

## **I. JUDUL DAN ABSTRAK**

### **1.1 Judul**

Pendeteksian Potensi Disharmoni pada Peraturan Perundang-Undangan Bidang Investasi Menggunakan *Topic Modeling Latent Dirichlet Allocation* (LDA) dan Model *Natural Language Inference* (NLI) IndoBERT

### **1.2 Abstrak**

Analisis menyeluruh terhadap keselarasan antar peraturan-peraturan dengan norma dasar dan hierarki perundang-undangan penting untuk memastikan implementasi perubahan undang-undang berjalan efisien dan konsisten dengan sistem hukum yang berlaku. Salah satu bidang yang menjadi fokus dalam penerapan undang-undang *omnibus law* yang sedang menjadi perhatian masyarakat Indonesia adalah bidang investasi. Pada penelitian ini akan digunakan *machine learning* dalam analisis keselarasan perundang-undangan untuk memfasilitasi identifikasi potensi disharmoni atau ketidakselarasan antar undang-undang bidang investasi dengan lebih efisien dan efektif. Metode yang akan digunakan untuk mengidentifikasi potensi disharmoni atau ketidakselarasan antar undang-undang adalah *Latent Dirichlet Allocation* (LDA) untuk *clustering* topik undang-undang dan *Model Natural Language Inference* (NLI) IndoBERT untuk menilai hubungan antar ayat. Hasil penelitian menunjukkan bahwa pemodelan klasifikasi menggunakan teknik *machine learning* dapat diaplikasikan untuk mempermudah pengidentifikasian ketidakselarasan dalam perundang-undangan. Metode LDA dapat mereduksi ruang pencarian dimana hanya peraturan yang membahas hal yang sama saja yang dibandingkan dan penggunaan model pemrosesan NLI IndoBERT dapat membantu untuk mengidentifikasi adanya potensi kontradiksi atau perbedaan dalam interpretasi antara berbagai teks ayat peraturan. Hasil penelitian ini diharapkan dapat memberikan kontribusi yang signifikan bagi pengembangan kebijakan dan praktik hukum di Indonesia, terutama dalam bidang investasi.

Kata kunci: IndoBERT, investasi, LDA, NLI, undang-undang

## II. PENDAHULUAN

### 2.1 Latar Belakang

Sistem hukum dan perundang-undangan memiliki peran krusial dalam mengatur kehidupan masyarakat dan menjaga stabilitas negara. Indonesia sebagai negara hukum memiliki perundang-undangan yang dibentuk dari berbagai tingkatan. Hierarki hukum ini mencerminkan pentingnya keselarasan dan keharmonisan antara peraturan-peraturan yang ada, agar pelaksanaan hukum berjalan efisien, adil, dan sesuai dengan prinsip hukum yang berlaku (Kemenkumham 2020). Dalam konteks ini, permasalahan keselarasan perundang-undangan menjadi hal yang kompleks dan menarik untuk diteliti. Diperlukan pemahaman mendalam dan analisis yang tepat untuk menemukan apakah terdapat kesesuaian antara ayat-ayat dalam peraturan-peraturan yang ada. Identifikasi potensi ketidakselarasan atau ketidakharmonisan antara ayat-ayat tersebut penting guna menjaga kepastian hukum dan mencegah terjadinya konflik hukum di masa yang akan datang.

Dalam upaya untuk mendorong pertumbuhan ekonomi, Presiden Joko Widodo menyampaikan komitmen pemerintah untuk menerbitkan dua undang-undang (UU) besar, yaitu UU Cipta Lapangan Kerja dan UU Pemberdayaan Usaha Makro, Kecil, dan Menengah (UMKM). Dua UU besar ini memiliki peran strategis dalam meningkatkan daya saing ekonomi Indonesia dan menciptakan lapangan kerja bagi masyarakat. Salah satu bidang yang menjadi fokus dalam penerapan *omnibus law* adalah bidang investasi. Penyederhanaan dan harmonisasi dalam peraturan investasi menjadi langkah krusial untuk menciptakan iklim investasi yang kondusif dan menarik bagi para investor baik dari dalam maupun luar negeri. Dengan menghadirkan peraturan yang saling mendukung dan tidak tumpang tindih, diharapkan akan memberikan kepastian hukum dan mendorong pertumbuhan investasi di Indonesia (Suriadinata 2019).

Dalam konteks harmonisasi Undang-Undang (UU) bidang investasi, konsep jenjang norma hukum (*Stufentheorie*) dari Kelsen (1961) menjadi relevan. Hans Kelsen menyatakan bahwa norma-norma hukum memiliki hierarki dan berjenjang, di mana norma yang lebih rendah bersumber dan berlaku berdasarkan pada norma yang lebih tinggi. Konsep ini mencerminkan pentingnya konsistensi dan keselarasan antara peraturan-peraturan yang ada. Dalam penilaian keselarasan perundang-undangan, norma dasar (*grundnorm*) berperan sebagai landasan yang menjadi acuan untuk norma-norma di bawahnya. Norma dasar ditetapkan oleh masyarakat sebagai norma yang mendasari

seluruh peraturan hukum. Sejalan dengan pandangan Indrati (2007), norma dasar menjadi *presupposed* atau digunakan sebagai asumsi dasar dalam konstruksi hukum.

Perlu dilakukan analisis menyeluruh terhadap keselarasan antar peraturan-peraturan yang ada dengan norma dasar dan hierarki perundang-undangan yang berlaku. Hal ini penting untuk memastikan bahwa setiap perubahan atau penyesuaian dalam UU tidak bertentangan dengan norma dasar dan peraturan di tingkatan yang lebih tinggi. Dengan begitu, harmonisasi dan koordinasi antar peraturan-peraturan dapat terjaga, dan implementasi perubahan UU bidang investasi dapat berjalan secara efisien dan konsisten dengan sistem hukum yang berlaku.

Penerapan teknologi *big data* dan *artificial intelligence* (AI) memberikan potensi besar untuk memfasilitasi analisis keselarasan perundang-undangan secara efisien dan akurat. Dengan memanfaatkan pemodelan *text analytic*, dapat ditemukan pola-pola dan hubungan antar ayat-ayat dalam peraturan, serta memvisualisasikan hasil analisis tersebut dalam bentuk yang lebih intuitif. Salah satu pemodelan *text analytical* yang dapat digunakan untuk mengidentifikasi potensi disharmoni atau ketidakselarasan antar undang-undang adalah *Latent Dirichlet Allocation* (LDA) untuk *clustering* topik undang-undang dan *Model Natural Language Inference* (NLI) IndoBERT untuk menilai hubungan antar ayat. Dengan memanfaatkan LDA, dapat ditemukan pola-pola dan hubungan antar ayat-ayat dalam peraturan, sehingga potensi ketidakselarasan dapat teridentifikasi lebih cepat. Kemudian Model NLI IndoBERT dapat digunakan untuk melakukan penilaian keselarasan antara ayat-ayat dalam peraturan dan norma dasar yang berlaku. Model ini dapat memahami hubungan antar kalimat secara lebih mendalam dan akurat, sehingga dapat mendeteksi ketidakharmonisan dengan tingkat akurasi yang tinggi (Labban *et al.* 2022).

## 2.2 Rumusan Masalah

Berdasarkan latar belakang di atas, rumusan masalah dari penelitian ini adalah:

1. Bagaimana penerapan teknologi *big data* dan *artificial intelligence* (AI) dalam analisis keselarasan perundang-undangan dapat memfasilitasi identifikasi potensi disharmoni atau ketidakselarasan antar undang-undang?

2. Bagaimana pemanfaatan pemodelan *Latent Dirichlet Allocation* (LDA), dapat membantu mengidentifikasi pola-pola dan hubungan antar ayat-ayat dalam peraturan untuk mendukung analisis keselarasan?
3. Bagaimana Model *Natural Language Inference* (NLI) IndoBERT dapat digunakan untuk menilai hubungan antar ayat-ayat dalam peraturan dan norma dasar yang berlaku sehingga dapat mendeteksi potensi disharmoni antar undang-undang secara efisien dan akurat?
4. Bagaimana relevansi dan dampak penerapan teknologi *big data* dan *artificial intelligence* dalam menyelesaikan masalah harmonisasi perundang-undangan di Indonesia?

### 2.3 Tujuan

Makalah ini memiliki beberapa tujuan untuk mencapai pemahaman yang komprehensif tentang keselarasan perundang-undangan di Indonesia terkait dengan bidang investasi. Beberapa tujuan tersebut adalah:

1. Mengkaji penerapan teknologi *big data* dan *artificial intelligence* (AI) dalam analisis keselarasan perundang-undangan untuk memfasilitasi identifikasi potensi disharmoni atau ketidakselarasan antar undang-undang dengan lebih efisien dan akurat.
2. Menjelaskan bagaimana pemodelan *text analytic*, khususnya *Latent Dirichlet Allocation* (LDA), dapat membantu mengidentifikasi pola-pola dan hubungan antar ayat-ayat dalam peraturan untuk mendukung analisis keselarasan secara lebih mendalam dan intuitif.
3. Mengimplementasikan model *Natural Language Inference* (NLI) IndoBERT untuk menilai hubungan antar ayat-ayat dalam peraturan dan norma dasar yang berlaku sehingga dapat mendeteksi potensi disharmoni antar undang-undang dengan tingkat akurasi yang tinggi.
4. Menganalisis manfaat dan potensi dampak positif dari penerapan teknologi *big data* dan *artificial intelligence* dalam mencapai keselarasan perundang-undangan, sehingga dapat menghasilkan rekomendasi dan solusi untuk meningkatkan keselarasan perundang-undangan di Indonesia, khususnya dalam bidang investasi.

## 2.4 Manfaat

Makalah ini diharapkan akan memberikan manfaat bagi berbagai pihak, termasuk pemerintah, akademisi, praktisi hukum, dan masyarakat umum. Beberapa manfaat dari makalah ini adalah:

1. Mengidentifikasi kontribusi teknologi *big data* dan *artificial intelligence* dalam meningkatkan efisiensi dan akurasi analisis harmonisasi perundang-undangan.
2. Memperkenalkan penerapan teknologi *Natural Language Processing* (NLP) dan pemodelan *text analytic* seperti LDA dan NLI IndoBERT dalam analisis perundang-undangan, sehingga dapat menjadi panduan bagi penelitian dan implementasi serupa di masa depan.
3. Mengoptimalkan penggunaan teknologi *big data* dan AI untuk menyusun undang-undang yang lebih harmonis dan selaras dengan norma dasar yang berlaku, sehingga tercipta lingkungan hukum yang lebih adil, efisien, dan sesuai dengan prinsip hukum yang berlaku.
4. Memberikan panduan dan rekomendasi bagi pembuat kebijakan dan praktisi hukum untuk meningkatkan harmonisasi perundang-undangan di Indonesia, khususnya dalam bidang investasi.

## III. METODOLOGI

### 3.1. Data

Data yang digunakan pada penelitian ini adalah data peraturan perundang-undangan yang terdiri dari empat jenis peraturan, yaitu Undang-undang, Peraturan Pemerintah, Peraturan Presiden, dan Peraturan Menteri yang berfokus pada investasi. Data ini mencakup periode tahun 1965 hingga 2021 dan terdiri dari 110 dokumen. Informasi yang akan digunakan pada setiap dokumen berupa jenis peraturan, nomor dan tahun peraturan, nomor pasal dan ayat, serta teks dari masing-masing ayat.

Pada penelitian ini juga digunakan dataset “IndoNLI” yang tersedia pada *website* <https://huggingface.co/datasets/indonli>. Dataset "IndoNLI" adalah sebuah kumpulan data bahasa Indonesia yang dikumpulkan untuk pemrosesan pemahaman bahasa alami (*Natural Language Inference* - NLI). Dataset ini dirancang untuk mendukung penelitian dan pengembangan dalam memahami hubungan antara dua kalimat dalam bahasa Indonesia, yaitu kalimat premis dan kalimat hipotesis. Pada penelitian ini, dataset "IndoNLI" akan

digunakan sebagai bagian dari tahapan *training* model NLI dengan menggunakan IndoBERT. Dataset "IndoNLI" akan menjadi dasar untuk pelatihan model agar implikasi antar ayat dalam peraturan perundang-undangan dapat dipahami dan dinilai dengan lebih baik.

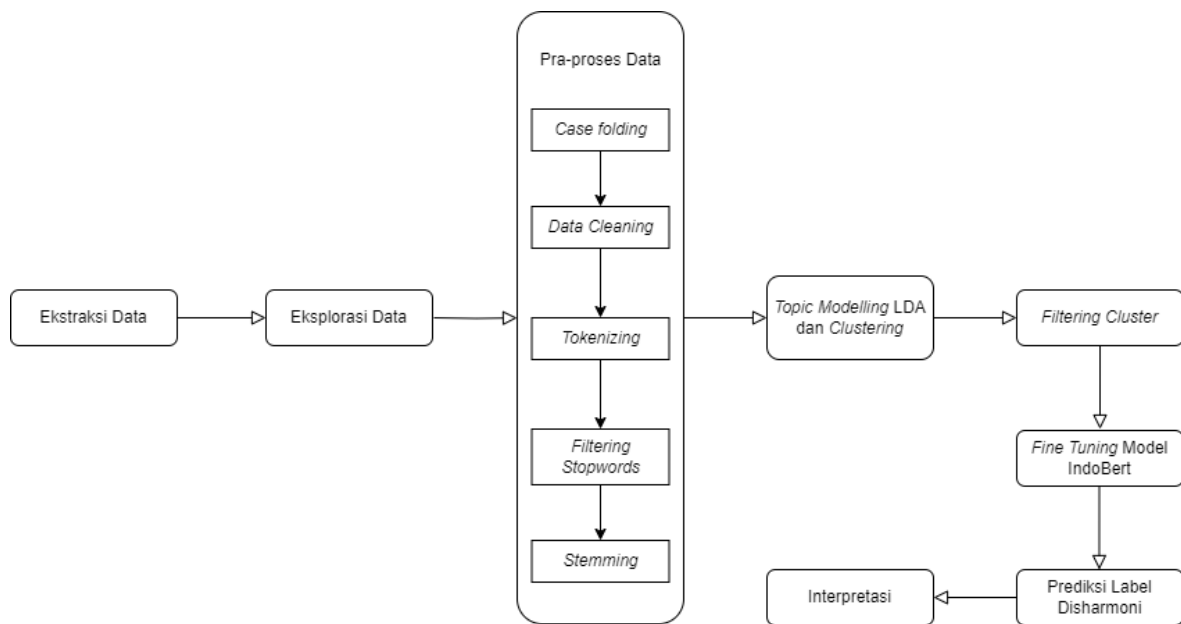
### 3.2. Kerangka Berpikir

Proses penelitian dimulai dengan memahami dokumen-dokumen peraturan perundang-undangan yang tersedia sebagai data *input* penelitian. Kemudian tahapan dilanjutkan dengan mengekstraksi dokumen-dokumen tersebut agar lebih mudah untuk dibaca dan dianalisis menggunakan komputer. Setelah proses ekstraksi selesai, dilakukan eksplorasi data untuk memahami pola dan gambaran umum dari dokumen-dokumen yang telah disesuaikan. Dokumen-dokumen ini akan melalui tahap *pre-processing* untuk membersihkan tiap teks dan mengubahnya menjadi format yang lebih sesuai untuk analisis. Setelah tahapan *pre-processing* akan didapatkan kolom berisi masing-masing ayat pada setiap pasal undang-undang yang disimpan pada suatu *data frame*. Selanjutnya, akan diidentifikasi topik-topik utama yang muncul dalam tiap ayat dokumen perundang-undangan. Dengan menggunakan teknik ini, ayat-ayat tersebut dapat dikelompokkan berdasarkan konteks dan topik yang dibahasnya, sehingga akan terbentuk beberapa kelompok atau *cluster* beserta nilai probabilitas dari ayat tersebut untuk masuk pada suatu *cluster*. Kemudian akan dilakukan tahap *filtering cluster* dengan menghitung jarak *Euclidean* dari nilai probabilitas tiap dua ayat dan ditentukan suatu nilai *cut off* agar didapatkan *cluster-cluster* yang hanya berisi kumpulan ayat dengan kemiripan yang tinggi. Proses tersebut dilakukan untuk mengurangi jumlah pasangan ayat yang akan dianalisis pada tahapan selanjutnya.

Selanjutnya, akan dilakukan analisis menggunakan bantuan komputer untuk mengevaluasi implikasi dan potensi kontradiksi antar ayat-ayat dalam *data frame* yang telah dihasilkan. Pasangan ayat yang dibandingkan hanya ayat-ayat yang berada dalam satu *cluster* dan “lolos” pada tahap *filtering* sebelumnya. Analisis potensi disharmoni antar ayat menghasilkan *output* berupa suatu kolom yang akan berisi keterangan “*neutral*” jika pasangan ayat tersebut tidak ada hubungan implikasi yang jelas, “*entailment*” jika pasangan ayat tersebut bersesuaian, atau “*contradiction*” jika pasangan ayat tersebut bertentangan. Dengan demikian, potensi disharmoni antar ayat-ayat yang terkait dengan investasi dapat dideteksi dan diinterpretasikan dengan lebih baik.

### 3.3. Prosedur Analisis Data

Gambar 1 di bawah menggambarkan alur analisis data yang digunakan untuk mendeteksi potensi disharmoni pada peraturan perundang-undangan di bidang investasi. Proses dimulai dengan mengumpulkan dan mengekstraksi dokumen peraturan perundang-undangan yang digunakan pada penelitian ini. Kemudian dilakukan eksplorasi untuk melihat pola dan gambaran umum dari dokumen perundang-undangan yang digunakan. Selanjutnya, dokumen-dokumen ini akan melalui tahap *pre-processing* untuk membersihkan teks dan mengubahnya menjadi format yang lebih sesuai untuk analisis.



**Gambar 1** Diagram alur proses analisis data

Setelah proses *pre-processing*, *data frame* berisi teks perundang-undangan tersebut akan digunakan untuk melatih model *Latent Dirichlet Allocation* (LDA). Model LDA akan mengidentifikasi topik-topik utama yang muncul dalam korpus dokumen perundang-undangan, sehingga dapat membagi dokumen menjadi kelompok-kelompok berdasarkan konteks yang dibahasnya. Kemudian akan dilakukan tahap *filtering* kelompok atau *cluster* dengan menghitung jarak *Euclidean* dari nilai probabilitas tiap dua objek (ayat) dan menetapkan nilai *cut off* agar hanya pasangan ayat dengan nilai jarak yang minimum (tingkat kemiripan tinggi) yang digunakan pada tahapan *Fine Tuning*. Selanjutnya, akan dilakukan tahap analisis menggunakan Model *Natural Language Inference* (NLI) IndoBERT. Model NLI ini akan membantu dalam mengevaluasi implikasi dan potensi kontradiksi antara ayat-ayat dalam dokumen perundang-undangan. Dengan

demikian, potensi disharmoni antara ayat-ayat dapat dideteksi dan diinterpretasikan dengan lebih baik.

#### 3.4. Ekstraksi Data

Data diekstraksi dari dokumen dengan format *pdf* menjadi suatu *data frame* menggunakan *package* “*pflplumber*” pada bahasa pemrograman Python. Setiap dokumen dibagi berdasarkan jenis peraturan serta nomor dan tahunnya, kemudian informasi pada dokumen akan dibaca per pasal yang ditandai dengan kata “pasal” yang terletak di antara dua *newline* dan selanjutnya diikuti oleh ayat yang ditandai dengan poin angka di dalam tanda kurung (contoh = (1)).

Berikut merupakan Tabel 1 yang memuat contoh hasil ekstraksi dokumen Undang-Undang Nomor 1 Tahun 2004 Pasal 1 Ayat 1 dan 2. Tabel di bawah juga menunjukkan bahwa unit yang akan digunakan dalam penelitian ini adalah ayat.

**Tabel 1** Contoh hasil ekstraksi data

Jenis	Nomor	Tahun	Pasal	Ayat	Teks
Undang-undang	1	2004	1	1	Perbendaharaan Negara adalah pengelolaan dan pertanggungjawaban keuangan negara, termasuk investasi dan kekayaan yang dipisahkan, yang ditetapkan dalam APBN dan APBD.
Undang-undang	1	2004	1	2	Kas Negara adalah tempat penyimpanan uang negara yang ditentukan oleh Menteri Keuangan selaku Bendahara Umum Negara untuk menampung seluruh penerimaan negara dan membayar seluruh pengeluaran negara.



### 3.5 Pra Proses Data

#### 3.5.1. *Case folding*

*Case folding* merupakan tahapan pada *text pre-processing* yang menyamaratakan penggunaan huruf kapital dengan mengubah semua huruf menjadi huruf kecil (*lowercase*). Hal ini dilakukan untuk menghindari terjadinya *double indexing* di mana dua kata yang sama terdeteksi sebagai dua objek yang berbeda karena penggunaan kapital yang berbeda. Berikut adalah Tabel 2 yang memuat contoh hasil *case folding*.

**Tabel 2** Hasil *case folding*

Teks sebelum <i>case folding</i>	Teks setelah <i>case folding</i>
Bank Sentral adalah sebagaimana dimaksud dalam Undang Undang Dasar 1945 Pasal 23D.	bank sentral adalah sebagaimana dimaksud dalam undang undang dasar 1945 pasal 23d.
Kepala Satuan Kerja Pengelola Keuangan Daerah adalah Bendahara Umum Daerah.	kepala satuan kerja pengelola keuangan daerah adalah bendahara umum daerah.

#### 3.5.2. *Data cleansing*

*Data cleansing* dilakukan untuk mengurangi *noise* pada data dengan menghapus karakter yang tidak relevan pada analisis, seperti tanda baca dan *whitespace*. Tahapan *data cleansing* pada penelitian ini dibagi menjadi dua bagian sesuai dengan kebutuhan metode analisis, yaitu:

- Untuk kebutuhan *Topic Modeling* diperlukan kata-kata “unik” pada setiap ayat agar dapat menggambarkan topik yang dibahas oleh setiap kalimat, sehingga *string* angka pada data akan dihapus. Contoh hasil *data cleansing* untuk *Topic Modeling* ditampilkan pada tabel berikut.

**Tabel 3** *Data cleansing* untuk *Topic Modeling*

Teks sebelum <i>cleansing</i>	Teks setelah <i>cleansing</i>
bank wajib memberikan keterangan sebagaimana dimaksud dalam pasal 42 dan pasal 43	bank wajib memberikan keterangan sebagaimana dimaksud dalam pasal dan pasal

penyaluran dana bagi hasil pbb dan bphtb sebagaimana dimaksud pada ayat (3) dan ayat (4) dilakukan sesuai dengan peraturan perundang undangan.	penyaluran dana bagi hasil pbb dan bphtb sebagaimana dimaksud pada ayat dan ayat dilakukan sesuai dengan peraturan perundang undangan
--	---

- Untuk pemodelan klasifikasi label, diperlukan informasi yang lengkap dari suatu ayat agar perbandingan antar ayat lebih tepat, sehingga *string* angka pada data akan dipertahankan. Berikut adalah Tabel 4 yang memuat contoh hasil *Data cleansing* untuk *Modeling*.

**Tabel 4** *Data cleansing* untuk *Modeling*

Teks sebelum <i>cleansing</i>	Teks setelah <i>cleansing</i>
WP terdiri atas: a. WUP; b. WPR; dan c. WPN.	wp terdiri atas a wup b wpr dan c wpn
Wilayah yang akan diusahakan sebagaimana dimaksud pada ayat (2) dan ayat (3) berubah statusnya menjadi WUPK.	wilayah yang akan diusahakan sebagaimana dimaksud pada ayat 2 dan ayat 3 berubah statusnya menjadi wupk

### 3.5.3. *Tokenizing*

*Tokenizing* atau tokenisasi merupakan proses pemisahan teks menjadi potongan-potongan yang disebut “token” untuk kemudian dianalisis pada tahap selanjutnya. Token dibentuk dengan memotong teks yang dipisahkan oleh *whitespace* atau tanda spasi yang kemudian disimpan dalam bentuk *list*. Contoh hasil *Tokenizing* teks ditampilkan pada tabel berikut.

**Tabel 5** Hasil *Tokenizing* teks

<b>Data sebelum <i>tokenizing</i></b>	<b>Data setelah <i>tokenizing</i></b>
bank wajib memberikan keterangan sebagaimana dimaksud dalam pasal 42 dan pasal 43	['bank', 'wajib', 'memberikan', 'keterangan', 'sebagaimana', 'dimaksud', '', 'dalam', 'pasal', '42', 'dan', 'pasal', '43']
penyaluran dana bagi hasil pbb dan bphtb sebagaimana dimaksud pada ayat 3 dan ayat 4 dilakukan sesuai dengan peraturan perundang undangan	['penyaluran', 'dana', 'dan', 'bagi', 'hasil', 'pbb', 'dan', 'bphtb', 'sebagaimana', 'dimaksud', 'pada', 'ayat', '3', 'dan', 'ayat', '4', 'dilakukan', 'sesuai', 'dengan', 'peraturan', 'perundang', 'undangan']

#### 3.5.4. *Filtering* (penghapusan *stopwords*)

*Stopwords* adalah kata-kata yang tidak memberikan informasi tambahan pada proses analisis. *Stopwords* dapat berupa kata penghubung atau kata ganti seperti 'dan', 'adalah', 'pada', dan 'saya'. Tahap *filtering* hanya dilakukan untuk analisis *Topic Modeling* karena pada klasifikasi diperlukan teks utuh agar model dapat memahami konteks dari kalimat secara lebih lengkap. Selain *stopwords* dasar, akan dihapus juga kata-kata yang sering muncul pada perundang undangan tetapi tidak menambah makna atau menggambarkan topik yang dibahas oleh ayat tersebut, seperti kata 'ayat', 'pasal', dan 'undang-undang'. Berikut adalah Tabel 6 yang memuat contoh hasil *filtering* untuk *Topic Modeling*.

**Tabel 6** Contoh hasil *filtering*

<b>Data sebelum <i>filtering</i></b>	<b>Data setelah <i>filtering</i></b>
['bank', 'wajib', 'memberikan', 'keterangan', 'sebagaimana', 'dimaksud', '', 'dalam', 'pasal', '42', 'dan', 'pasal', '43']	['bank', 'wajib', 'memberikan', 'keterangan']

['penyaluran', 'dana', 'dan', 'bagi', 'hasil', 'pbb', 'dan', 'bphtb', 'sebagaimana', 'dimaksud', 'pada', 'ayat', '3', 'dan', 'ayat', '4', 'dilakukan', 'sesuai', 'dengan', 'peraturan', 'perundang', 'undangan']	['penyaluran', 'dana', 'bagi', 'hasil', 'pbb', 'bphtb', 'sesuai', 'peraturan', 'perundang', 'undangan']
---	---

### 3.5.5. Lemmatizing

*Lemmatizing* adalah proses mengubah kata berimbuhan menjadi kata dasar, yang bertujuan untuk infleksi kata yang tidak diperlukan, sehingga kata yang memiliki akar yang sama dapat dikenali sebagai objek yang sama dengan memperhatikan konteks agar menjadi kata yang bermakna. Proses *lemmatizing* hanya dilakukan untuk analisis *Topic Modeling* karena penambahan imbuhan bisa memberi pemahaman kontekstual yang berbeda sehingga perlu dipertahankan untuk membandingkan keselarasan dua ayat. Contoh hasil *lemmatizing* ditunjukkan pada tabel berikut.

**Tabel 7** Contoh hasil *lemmatizing*

Data sebelum <i>lemmatization</i>	Data setelah <i>lemmatization</i>
['bank', 'wajib', 'memberikan', 'keterangan', 'pasal', '42', 'pasal', '43']	['bank', 'wajib', 'beri', 'keterangan', 'pasal', '42', 'pasal', '43']
['penyaluran', 'dana', 'bagi', 'hasil', 'pbb', 'bphtb', 'ayat', '3', 'ayat', '4', 'dilakukan', 'sesuai', 'peraturan', 'perundang', 'undangan']	['salur', 'dana', 'bagi', 'hasil', 'pbb', 'bphtb', 'ayat', '3', 'ayat', '4', 'laku', 'sesuai', 'peraturan', 'undang', 'undang']

### 3.6 Topic Modeling: Latent Dirichlet Allocation (LDA)

*Latent Dirichlet Allocation* (LDA) adalah salah satu teknik dalam analisis teks yang populer untuk melakukan *topic modeling*. Ide utama dari algoritma ini adalah menganggap bahwa setiap dokumen terdiri dari beberapa topik, dan setiap topik memiliki kata-kata yang berhubungan atau berkorelasi. Namun, topik-topik ini tidak dapat diamati secara

langsung dan disebut sebagai "*latent*" karena topik yang dibahas dalam setiap dokumen tidak dapat diketahui secara eksplisit. Oleh karena itu, LDA berfungsi untuk mengungkapkan topik-topik tersebut dari distribusi kata-kata dalam korpus.

LDA mencari representasi dari  $k$  topik ke dalam distribusi multinomial dari kumpulan kata-kata (Blei *et al.* 2018). Proses pelatihan LDA dimulai dengan menentukan jumlah topik yang diinginkan sebelumnya. Selanjutnya, LDA akan menghitung probabilitas distribusi kata-kata dalam setiap topik dan probabilitas distribusi topik dalam setiap dokumen melalui iterasi. Proses ini berfokus pada mengoptimalkan nilai-nilai probabilitas berdasarkan data yang ada. Setelah selesai proses pelatihan, LDA akan mengasosiasikan setiap kata dalam setiap dokumen dengan salah satu topik berdasarkan probabilitasnya. Sebagai hasilnya, dokumen-dokumen yang membahas topik yang serupa akan tergabung dalam kelompok yang sama, yang mempermudah analisis dan pemahaman terhadap isi dari korpus teks tersebut.

Dalam LDA, terdapat dua parameter yang diprediksi, yaitu  $\phi$  (*parts-versus-topics*) dan  $\theta$  (*composites-versus-topics*). Parameter  $\phi$  dan  $\theta$  dihitung dari parameter  $\alpha$  (*document-topic density*) dan  $\beta$  (*topic-word density*). Metode ini memungkinkan kita untuk menemukan pola-pola topik dalam korpus teks tanpa adanya label topik yang eksplisit sebelumnya. Dengan LDA, kita dapat memahami topik-topik apa saja yang muncul dalam suatu korpus teks dan bagaimana kata-kata tersebut berkorelasi dengan setiap topik, sehingga dapat membantu dalam analisis teks dan pemahaman lebih mendalam tentang isi dari dokumen-dokumen yang ada (Wu 2012).

### 3.7 Model Natural Language Inference (NLI) IndoBERT

*Natural Language Inference* (NLI) adalah metode dalam pemrosesan bahasa alami (NLP) yang bertujuan menentukan hubungan antara dua kalimat: kalimat asal (*premise*) dan kalimat hipotesis (*hypothesis*). Konsep dasar metode ini adalah menemukan apakah hipotesis dapat dianggap sebagai "benar" (*entailment*), "salah" (*contradiction*), atau "tidak dapat dikonfirmasi" (*neutral*) berdasarkan informasi yang terkandung dalam kalimat asal. Model NLI bertujuan untuk mengidentifikasi hubungan semantik antara dua kalimat dan menilai apakah kalimat hipotesis memiliki implikasi yang sesuai dengan kalimat asal.

Salah satu model bahasa yang dapat digunakan dalam pemrosesan NLI adalah IndoBERT. Konsep IndoBERT dalam metode NLI adalah mengenali hubungan logis antara

dua kalimat, yaitu premis (*premise*) dan hipotesis (*hypothesis*). Dalam pendekatan NLI, IndoBERT dilatih dengan menggunakan dataset yang berisi pasangan premis dan hipotesis yang telah diberi label untuk salah satu dari tiga kategori: *entailment*, *contradiction*, atau *neutral*. Proses *pre-training* IndoBERT telah memberikan model pemahaman yang mendalam tentang bahasa Indonesia secara umum, dan *fine-tuning* pada pemrosesan NLI akan mengkustomisasi model untuk memahami hubungan logis antara premis dan hipotesis.

Selama tahap *fine-tuning*, model dilatih untuk mengenali konteks dan implikasi antara premis dan hipotesis dalam berbagai kalimat. Model belajar mengenali pola dan fitur yang mengindikasikan apakah hipotesis dapat disimpulkan dari premis (*entailment*), bertentangan dengan premis (*contradiction*), atau tidak ada hubungan implikasi yang jelas (*neutral*). Dengan *fine-tuning* pada pemrosesan NLI, IndoBERT dapat digunakan untuk melakukan analisis dan pemahaman lebih lanjut pada teks-teks dalam bahasa Indonesia. Pendekatan *fine-tuning* pada IndoBERT untuk pemrosesan NLI memungkinkan model untuk menguasai representasi konteks dan relasi antar kalimat dengan lebih baik, sehingga memberikan performa yang lebih baik dalam tugas analisis teks dan pemrosesan bahasa alami dalam konteks analisis implikasi dan kontradiksi dalam bahasa Indonesia (Mahendra *et al.* 2021).

## IV. PEMBAHASAN

### 4.1. Ekstraksi Data

Pada tahapan ekstraksi data, data telah berhasil diubah menjadi sebuah *data frame* yang memudahkan analisis lebih lanjut. *Data frame* tersebut memiliki beberapa kolom, yaitu "Jenis Peraturan", "Nomor Peraturan", "Tahun Peraturan", "Pasal", "Ayat", dan "Teks". Kolom "Teks" berisi ayat yang terdapat dalam setiap pasal undang-undang dan merupakan kolom yang menjadi fokus *input* pada tahapan analisis data selanjutnya. Selanjutnya, data ini akan digunakan untuk analisis lebih lanjut, termasuk penerapan teknik *topic modeling* dan NLI untuk mendeteksi potensi disharmoni pada peraturan perundang-undangan di bidang investasi.

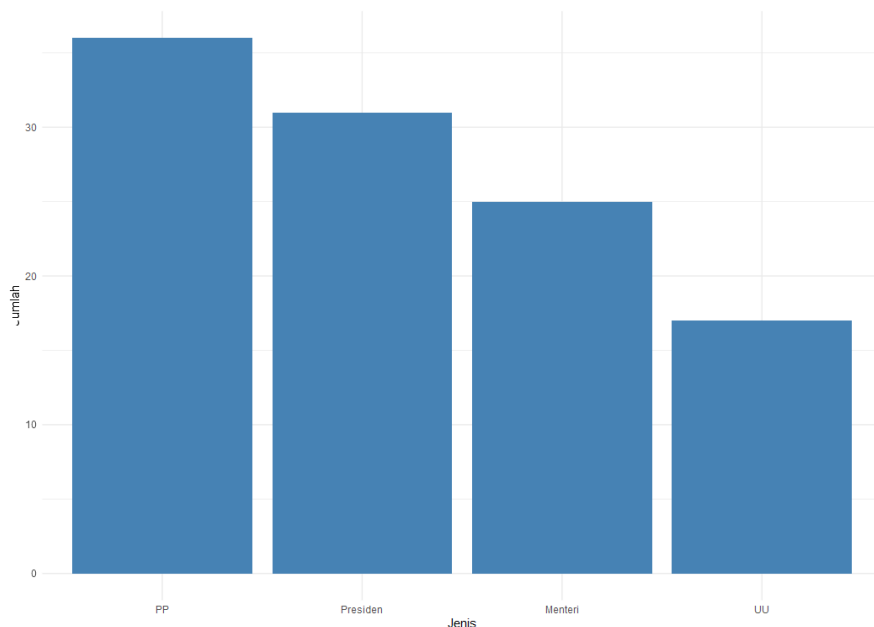
Pada seluruh dokumen perundang-undangan yang digunakan, terdapat enam dokumen dengan format berbeda yang harus diekstraksi dengan khusus, di mana tulisan pada dokumen tersebut terdeteksi sebagai gambar sehingga memerlukan tahapan ekstraksi

menggunakan *Optical Character Recognition* (OCR) sebelum diekstraksi ke dalam bentuk *data frame* Python. 6 dokumen perundang-undangan tersebut adalah sebagai berikut:

1. Instruksi Presiden Nomor 2 Tahun 2021
2. Instruksi Presiden Nomor 6 Tahun 2016
3. Keputusan Menteri Perhubungan Nomor 45 Tahun 2015
4. Peraturan Presiden Nomor 49 Tahun 2011
5. Peraturan Menteri Keuangan Nomor 103 Tahun 2019
6. Peraturan Pemerintah Nomor 33 Tahun 2013

#### 4.2. Eksplorasi Data

Berdasarkan Gambar 2 terlihat bahwa dari total 110 dokumen yang dianalisis, sebagian besar dokumen adalah Peraturan Pemerintah dengan jumlah sebanyak 36 dokumen. Sedangkan jumlah dokumen Undang-Undang paling sedikit, hanya sebanyak 17 dokumen. Dari persebaran ini, dapat disimpulkan bahwa Peraturan Pemerintah merupakan jenis peraturan yang paling dominan dalam korpus dokumen yang dianalisis, diikuti oleh Peraturan Presiden dan Peraturan Menteri dengan jumlah yang hampir sama, sementara Undang-Undang memiliki jumlah yang paling sedikit. Gambar di bawah juga menunjukkan potensi adanya kebutuhan untuk lebih fokus pada analisis dan pemahaman Peraturan Pemerintah karena dominasi jumlah dokumen tersebut.

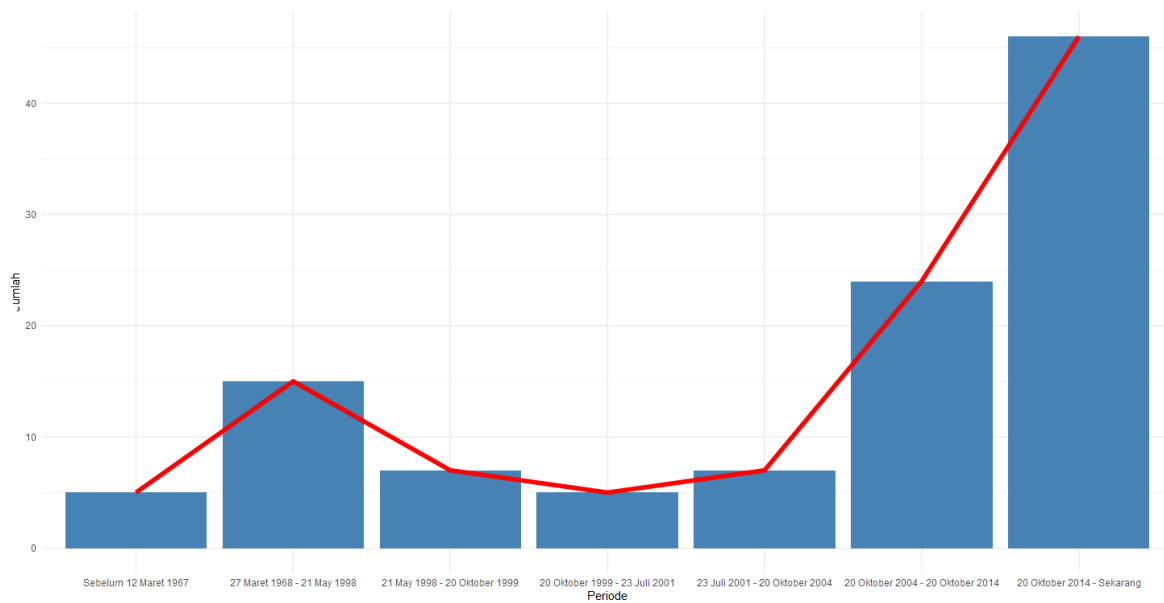


Gambar 2 *Barchart* persebaran jumlah dokumen perundang-undangan

Gambar 3 di bawah menunjukkan bahwa pada periode Oktober 2014 sampai 2022 (periode Presiden Joko Widodo (Jokowi)) jumlah dokumen perundang-undangan pada bidang investasi cenderung lebih banyak dibandingkan periode tahun lainnya. Dari Gambar 3 yang menunjukkan persebaran tahun rilis dokumen perundang-undangan yang digunakan pada penelitian ini juga terdapat beberapa *insight* yang dapat diperoleh yaitu:

1. Perubahan Kebijakan pada Masa Pemerintahan Tertentu: Terdapat pola dalam persebaran dokumen perundang-undangan berdasarkan tahun rilis yang terkait dengan periode presiden. Hal ini menunjukkan adanya perubahan kebijakan atau fokus dalam pembuatan regulasi cukup berkaitan dengan kepemimpinan presiden pada periode tertentu.
2. Fokus pada Era Tertentu: Jumlah dokumen perundang-undangan yang lebih tinggi pada Periode Jokowi dapat menunjukkan bahwa pemerintahan lebih berfokus pada bidang investasi pada era tersebut, atau mungkin ada peristiwa atau kejadian penting yang mempengaruhi pembuatan regulasi dalam bidang tersebut. Misalnya, memasuki periode kedua, Presiden Jokowi meluncurkan gagasan pembentukan *Omnibus Law* yang sangat revolusioner, sekaligus juga kontroversial, yaitu pada Undang-Undang Cipta Kerja (Undang Undang Nomor 11 Tahun 2020).
3. Konsistensi Regulasi: Terdapat distribusi dokumen perundang-undangan yang merata di beberapa periode, yaitu periode BJ. Habibie, Gus Dur, dan Megawati. Hal ini dapat menunjukkan adanya konsistensi dalam pembuatan peraturan di bidang investasi di berbagai periode pemerintahan.
4. Potensi Evaluasi Kebijakan: *Barchart* di bawah juga dapat digunakan sebagai dasar untuk melakukan evaluasi kebijakan hukum di masa lalu, terutama untuk menilai efektivitas regulasi yang telah diberlakukan pada periode-presiden tertentu dan apakah peraturan tersebut tetap relevan atau memerlukan revisi.





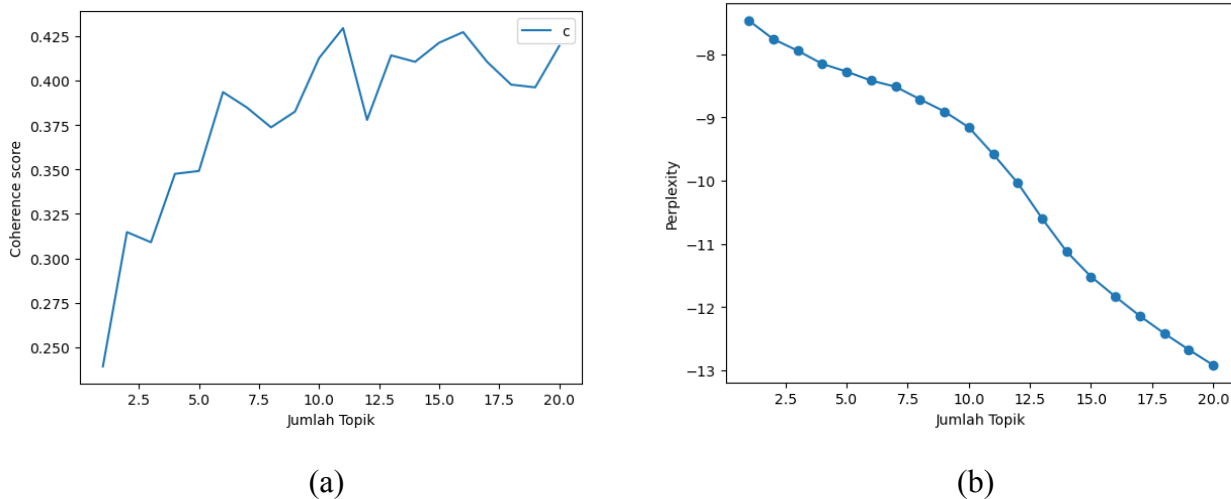
Gambar 3 *Barchart* persebaran tahun rilis dokumen perundang-undangan

#### 4.3. *Topic Modeling* dan *Clustering*

*Topic Modeling* bertujuan untuk mengidentifikasi topik-topik utama yang terdapat dalam seluruh perundang-undangan yang digunakan sebagai data *input*. Dalam penelitian ini, akan digunakan modul *gensim* pada Python untuk melakukan *Topic Modeling* pada seluruh dokumen peraturan perundang-undangan, sehingga dokumen-dokumen tersebut dapat dikelompokkan ke dalam beberapa kelompok berdasarkan topik yang dibahasnya.

Proses *Topic Modeling* dimulai dengan pemrosesan teks, seperti tokenisasi, pembersihan, dan pembentukan kamus kata-kata. Setelah itu, akan digunakan algoritma LDA (*Latent Dirichlet Allocation*) untuk memodelkan topik-topik yang tersirat dalam seluruh dokumen tersebut. Sebelum membuat model, perlu dilakukan penentuan jumlah topik yang optimal. Untuk itu, akan digunakan dua metode evaluasi, yaitu *perplexity* dan *coherence score*. *Perplexity* adalah nilai *normalized log-likelihood* yang digunakan untuk mengukur seberapa baik model dapat menggeneralisasi data yang belum pernah dilihat sebelumnya atau data *train*. Semakin rendah nilai *perplexity*, semakin baik model dalam menggeneralisasi data. Sementara itu, *coherence score* digunakan untuk mengukur kualitas topik yang dihasilkan oleh model. *Coherence score* mengukur seberapa mudah suatu topik dapat dipahami dan ditafsirkan oleh manusia. Semakin tinggi nilai *coherence score*, semakin baik kualitas topik yang dihasilkan oleh model.

Dengan menentukan jumlah topik yang optimal berdasarkan nilai *perplexity* dan *coherence score*, kita dapat membagi seluruh dokumen peraturan perundang-undangan ke dalam kelompok-kelompok yang sesuai dengan topik-topik utama yang dibahasnya. Hal ini akan mempermudah analisis dan perbandingan antar ayat yang sejenis, sehingga isi dari dokumen-dokumen peraturan perundang-undangan tersebut dapat dipahami dengan lebih efisien dan efektif.

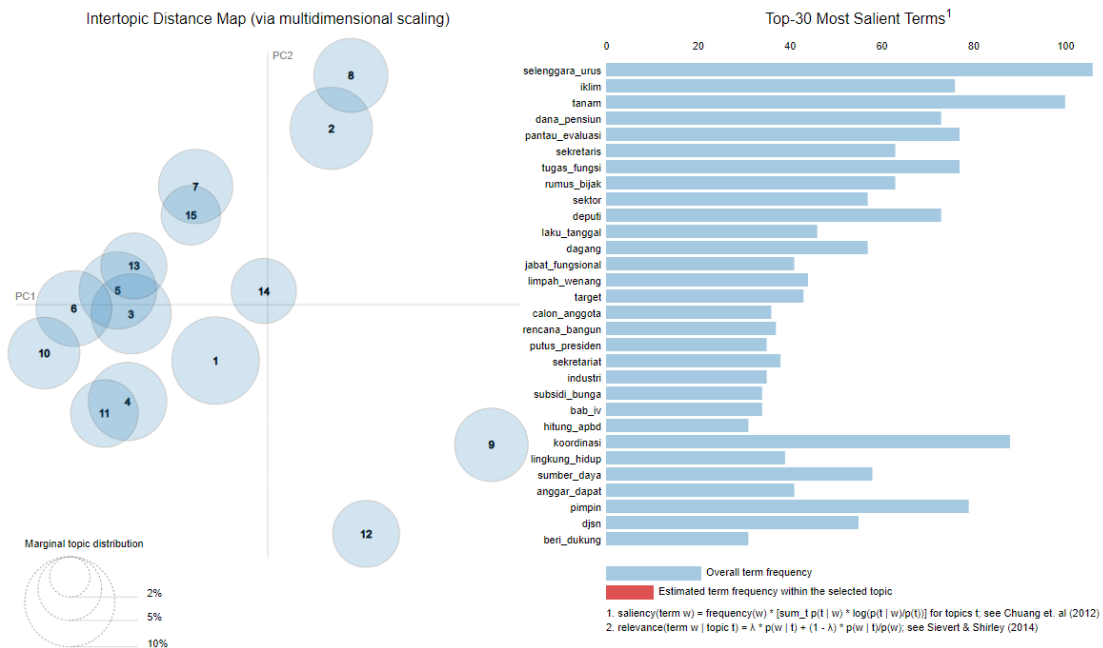


**Gambar 4** Nilai evaluasi metrik modeling LDA  
*coherence score* (a) dan *perplexity* (b)

Gambar 4 menunjukkan bahwa terdapat hubungan antara jumlah topik yang digunakan dalam model dengan nilai *coherence score* dan *perplexity*. Seiring meningkatnya jumlah topik, nilai *coherence score* dan *perplexity* cenderung meningkat hingga mencapai suatu titik optimum. Terlihat bahwa titik optimum diperoleh ketika jumlah topik adalah 15, di mana pada titik ini nilai *coherence score* mencapai 0,42 dan nilai *perplexity* mencapai -11,51. Pada titik ini, model memiliki keseimbangan yang baik antara kemampuan untuk menghasilkan topik-topik yang mudah dipahami dan diinterpretasikan oleh manusia (dengan nilai *coherence score* yang tinggi) serta kemampuan untuk menggeneralisasi data yang belum pernah dilihat sebelumnya dengan baik (dengan nilai *perplexity* yang rendah).

Dengan menggunakan jumlah topik sebanyak 15, model dapat mengidentifikasi dan memodelkan topik-topik utama secara efektif yang terdapat dalam seluruh dokumen peraturan perundang-undangan. Topik-topik ini dapat digunakan untuk mengelompokkan

dokumen-dokumen yang membahas hal yang serupa, sehingga mempermudah analisis dan perbandingan antar ayat yang memiliki konteks yang sama. Sebagai hasilnya, model dapat memberikan wawasan yang lebih mendalam tentang isi dan struktur dari dokumen peraturan perundang-undangan tersebut.



**Gambar 5** Jarak antar topik dan 30 kata yang paling menonjol pada dokumen

Penjelasan mengenai hubungan antara *cluster*-topik berdasarkan informasi yang diberikan adalah sebagai berikut.

#### 1. *Cluster* 1 dan 14: Teknis dan Evaluasi

Kedua *cluster* ini berdekatan karena membahas topik yang serupa, yaitu masalah teknis dan evaluasi. Topik ini mungkin melibatkan metode, proses, atau alat yang digunakan untuk mengevaluasi sesuatu, seperti proyek atau program.

#### 2. *Cluster* 12 dan 9: Perangkat Pemerintahan dan Iklim

*Cluster* 12 dan 9 berjarak cukup jauh karena membahas topik yang berbeda. *Cluster* 12 berfokus pada perangkat pemerintahan seperti pejabat, petugas, dan calon anggota, sedangkan *cluster* 9 membahas tentang iklim, yang mungkin melibatkan perubahan iklim dan dampaknya.

#### 3. *Cluster* 2 dan 8: Industri, Sumber Daya, dan Lingkungan Hidup

Kedua cluster ini bersinggungan karena mengandung topik yang serupa, yakni industri, sumber daya, dan lingkungan hidup. Topik ini mungkin membahas tentang dampak industri terhadap lingkungan dan pengelolaan sumber daya secara berkelanjutan.

4. *Cluster* 10, 6, 5, 13, dan 3: Investasi, Modal, Kredit, dan Perbankan

*Cluster* 10, 6, 5, 13, dan 3 berdekatan karena memiliki fokus yang mirip, yaitu berbicara tentang investasi, modal, kredit, dan perbankan. Topik-topik ini berkaitan dengan aspek keuangan dan ekonomi yang melibatkan investasi dan sumber modal, termasuk kredit dari perbankan.

5. *Cluster* 7 dan 15: Kebijakan Modal, Subsidi, dan Laut

*Cluster* 7 dan 15 bertumpuk karena memiliki topik-topik yang berhubungan erat. *Cluster* 7 mungkin membahas tentang kebijakan modal, yang mencakup keputusan mengenai penggunaan dana dan investasi. Sementara itu, *cluster* 15 membahas tentang subsidi dan laut, yang mungkin berkaitan dengan kebijakan pemerintah dalam memberikan subsidi dan isu-isu laut, seperti keberlanjutan perikanan.

6. *Cluster* 4 dan 11: Wewenang, Penggunaan Anggaran, Akuntansi, dan Inventarisasi

*Cluster* 4 dan 11 bertumpuk karena berisi topik yang terkait erat. *Cluster* 4 berfokus pada aspek wewenang dan penggunaan anggaran, mungkin membahas tentang bagaimana anggaran digunakan dan dikelola oleh entitas tertentu. Di sisi lain, *cluster* 11 membahas tentang akuntansi dan inventarisasi, yang berhubungan dengan pencatatan dan pengelolaan aset dan barang inventaris.

#### 4.4. *Tuning Pre-Trained Model*

Model yang digunakan untuk mengklasifikasi dalam penelitian ini adalah model IndoBERT-large-p2 yang dikembangkan oleh indobenchmark. Model ini merupakan varian dari model BERT yang diadaptasi khusus untuk bahasa Indonesia. Untuk mendapatkan model ini, perlu dilakukan proses *training* ulang menggunakan dataset Indo4B yang terdiri dari 23,43 GB dokumen, dengan total 4 miliar kata bahasa Indonesia yang berasal dari 15 sumber yang berbeda. Model ini awalnya didesain untuk analisis *Next Sentence Prediction* (NSP), tetapi dalam penelitian ini akan digunakan untuk analisis *Natural Language Inference* (NLI).

Model IndoBERT-large-p2 kemudian di-*training* ulang menggunakan dataset IndoNLI untuk menyesuaikan model dengan kebutuhan analisis NLI. Dataset IndoNLI dikhususkan untuk pemrosesan NLI, yaitu mengklasifikasikan hubungan antara dua teks,

yaitu premis dan hipotesis (suatu pasangan ayat). Dataset ini terdiri dari 10 ribu data *train* dan 2 ribu data *test* yang dilabeli oleh ahli bahasa, sehingga dapat diandalkan untuk tujuan analisis NLI.

Dengan melakukan *training* ulang menggunakan dataset IndoNLI, model IndoBERT-large-p2 dapat menjadi lebih akurat dan efektif dalam melakukan klasifikasi hubungan antara dua teks, khususnya pada tahap pendeteksian potensi disharmoni pada pasangan ayat dalam *data frame*. Proses *training* ulang ini memungkinkan model untuk lebih baik dalam mengevaluasi implikasi dan kontradiksi antara ayat-ayat dalam dokumen perundang-undangan, sehingga potensi disharmoni dapat dideteksi dengan lebih baik dan dapat mendukung analisis yang lebih mendalam dalam konteks regulasi investasi. Adapun konfigurasi *hyperparameter* yang digunakan dalam *training* model dapat dilihat pada Tabel 8.

**Tabel 8** *Hyperparameter training model*

Hyperparameter	Nilai
Batch Size	32
Epochs	10
Learning Rate	1e-5
Optimizer	AdamW
Warmup Ratio	0,1
Weight Decay	0,01

Dari hasil *training* yang telah dilakukan, model yang digunakan telah berhasil mencapai tingkat akurasi sebesar 0.765 dalam melakukan klasifikasi. Nilai akurasi tersebut menunjukkan seberapa baik model dapat mengklasifikasikan hubungan antar pasangan teks. Tingkat akurasi sebesar 0.765 dapat dianggap sebagai hasil yang cukup baik, terutama mengingat kompleksitas proses NLI dalam mendeteksi implikasi dan potensi kontradiksi. Meskipun nilai akurasi tersebut sudah mencapai tingkat yang dapat diandalkan, masih ada ruang untuk peningkatan performa model melalui *fine-tuning* dan optimasi lebih lanjut.

#### 4.4. Prediksi Label

Prediksi label dilakukan dengan memasangkan dua ayat yang berada pada satu *cluster*, kemudian di antara ayat tersebut akan dihitung jarak *Euclidean* menggunakan vektor yang berisi probabilitas teks tersebut untuk masuk ke dalam suatu topik, sehingga masing-masing teks akan memiliki vektor dengan 15 elemen. Hal ini dilakukan untuk memperkecil ruang pencarian sehingga hanya ayat-ayat yang memiliki nilai similaritas yang tinggi yang akan dibandingkan.

Tabel 9 Contoh hasil prediksi label

No.	Teks Ayat 1	Teks Ayat 2	Label	Prob
1.	penyertaan modal adalah bentuk investasi pemerintah pada badan usaha dengan mendapat hak kepemilikan termasuk pendirian perseroan terbatas dan atau pengambilalihan perseroan terbatas	penyertaan modal adalah bentuk investasi pemerintah pada badan usaha dengan mendapat hak kepemilikan	<i>entailment</i>	0.844
2.	dalam penyelenggaraan penelitian dan pengembangan pendidikan dan latihan serta penyuluhan kehutanan wajib memperhatikan ilmu pengetahuan dan teknologi kearifan tradisional serta kondisi sosial budaya masyarakat	pendidikan dan latihan kehutanan bertujuan untuk membentuk sumber daya manusia yang menguasai serta mampu memanfaatkan dan mengembangkan ilmu pengetahuan dan teknologi dalam pengurusan hutan secara adil dan lestari didasari iman dan taqwa kepada tuhan yang maha esa	<i>entailment</i>	0.755
3.	dana operasional yang dapat diambil dari dana jaminan sosial ketenagakerjaan sebagaimana dimaksud dalam pasal 12 ayat 5 ditetapkan paling tinggi a 10 sepuluh persen dari iuran jaminan kecelakaan	dana operasional yang dapat diambil dari dana jaminan sosial ketenagakerjaan sebagaimana dimaksud dalam pasal 12 ayat 5 ditetapkan paling tinggi a 10 sepuluh persen dari Iuran jaminan kecelakaan	<i>contradiction</i>	0.679

	kerja dan jaminan kematian dan b 10 sepuluh persen dari Iuran yang diterima dan 10 sepuluh persen dari dana hasil pengembangan jaminan hari tua dan jaminan pensiun	kerja dan jaminan kematian dan b 2 dua persen dari akumulasi iuran dan dana hasil pengembangan jaminan hari tua		
4.	dalam penentuan persentase dana operasional sebagaimana dimaksud dalam pasal 13 ayat 2 bpjs ketenagakerjaan mengajukan usulan besaran persentase dana operasional kepada menteri paling lambat 3 tiga bulan sebelum tahun anggaran yang bersangkutan dengan melampirkan rencana kerja anggaran tahunan bpjs ketenagakerjaan	dalam penentuan persentase dana operasional sebagaimana dimaksud dalam pasal 13 ayat 2 bpjs ketenagakerjaan mengajukan usulan besaran persentase dana operasional kepada menteri dalam jangka waktu paling lambat 4 empat bulan sebelum tahun anggaran yang bersangkutan dengan melampirkan rancangan rencana kerja anggaran tahunan bpjs ketenagakerjaan	<i>contradiction</i>	0.728
5.	dana tapera sebagaimana dimaksud dalam pasal 60 huruf bersumber dari hasil penghimpunan simpanan peserta hasil pemupukan simpanan peserta hasil pengembalian kredit pembiayaan dari peserta hasil pengalihan aset tabungan perumahan pegawai negeri sipil yang dikelola oleh badan pertimbangan tabungan perumahan pegawai negeri sipil dana wakaf dan dana lainnya yang sah sesuai dengan ketentuan peraturan perundang undangan	bentuk dan isi laporan pengelolaan program sebagaimana dimaksud pada ayat 1 diusulkan oleh bp tapera setelah berkonsultasi dengan komite tapera 4 laporan keuangan bp tapera sebagaimana dimaksud pada ayat 1 disusun dan disajikan sesuai dengan standar akuntansi keuangan yang berlaku	<i>neutral</i>	0.549

**Keterangan sumber dokumen perundang-undangan:**

1. Teks 1: Peraturan Pemerintah No. 49 Tahun 2011  
Teks 2: Peraturan Pemerintah No. 1 Tahun 2008
2. Teks 1 dan 2: Undang-Undang No. 41 Tahun 1999
3. Teks 1: Peraturan Pemerintah No. 55 Tahun 2015  
Teks 2: Peraturan Pemerintah No. 99 Tahun 2013
4. Teks 1: Peraturan Pemerintah No. 55 Tahun 2015  
Teks 2: Peraturan Pemerintah No. 99 Tahun 2013
5. Teks 1 dan 2: Undang-Undang No. 4 Tahun 2016

Tabel 9 berisi contoh hasil prediksi label untuk beberapa pasang teks ayat. Pada data pertama (No. 1), teks ayat kedua secara implisit mengandung informasi yang ada di teks ayat pertama, sehingga diprediksi dengan label "*entailment*" (terkandung) dan mendapat nilai *probability* sangat tinggi dari model. Pada data kedua (No. 2), teks ayat pertama dan teks ayat kedua juga terkait erat karena mengandung informasi yang saling berhubungan, dan oleh karena itu, diprediksi dengan label "*entailment*." Data ketiga (No. 3), menunjukkan kontradiksi antara teks ayat pertama dan teks ayat kedua karena menyajikan persentase yang berbeda dalam pengambilan dana operasional, sehingga diprediksi dengan label "*contradiction*" (kontradiksi) dari model. Sementara itu, pada data keempat (No. 4) juga terjadi kontradiksi karena perbedaan waktu pengajuan usulan persentase dana operasional, sehingga diberi label "*contradiction*". Pada data kelima (No. 5), teks ayat pertama dan teks ayat kedua tidak memiliki keterkaitan yang erat, sehingga diberi label "*neutral*" (netral) karena membahas topik yang berbeda meskipun masih berkaitan dengan Tapera.

Tabel 9 juga menunjukkan adanya perubahan atau revisi perundang-undangan pada periode kepresidenan yang berbeda, khususnya terkait dengan Peraturan Pemerintah No. 99 Tahun 2013 dan Peraturan Pemerintah No. 55 Tahun 2015. Peraturan Pemerintah No. 99 Tahun 2013 dipublikasikan pada periode kepresidenan Presiden Susilo Bambang Yudhoyono (SBY), sementara Peraturan Pemerintah No. 55 Tahun 2015 direvisi dan dirilis pada periode kepresidenan Presiden Joko Widodo (Jokowi). Terlihat bahwa ayat pada tabel di atas yang mengandung revisi (No. 3 dan No. 4) diberi label "*contradiction*" oleh model.



Hal ini berarti model mengenali bahwa terdapat perbedaan atau kontradiksi antara teks ayat yang satu dengan yang lainnya.

Revisi perundang-undangan merupakan hal yang umum terjadi untuk mengatasi permasalahan atau menyesuaikan dengan perkembangan dan kebutuhan yang berkembang di masyarakat. Dalam konteks perundang-undangan di Indonesia, revisi biasanya dilakukan untuk meningkatkan kejelasan, efisiensi, dan relevansi peraturan guna mencapai tujuan tertentu, seperti perbandingan pada data pertama yaitu ayat pada Peraturan Pemerintah No. 49 Tahun 2011 memperjelas secara lebih rinci Peraturan Pemerintah No. 1 Tahun 2008. Pentingnya perubahan atau revisi perundang-undangan ini menegaskan pentingnya pemahaman yang tepat dan akurat mengenai isi peraturan yang berlaku. Oleh karena itu, penggunaan model prediksi seperti yang tercantum dalam Tabel 9 dapat membantu untuk mengidentifikasi adanya potensi kontradiksi atau perbedaan dalam interpretasi antara berbagai teks ayat peraturan, dan memberikan panduan bagi pembuat kebijakan untuk mengatasi permasalahan tersebut.

## **V. PENUTUP**

### **5.1 Kesimpulan**

Pemodelan klasifikasi menggunakan teknik *machine learning* dapat diaplikasikan untuk mempermudah pengidentifikasian ketidakselarasan dalam perundang-undangan. Metode *Topic Modelling* dapat mereduksi ruang pencarian dimana hanya peraturan yang membahas hal yang sama saja yang dibandingkan, hal ini dapat mereduksi kemungkinan pasangan ayat yang awalnya berjumlah jutaan kombinasi menjadi ribuan kombinasi. Selanjutnya Model IndoBERT dapat memberikan dugaan hubungan antara dua peraturan, sehingga dapat mempermudah para ahli dalam melakukan pemeriksaan ulang, dimana peraturan-peraturan yang dilabeli “*contradiction*” memiliki potensi yang besar untuk terjadi ketidakselarasan atau disharmoni.

### **5.2 Saran**

Saran dan rekomendasi yang dapat diberikan untuk pemerintah maupun peneliti selanjutnya berdasarkan hasil penelitian di atas adalah sebagai berikut.

1. Pemanfaatan Metode *Topic Modeling*: Hasil pembagian beberapa topik utama pada dokumen perundang-undangan bidang investasi yang baik pada penelitian ini

menunjukkan bahwa peneliti selanjutnya dapat menggunakan metode *Topic Modeling* LDA dalam berbagai domain hukum, ekonomi, kesehatan, dan lainnya untuk mengidentifikasi pola-pola dan topik-topik penting dalam dokumen teks yang besar dan kompleks.

2. *Fine-tuning* IndoBERT untuk NLI: Pemerintah dan peneliti dapat melanjutkan penelitian dengan *fine-tuning* model IndoBERT menggunakan dataset NLI bahasa Indonesia lainnya untuk meningkatkan performa model dalam pemrosesan *Natural Language Inference* (NLI). Dengan *fine-tuning* yang lebih spesifik, model dapat lebih optimal dalam mengevaluasi hubungan implikasi dan kontradiksi antar kalimat atau ayat dalam berbagai konteks teks.

3. Integrasi dengan Sistem Hukum Digital: Pemerintah dapat mempertimbangkan integrasi hasil penelitian ini dengan sistem hukum digital, seperti *platform* penelusuran hukum dan *database* peraturan perundang-undangan, sehingga dapat memberikan kemudahan akses dan analisis yang lebih efisien bagi para pengguna sistem hukum.

4. Pengembangan Aplikasi Pendeteksi Disharmoni: Berdasarkan hasil analisis potensi disharmoni antar ayat yang dihasilkan, peneliti atau pemerintah dapat mengembangkan aplikasi atau alat bantu berbasis teknologi yang dapat secara otomatis mendeteksi disharmoni antar pasal atau ayat dalam dokumen peraturan perundang-undangan. Hal ini dapat membantu para pengguna hukum untuk lebih mudah dan cepat mengidentifikasi potensi konflik atau ketidaksesuaian dalam aturan-aturan yang ada.

5. Kolaborasi dengan Para Ahli Hukum: Pemerintah atau peneliti dapat bekerja sama dengan para ahli hukum dan praktisi di bidang investasi untuk memvalidasi hasil analisis potensi disharmoni dan memberikan perspektif hukum yang lebih mendalam terkait hasil temuan. Kolaborasi ini akan memberikan kesempatan untuk menguji kebenaran dan relevansi hasil penelitian dalam konteks praktik hukum yang sebenarnya.

Dengan mengimplementasikan saran-saran di atas, diharapkan penelitian dan penggunaan teknologi pemrosesan bahasa alami, seperti *Topic Modeling* dan NLI, akan semakin berkembang dan memberikan kontribusi yang signifikan bagi kebijakan dan praktik hukum dalam berbagai bidang di Indonesia. Selain itu, adopsi ide dan transfer gagasan dari penelitian ini dapat berpotensi menjadi model bagi negara-negara lain yang ingin mengadopsi teknologi dalam analisis dan pemahaman peraturan perundang-undangan.

## VI. REFERENSI

- Blei DM, Ng AY, Jordan MI. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.
- Indrati MF. 2007. *Ilmu Perundang-Undangan*. Yogyakarta (ID): Kanisius.
- Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, Zhao L. 2018. Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*. 78(11).
- Kelsen H. 1961. *General Theory of Law and State*. New York (AS): Russell & Russell.
- [KEMENKUMHAM RI] Kementerian Hukum dan HAM RI. 2020. Analisis dan Evaluasi Peraturan Perundang-Undangan. Jakarta (ID): Pusat Analisis dan Evaluasi Hukum Nasional, Kementerian Hukum dan HAM RI.
- Laban P, Schnabel T, Paul N. Bennett, Marti A. Hearst. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics* 2022. 10:163–177. doi: [https://doi.org/10.1162/tacl\\_a\\_00453](https://doi.org/10.1162/tacl_a_00453).
- Liu NF, Schwartz R, Smith NA. 2019. *Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets*. Minneapolis (AS): NAACL.
- Mahendra R, Aji AF, Louvan S, Rahman F, dan Vania C. 2021. *IndoNLI: A Natural Language Inference Dataset for Indonesian*. Scotland (UK): Cornell University.
- Suriadinata V. 2019. Penyusunan undang-undang di bidang investasi: kajian pembentukan omnibus law di Indonesia. *Jurnal Ilmu Hukum*. 4(1):115–134.
- Wu J. 2012. *Advances in K-means Clustering: A Data Mining Thinking*. Berlin (DE): Springer Publishing Company.