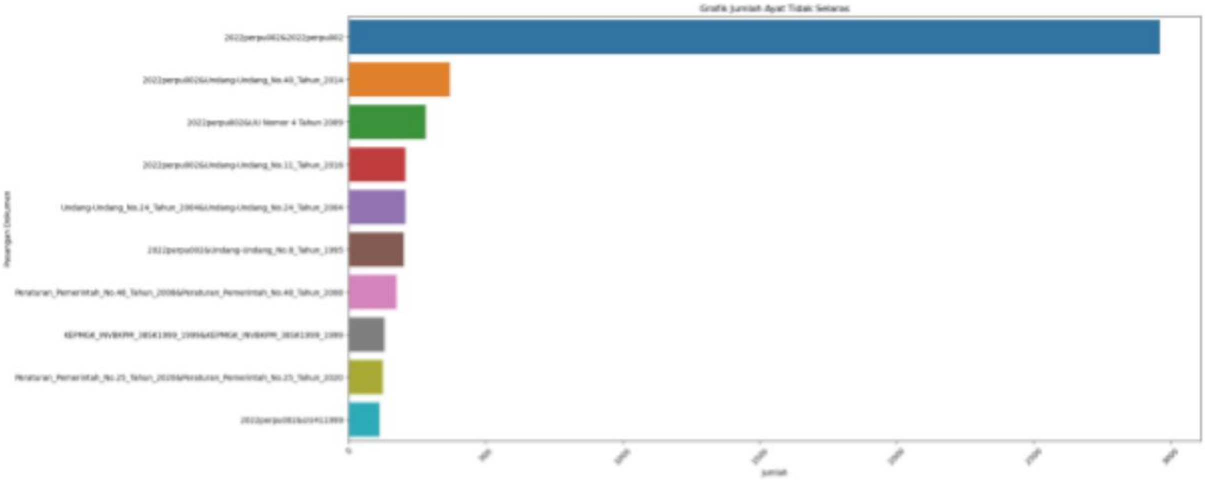
The x-axis in the graph above depicts the scores of the fine ***-tuned*** c-TF-IDF. with KeyBERT. While the y-axis depicts 5 words or ***keywords*** with a score the highest representing the related topic. Based on the graph, it can be seen that Most of the verses in the laws and regulations that are given discuss the related matters. criminal penalties or imprisonment with the keywords being prison, criminal, billion, million, and victim.

Based on the predictions generated by the model, as many as 169,900 pairs of verses or about 95% of the verse pairs analyzed have a harmonious relationship. If the percentage calculated for all existing pairs of verses is 64,519,120 pairs of verses, So the percentage of pairs of verses that are not in harmony is only around 0.01%.



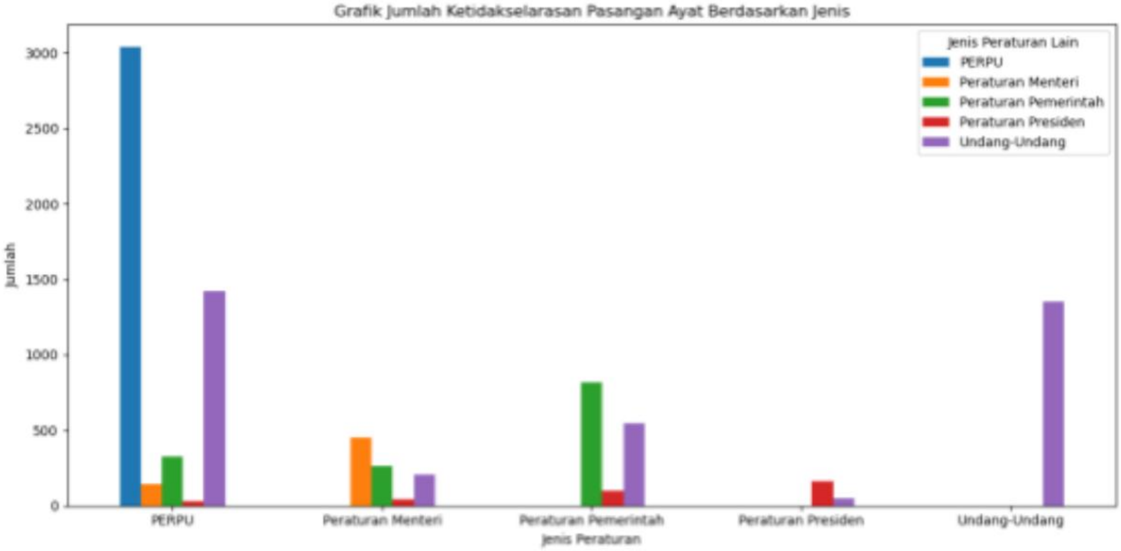
Gambar 8. Grafik Prediksi Model

Meanwhile, if a pair of verses that have an inconsistent relationship are analyzed furthermore, it was found that the pair of verses in the Government Regulation in Lieu of Law Number 2 of 2022 has the most pairs of inconsistent verses namely 2961 pairs.



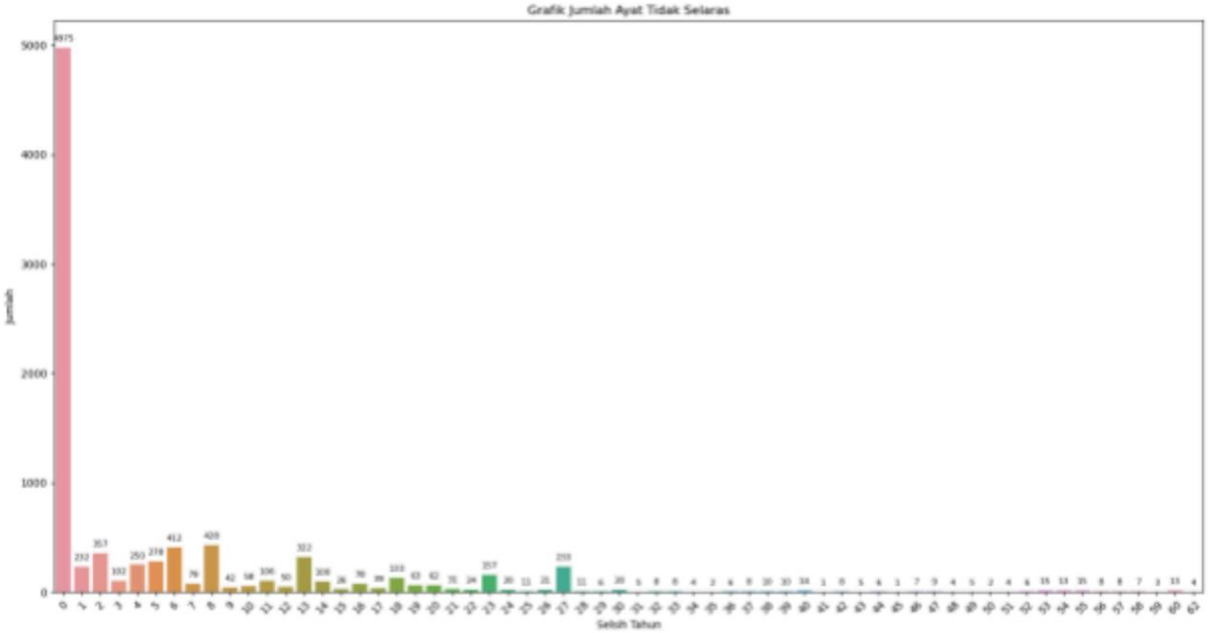
Gambar 9. Grafik Jumlah Ayat Tidak Selaras

If we look at the type of legislation, it can be seen that that the model tends to find pairs of verses that are incongruent in type the same laws and regulations.



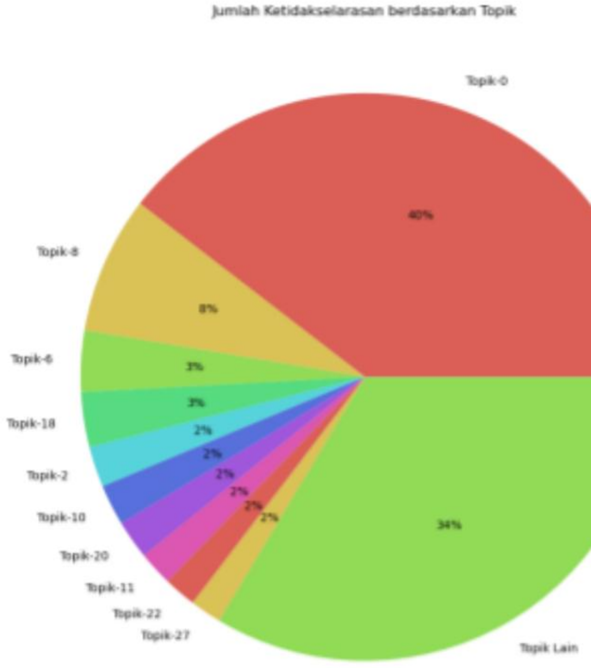
Gambar 10. Jumlah Ketidakselarasan Berdasarkan Jenis

In the pairs of verses that are not in harmony it is also seen that both tend to be passed in the same year, as seen in the following graph.



Gambar 11. Selisih Tahun pada Pasangan Ayat Tidak Selaras

Based on the topic, pairs of verses that are not in harmony tend to have the same topic. number 0 in **Figure 11,** which discusses criminal penalties or imprisonment where around 40% of the total pairs of incongruent verses.



Gambar 12. Jumlah Ayat Tidak Selaras Berdasarkan Topik

On a dataset containing 1057 rows of manually ***labeled*** data, the model which is used has an accuracy of 93.34%. However, there are still quite a lot of ***false positive*** on the predictions generated by the model where the sentence should not be contains a contradiction, but is detected to contain a contradiction. This tends to happen in complex regulatory documents, one of which is the Regulation Government in Lieu of Law Number 2 of 2020 which has more than 1000 pages that cause the model to have difficulty understanding the context of the regulations. the legislation. For that reason, the model can be further developed by train it on a dataset containing statements in the regulations legislation. This can make the model better at understanding and analyze sentences in laws.

VI. CLOSING

6.1 Conclusion

Investment laws and regulations tend to have a high level of good harmony or alignment because only a small part of the verse pairs which contains contradictions. The Indo-roberta-indonli model also has a good performance. quite good at identifying alignment with related laws and regulations investment. However, the model has a less than good performance on complex legal documents. In addition, the use of BERTopic also proven to help the process of analyzing the harmonization of legislation to become more efficient.

6.2 Suggestions

As a country based on law, harmony or alignment of laws is a things that need to be considered and analyzed carefully. Harmony analysis laws are very important because laws are the legal basis which regulates the lives of various levels of society. For this reason, Indonesia should be able to provide and store data on statutory regulations better as well involving people who are experts in the legal field in collecting or ***labeling*** data. With this data, a more accurate system or model can be created in analyzing the alignment of laws and regulations to reduce uncertainty law in Indonesia.

VII. REFERENCES

Bureau of Public Relations, Law, and Cooperation of the Ministry of Law and Human Rights. (2020). ***Minister of Law and Human Rights: Convenience Investment For the sake of Prosperity Public Indonesia.*** https://www.kemenkumham.go.id/berita-utama/menkumham-keeasy-investasi-demi-kem Indonesian-society-accommodation

Saputra, A. (2016, September 29). ***Legal Reform Package, Jokowi Must Complete It Egosectoral Between Institution.*** Detik news. https://news.detik.com/berita/d-5995527/peraturan-perundang-undangan-pengertian-jen-hin gga-load-material

Stanford NLP Group. (nd). ***The Stanford Natural Language Inference (SNLI) Corpus.*** https://nlp.stanford.edu/projects/snli/

Wang, S., Fang, H., Khabsa, M., Mao, H., Ma, H. (2021). ***Entailment as Few-Shot Learner.*** https://arxiv.org/pdf/2104.14690

Grootendorst, M. (2022). ***BERTopic: Neural topic modeling with a class-based basis TF-IDF procedure.*** https://arxiv.org/abs/2203.05794

Singh, A. (2021). ***Evolving with BERT: Introduction to RoBERTa.*** https://medium.com/analytics-vidhya/evolving-with-bert-introduction-to-roberta-5174ec0e7c8 2

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). ***RoBERTa: A Robustly Optimized BERT Pretraining Approach.*** https://arxiv.org/abs/1907.11692

Lemon Corn, S. (2021). ***Indo-roberta-indonli.*** HuggingFace. https://huggingface.co/StevenLimcorn/indo-roberta-indonli

Wongso, W., Limcorn, S., Rahmadani, S., Wah, CK (2021). ***Indonesian RoBERTa Base.*** HuggingFace. https://huggingface.co/flax-community/indonesian-roberta-base

Mahendra, Rahmad and Aji, Fikri, A. and Louvan, Samuel and Rahman, Fahrurrozi and Vania, Clara. (2021, December 5). ***IndoNLI: A Natural Language Inference Dataset for Indonesian.*** https://github.com/ir-nlp-csui/indonli/tree/main