Journal of Information Systems and Informatics
Vol. 6, No. 3, September 2024 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v6i3.861 Published By DRPM-UBD

Leveraging NLP to Analyze Regulatory Document Interconnections: A Systematic Review

Yudi Agusta[1], I Gusti Ayu Aprilia Santi[2], Ni Putu Putri Intan Maharani[3]

[1,2,3]Information System Department, Institute of Technology and Business STIKOM Bali, Indonesia
Email: [1]yudi@stikom-bali.ac.id, [2]ayu.apriliasanti2@gmail.com, [3]maharaniintan274@gmail.com

Abstract

A sustainable digital village requires an effective policy management mechanism to deliver relevant regulatory information to the community. Management information systems for regulations play a crucial role in achieving this. However, communities still face challenges in understanding and navigating the relationships between various regulations. To address this issue, this study conducts a systematic review of the components found in regulatory documents and the methods used to analyze them. The review identifies eight key components in regulatory documents: topic, structure, category, initiator, level, considerations, related regulations, and content. Natural Language Processing (NLP) techniques can be employed for data preprocessing, including tokenization, lowercasing, stop word removal, stemming, filtering, part-of-speech tagging, lemmatization, and chunking. For feature extraction, methods such as TF-IDF, bag-of-words, WordCount, N-grams, and word embeddings can be applied. To measure the interconnection between regulations, techniques like cosine similarity and K-Means clustering can be utilized. Experimental results demonstrate that combining different methods significantly influences the accuracy of identifying regulatory interconnections. The choice of methods whether simple or complex depends on the context, and confirmation through manual analysis is often required to ensure accuracy.

Keywords: Regulatory Documents, Natural Language Processing, Text Mining, Systematic Literature Review

1. INTRODUCTION

Big Data has become a critical component in the process of knowledge formation, especially in decision-making processes. One significant source of big data in daily life is regulatory and legal documents. Regulatory documents refer to official rules or regulations issued by governments or competent authorities [1]. These documents serve to outline provisions, procedures, and guidelines to govern specific sectors or activities. However, as the volume and complexity of these documents grow, managing and analyzing them effectively presents a significant challenge [2].

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

The increasing volume and complexity of regulatory documents pose serious challenges to their efficient management. Regulatory documents are not homogeneous; they include diverse elements such as themes, considerations, reference regulations, regulatory content, and sectors. This diversity, coupled with their sheer volume, makes it difficult to ensure consistency and efficiency within regulatory frameworks. Without proper analysis methods, interrelationships between documents are often missed, leading to inefficiencies, inconsistencies, or even gaps in regulatory oversight. These issues can have far-reaching effects, potentially undermining decision-making processes and the implementation of policies [3].

To address these challenges, a method to analyze the interconnection between regulatory documents is needed. Such an analysis would enable stakeholders to better understand how different regulations relate to each other and could help improve the management, categorization, and utilization of these important resources. Effective document management would facilitate more streamlined policy implementation and decision-making by ensuring that regulatory documents are better organized and more easily navigable. This could help avoid redundancies and conflicts between regulations, ensuring a more coherent regulatory environment [4].

Recent advancements in information technology, particularly in natural language processing (NLP), offer promising solutions for regulatory document analysis [5]. NLP, a subset of big data analysis, has already been successfully applied in many areas of document analysis, providing significant advantages in handling large datasets and extracting meaningful patterns. Despite its success in general document analysis, NLP has not yet been widely applied to regulatory documents. The potential benefits of applying NLP to these types of documents, such as automating the identification of interconnections and improving the overall management process, have yet to be fully explored [6].

This paper aims to fill that gap by reviewing natural language-based document analysis techniques. It will examine the key components of regulatory documents and explore various NLP and text mining methods that can be used for data preprocessing and feature extraction. Additionally, the study will review methods for calculating interconnections between regulatory documents, providing insights into how these documents can be better managed. By doing so, this paper will contribute to the development of more effective tools for regulatory document management and lay a foundation for future research and practical applications in this field.

2. METHODS

Systematic Literature Review (SLR) is a method employed to systematically collect sources or data related to a specific research topic for further in-depth analysis of the content. This method enables researchers to critically assess the materials, ensuring a comprehensive understanding of the subject. Sources for SLR can be derived from various national and international journals [4].

2.1. Research Object

The object of this research comprises regulatory documents. The selection of regulatory manuscripts as the primary focus is based on several considerations:

1. Regulatory documents often exhibit a certain level of complexity that makes them difficult for the general public to easily comprehend.
2. These documents are interconnected both vertically and horizontally, meaning that understanding one may require consideration of others.
3. The volume of regulatory documents has been continuously increasing, adding to the challenge of their analysis.

2.2. Phase of the Research Method

The phases of the research method are outlined in Figure 1**.** It demonstrates the process of the Systematic Literature Review (SLR) as applied in this research. These stages are described in further detail as follows:

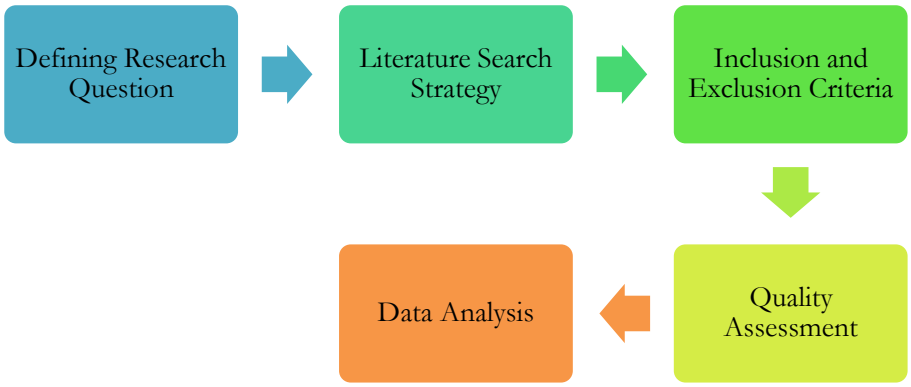

Figure 1. Systematic Literature Review Applied in The Research

The steps involved in conducting the Systematic Literature Review (SLR) are systematically depicted in Figure 1, which highlights the specific actions taken at each phase of the process. The following section will provide a detailed explanation of these steps.

2.2.1. Defining Research Question

In the Research Question (RQ), several research questions are set related to the research topic. The research questions investigated are:

RQ1. Which items or components in a regulatory document can be used to analyze the degree of interconnection between the regulatory documents?
RQ2. Which methods can be used to extract the regulatory document component and how are their capacities?
RQ3. What methods can be used to analyze the interconnection between regulatory documents and how are their capacities?

2.2.2. Literature Search Strategy

The literature search is conducted through the Google Scholar website. The process finds reliable sources that can answer research questions. Terms used for searching are related to the research questions including terms of "regulatory books", "methods of analysis of correlation", "natural language processing", "cosine similarity", "K-means", and "feature extraction".

2.2.3. Inclusion and Exclusion Criteria

A search process is carried out to find literature that is suitable to be used in this research. A study is eligible to be chosen if it meets the following criteria:

1. The data used is predominantly from the last 6 years, or around the years 2018-2024. Using references from the last 6 years helps ensure that the data used is still relevant, as newer references tend to be more accurate.
2. The data used are those that are related to regulatory documents, document text preprocessing, and document content feature extraction.

2.2.4. Quality Assessment

The Quality Assessment stages are carried out to identify the data being evaluated by adjusting the quality assessment criteria. The following are some questions that will serve as a reference in selecting relevant data are as follows:

QA1. Does literature consist of the components of the regulatory document that need to be used to analyze the interconnection between regulatory documents?
QA2. Does the literature contain a corresponding method used in extracting items or components of the regulatory document?
QA3. Does the literature contain a suitable method used in the development of methods for analyzing interconnection between regulatory documents?

2.2.5. Data Analysis

In this phase, the collected data will be analyzed to show that the data is suitable for the following:

1. Items or components of a regulatory document that are important to be used to analyze the interconnection between regulatory documents (refers to RQ1).
2. Methods that can be used to extract items or components of the regulatory document (refers to RQ2).
3. Methods that can be used to analyze the interconnection between regulatory documents (refers to RQ3).

3. RESULTS AND DISCUSSION

This section describes the process of searching for literature and discussing the content of literature used as a reference.

3.1. Search Process

Based on the search process through Google Scholar, 60 documents are found that can be reviewed further.

3.2. Inclusion and Exclusion Criteria Selection Results

The result of the search process will be selected based on inclusion and exclusion criteria. With this process, 42 documents are selected. Two journals do not meet the year criteria but meet the data criteria related to the regulatory document and the development of information systems and technology. The two documents are used to support the analysis. Table 1. displays the journals used as the source articles.

Table 1. The List the journal as a review source

No	Journal	Amount
1	Teknika	1
2	Techno.COM	1
3	e-Proceeding of Engineering	2
4	JINTEKS (Jurnal Informatika Teknologi dan Sains	1
5	Jurnal Teknik Elektro	1
6	Jurnal Komtika – Komputasi dan Informatika	1
7	Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer	2
8	Jurnal Privat Law	1
9	DINAMIKA	1
10	Jurnal Gema Keadilan	1

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

No	Journal	Amount
11	Lex Jurnalica	1
12	Jurnal Education and Development	1
13	ACADEMIA: Jurnal Inovasi Riset Akademik	1
14	Jurnal Konstitusi	1
15	Jurnal Komunikasi Nusantara	1
16	Masalah-Masalah Hukum	1
17	Lex Administratum	1
18	Res Publica	1
19	Jurnal Legislasi Indonesia	1
20	Jurnal Dialog	1
21	JURIKOM (Jurnal Riset Komputer)	1
22	Bina Darma Conference on Computer Science	1
23	Jurnal Informatika Polinema	1
24	Jurnal Teknologi Pintar	1
25	Jurnal Coding	1
26	ICON	1
27	JBegaTI	1
28	Jurnal Teknologi	1
29	Journal of Network and Computer Applications	1
30	Jurnal Jupiter	1
31	Citec Journal	1
32	Jurnal Tematika	1
33	Building of Informatics, Technology and Science (BITS)	1
34	IEEE/ACM Transactions on Audio, Speech, and Language Processing	1
35	Jurnal Teknologi dan Manajemen Informatika	1
36	IEEE Computational Intelligence Magazine	1
37	IT Journal Research and Development	1
38	Jurnal EKSPONENSIAL	1
39	Jurnal Teknik Informatika	1
40	Jurnal Linguistik Komputasional (JLK)	1

3.3. Quality Assessment Results

The research evaluated a total of 42 journals by applying inclusion and exclusion criteria to filter and assess relevant documents. The primary goal of this evaluation was to determine how well the documents addressed the research questions through a systematic quality assessment process, which was divided into three phases. Each phase corresponded to a specific research question, and the findings from these assessments are summarized below:

1. **Quality Assessment 1 (QA1)**
2. This phase evaluated the documents' relevance to Research Question 1 (RQ1). After applying the criteria, 13 out of the 42 journals were found to be related to RQ1. These documents provide valuable insights and data pertinent to the first research question, indicating that there is moderate availability of sources related to this aspect of the research.
3. **Quality Assessment 2 (QA2)**
   QA2 focused on evaluating the documents in relation to Research Question 2 (RQ2). This resulted in 22 relevant documents, which represents the highest number of relevant sources among the three assessments. The significant volume of literature related to RQ2 suggests that this particular research question covers a broader topic or area with more available scholarly resources.
4. **Quality Assessment 3 (QA3)**
   The final phase, QA3, assessed documents concerning Research Question 3 (RQ3). Only 7 documents were found to be relevant, which suggests that RQ3 either deals with a more niche topic or an area that is underexplored in existing literature. This could present an opportunity for further research in this particular domain.

Figure 2 provides a statistical breakdown of the number of journals associated with each research question based on the quality assessment review. It visually displays how the 42 selected journals are distributed across the three research questions, giving an at-a-glance understanding of the literature's focus.

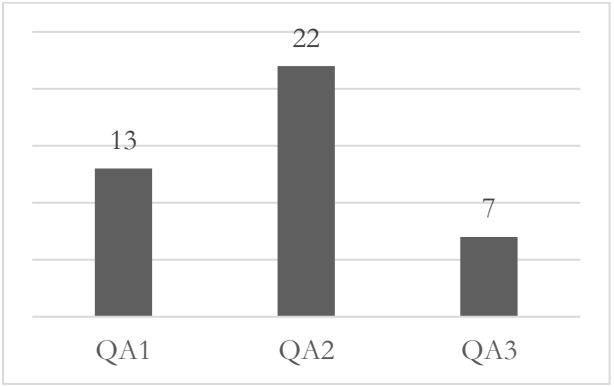

Figure 2. Quality Assessment Review Results

3.4. Data Analysis

The further analysis of the selected documents, in total, there are 13 journals dealing with items or components contained in the regulations. Several journals discuss or implement methods related to processes involving natural language

processing and text mining. There are 7 journals consisting of contents about the preprocessing method, and 13 journals about feature extraction method. Literature using the K-Means method was found in four journals, and the cosine similarity method exists in three journals, as shown in Figure 3.

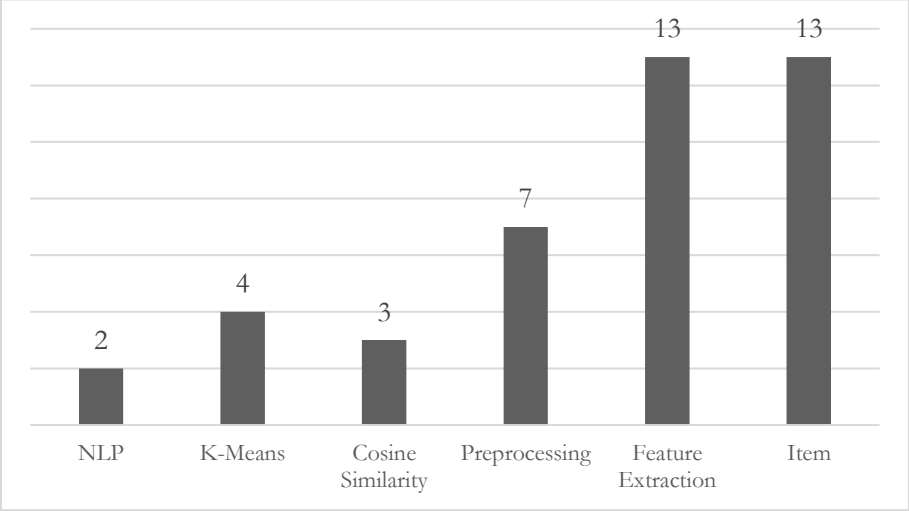

Figure 3. Total Review of Relevant Journals

Other literature discusses the application of interconnection calculation methods that are not widely used, such as principal component analysis (PCA), synonym recognition method, simple random sampling method, market basket analysis, ratcliff/obershelp algorithm, quantitative methods, input-output analysis methods, qualitative methods (qualitative research), combined method or mix method, multivariate analysis method, associative descriptive method, descriptive method, associative method, normative juridical method, active fuzzy constrained clustering (AFCC) algorithm, and hierarchical clustering method. In some literature, the NLP method is considered to relate to how machines understand human language to interact with each other [5]. It all contributed to the development of a more intelligent model and capable of delivering more qualitative experimental results [6].

To carry out interconnection analysis between end-level documents, such as document grouping analysis, one method that can be used is the K-means clustering method [7]. The K-Means algorithm is believed to be able to produce clusters optimally. Before feature extraction activities, the text preprocessing step also needs to be carried out. Several stages of text preprocessing include case folding, tokenizing, filtering, and stemming [8]. For feature extraction, the document content can be formed into a bag of words representations, TF-IDF, WordCount, N-Gram, and Word Embeddings [9], [10]. For analyzing the

interconnection between end-level documents, methods of common calculation such as cosine similarity can be used [11]. This method is widely used in NLP applications, document recommendation systems, and document grouping. In a case, the cosine similarity method can be applied with steps starting with preprocessing, then searching for synonymous words, and continued by calculating the values of TF-IDF, ending with calculating cosine similarity values [12].

3.5. Research Questions

This section discusses the answers to the three research questions (RQs) that have been set for this study.

1. Which items or components in a regulatory document can be used to analyze the degree of interconnection between the regulatory documents?

In analyzing the interconnection between regulation documents, there are eight items or components that can be found and recommended to be used for representing regulation documents. They are initiator, regulation level, considerations, related regulation, regulation content, rules category, regulatory topics, and regulatory structure. These items can be used to understand the content, context, and characteristics of the document and to evaluate how the regulatory document interconnects with other regulations.

Initiator: An initiator is a state agency or authority that takes the initiative to submit a draft of legislative regulations. The minister or the head of a non-ministerial body who submits proposals for law drafts, government regulations drafts to replace the laws and draft regulations of the government, and draft presidential rules the Minister or head of the non-departmental body who is the author of draft legislative regulations whose submissions are within the scope of his duties and responsibilities [13]. For local regulations, initiation can be made by the DPRD or the head of the district. Proposals for draft provincial-level or district-level or city-level regional regulations are based on research or examination of law and other research results within the appropriate territorial coverage [14].

Regulatory level: The hierarchy of regulations in legislation, or what is meant by the level of regulation, becomes one of the important foundations in the process and technique of drafting legislative regulations [14], [15]. Under Law (UU) no. 12 of 2011 article 7 section 1, the types and hierarchy of statutory regulations consist of: (1) The 1945 Constitution of the Republic of Indonesia (UUD45), (2) Decree of the People's Consultative Assembly (Tap MPR), (3) Laws/Government Regulations in lieu of Laws (UU/PERPU), (4) Government Regulations (PP), (5) Presidential Regulations (Perpres), (6) Provincial Regional Regulations (Perda

Prov), and (7) Regulations Regency/City Region (Perda Kab/Kota). The types of regulations other than those referred to in Article 7 Section 1 include those established by the People's Assembly, House of Representatives, Regional Representative Board, the Supreme Court, the Constitutional Court, the Supreme Audit Agency, the Judicial Commission, the Bank of Indonesia, the Minister, body, institution, or commission of the same level established under the Law or Government on the order of the Law, the Provincial People's Representative Council, the Governor, the Council of people's representatives of districts or cities, the mayor or mayor, heads of villages or the equivalent [16]. The regulatory level can be used to analyze the relationship between regulatory documents. If a regulatory document and another regulatory document have the same level, then the texts are assumed to be related. If a regulatory document and another regulatory document have a different level, then the texts are assumed to have no relationship or have a lower relationship.

Considerations: Considerations in the regulation of the laws containing a brief description of the ideas that form the background and the reasons for making the regulations of those laws. The establishment of mandatory legislative regulations covers the stages specified in Law No. 12 of 2011. The established phase consists of the phases of planning, preparation, discussion, validation or establishment, and legislation. The regulations state that the establishment of mandatory legislation requires consideration. In the V edition of the Great Dictionary of Indonesian Language (KBBI), consideration is understood as a basic consideration of decision-making, regulation, and so on. Considerations in a rule manuscript are divided into three types of foundations, consisting of philosophical foundations, juridic foundations, and sociological foundations [17].

The philosophical basis is one of the foundations that describes or expresses that the rules that are formed consider the conscious view of life as well as legal ideals, which include the atmosphere of mysticism and Indonesian philosophy. The philosophical foundation comes from the Foundation of the Republic of Indonesia, namely Pancasila, as stated in the Preamble to the 45th Constitution. The sociological basis is a consideration or reason that illustrates that a rule that is formed is used to meet the needs of society in various aspects. This foundation is supported by empirical facts regarding the development of problems and needs in society and the country. The juridical basis is a consideration or reason that illustrates that the regulations formed are rules to overcome a legal problem or fill a legal vacuum. This basis includes consideration of existing regulations, those that will be changed, or those that will be revoked. This weighing content is needed to ensure legal certainty and society's sense of justice. These three foundations are important considerations in drafting legal regulations [18].

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

One part of the activity of drafting a law, especially in terms of considerations, is that it requires an academic text whose preparation is based on a philosophical basis, a juridical basis, and a sociological basis. This study will accompany the draft legislative text with considerations. This is a necessity for legislative and executive institutions when drafting legislation. With this academic text, stakeholders will be able to review or study whether the draft legislation that is being prepared is worthy of being proposed or not. This is important because the preparation of regulations ultimately aims to create a comprehensive system of rules from both philosophical, sociological, and juridical aspects.

Related Regulations: The remembering section, also known as the legal basis for regulations, is a juridical basis for the formation of statutory regulations. The remembering section contains the legal basis in the form of the authority to form the statutory regulations and the statutory regulations ordering the formation. In regulations at the law level, the first part of "remembering" includes Article 18, Article 18A, Article 18B section (2), Article 20, Article 21, and Article 22D section (2) of the UUD45. In PERPU level regulations, the first part of "remembering" is Article 22 paragraph (1) of the UUD45. In PP level regulations, the first part of "remembering" is Article 5 paragraph (2) of UUD45. At the Presidential Decree level, the first part of "remembering" is Article 4 paragraph (1) of UUD45, and the legal basis contained in the "remembering" provisions of the Regional Regulation consists of Article 18 paragraph (6) of UUD45, the Law on Regional Establishment, and the Law on Regional Government [19]. In analyzing the relationship between regulatory documents, the related regulations contained in the regulatory documents are also used as a basis. If a regulatory document and another regulatory document have the same related regulations, then the regulatory document has a higher relationship than a regulatory document that has different related regulations.

Regulatory Content: Content is the subject, type, or unit of digital information. Content can be text, images, graphics, videos, voices, documents, reports, etc. [20]. In other words, content is everything presented or published in a document. In a regulatory document, the content is presented in the form of text. In the formulation of legislative regulations, the formulator must understand the question fundamentally to be able to pour ideas into the language of the text of the legislation [21].

Regulation Category: The category of regulation can be divided into statutory regulations, other statutory regulations, and other legal instruments. Determining the category of regulation can be done by looking at where the enactment of the legal product is carried out. Statutory regulations in Indonesia consist of seven types according to their hierarchical level, as explained in the level of regulations. Meanwhile, other laws and regulations are those that are recognized for their

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

existence and have binding legal force if they are ordered by higher laws and regulations or formed based on authority [22]. Other types of laws and regulations exist other than those contained in statutory regulations, namely the regulations contained in Article 8 paragraph 1 [23]. Then what includes other legal instruments include decisions, instructions, and circulars.

Regulation Topic: The types of topics or themes discussed in the regulation are very diverse, each of which has broad and varied implications for the society and economy of a country. Some examples of regulatory topics are education, health, the environment, agriculture, employment, taxation, investment, transportation, finance, and others. In analyzing the correlation between the regulatory document and the subject of the regulation that exists in the regulation text, if the topic of one regulatory document is the same as that of another, then the relationship between those regulatory documents is higher than that of the regulatory manuscript, which has different regulation topics.

Regulatory Structure: The regulatory structure of legislation is the framework or arrangement that regulates how laws are established, interpreted, and applied in a particular country or jurisdiction. The framework that becomes a concept in the formulation of regulatory laws is an important process in determining legal justice for a society [24]. In analyzing the interrelationship between regulations, the regulatory structure contained in the regulations should be observed. If a regulatory document has the same structure as another regulatory document, then the degree of correlation between the regulatory documents is assumed to be higher.

The basic framework that forms the essential parts of a legislative regulation is determined by reference to the provisions in Appendix I of Press No. 188 of 1998 on Techniques for the Preparation of Legislations. The legislative framework consists of: (1) Title, (2) Opening, (3) Body Strips, (4) General Regulations, (5) Provisions governing content material, (6) Penal Provision, (7) Transitional Regulation, (8) Closing Provision, (9) Closure, (10) Explanation (if required), and (11) Appendices (if necessary) [25].

The title contains a description of the type, number, year of legislation or establishment, and the name of the rule of law. The opening of the regulations of the laws consists of specific phrases: the phrase with the grace of the Lord who has it, the office of regulating the laws, the consideration, the basis of the law, and the dictum. The body of the rule is grouped into several sections, including general provisions, regulated substances, criminal provisions (if required), transitional provisions (if necessary), and closing provisions. Closure is the final part of a regulation containing a legislative order and its placement, which is categorized as placement in the LN (State Gazette), BN (National News), LD (Distance Gazette),

or BD (District News), and the signature of the approval or regulatory establishment. An explanation is an official interpretation of the formation of legislative regulations on certain norms in the body and should not be used as a legal basis for making further regulations. An attachment may contain descriptions, lists, tables, images, maps, and/or sketches. If a regulation requires an annex, it shall be stated in the body that the annex is an integral part of the regulation.

2. Which methods can be used to extract the regulatory document component and how are their capacities?

a) Text Preprocessing

Preprocessing is a phase used to prepare original data that is available and in raw condition, ready for processing [26]. This also applies to text documents. The purpose of text preprocessing is to prepare text into data that will undergo further processing or analysis. At this stage, several processes can be performed, including tokenizing, lowercasing, filtering, and stemming [12]. Alternatively, there are also some parts of preprocessing that are beneficial, such as part-of-speech tagging, lemmatization, and chunking.

This phase is generally begun by breaking down a set of statements into words using the concept of tokenizing [27]. Tokens can be words, phrases, or reading marks. The purpose of this formation is to break down the text into units that are easier to analyze. One step to implement is lowercase conversion to convert all letters into small letters. The process is continued with the filtering process, which is the phase for taking important words from the token result [27]. This stage also involves removing special characters, reading marks, and irrelevant or common words (stopwords) from the text. Text cleaning can also include normalizing text, such as turning words into their basic form. This phase is often referred to as stemming, which is the phase of searching for the underlying words of each filtered word. In this phase, the process of returning various punches of words to the same basic representation is carried out.

In addition to the text processing methods described above, there are several other stages of preprocessing that can help in analyzing and understanding text better. Part-of-speech tagging (POS-tagging) is a process of giving a word class mark to every word in the corpus. The word class referred to is, among other things, word work, word object, word character, word description, and some others [28]. POS-tags can be done either automatically or manually. POS-taggings are done manually using one or several linguists that give a suitable tag for each word in a text or corpus [29]. This method can also be done automatically using mathematical or other methods [28]. One other method, which is lemmatization, is a morphological transformation to transform words that appear in text into a basic form or

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

dictionary, known as a lemma. By reducing the number of different terms, lemmatization reduces the complexity of the text analyzed and therefore brings important benefits to the text processing component [30]. There is also the chunking phase, which is a process of grouping words into larger units, such as an object phrase or a word phrase. The chunking process is obtained from the results of the POS-tagging process. In the chunker process, the word class of POS-tags will be extracted into six phrasing levels, such as NP, VP, Preposition Phrase (PP), Adverb Phrase (ADVP), Adjective Phrases (AP), and Numerical Phrases (NUMP), based on rules created manually [31].

b) Feature Extraction

The next step that is part of text mining and needs to be implemented is the feature extraction process. One form of feature that can be extracted from document data is a Bag of Words (BoW). This method is a model that studies the vocabulary of the entire document and then models each document by counting the occurrence of each word [32]. In this model, text, whether in a sentence or a document, is represented as a multi-set bag of words contained in it. This representation does not pay attention to word order or grammar but still highlights the word diversity contained in the text. There are two methods in the Bag of Words. The first method is to count the words that appear in the sentence with the count vectorizer, and the second is to calculate the frequency of the words appearing in a sentence multiplied by the number of documents that contain a word in the form of TF-IDF. The result of the algorithm is a matrix whose lines are sentences and columns are unique keywords whose origin has been calculated. Because of its simple working principle, the Bag of Words algorithm is often used for categorizing text data [33].

For the Bag of Words method, there are some appearing weaknesses, including the result of the Bag of Words representation removing the sequence of a sentence and the meaning of the sentence. If there are many keywords existing, the size of the matrix will be enormous. The larger the size or dimension, the more data is needed to generalize the document accurately. For this, the process of categorizing big data dimensions takes a long time. It requires the vector representation of sentences to take the shape of smaller dimensions so that a better performance can be achieved [33]. The next method to be used is the Term Frequency-Inverse Document Frequencies (TF-IDF). TF-IDF is often used in various natural language processing applications, such as document classification, document extraction, and document grouping. The term frequency (TF) measures how often a word appears in a document. A way to calculate the TF is to divide the number of occurrences of a certain word in the document by the total number of words in that document. The formula used in the calculation of TF is as shown in Equation 1.

$$\text{TF}(t, d) = \frac{(\text{number of occurrences of word } t \text{ in document } d)}{(\text{total number of words in document } d)} \quad (1)$$

Inverse Document Frequency (IDF) measures how important a word is in a document collection. Words that appear in many documents have lower IDF values, while words that rarely occur in the document collection have a higher IDF value. The formula used in the calculation of IDF is as shown in Equation 2.

$$\text{IDF}(t, d) = \log \frac{(\text{total number of documents in corpus } d)}{(\text{the number of documents in the corpus containing the word } t)} \quad (2)$$

The TF-IDF score for a word in a document is the result of the multiplication between TF and IDF for the word. TF-IDF gives a higher weight to words that often appear in a particular document but rarely appear within the entire document collection. With TF-IDF, how important a word in a document, relative to the entire document collection, can be evaluated. Words with a high TF-IDF score in a document tend to be keywords or main topics in the document.

WordCount is a function or command used to count the number of words or characters in a document. The usefulness of WordCount is to help in knowing how long a document or piece of content has been produced [34]. Wordcount is easy to implement but does not consider word sequence, it only counts frequency. WordCount can be done in two ways, namely sequentially without MapReduce and with MapReduce using Hadoop [35].

The next method is the N-Gram method. An N-gram is used for the word sequence itself or a predictive model that assigns a probability [36]. The N-Gram process is based on separating text into strings with a length of n, starting from a specific position in the text. The next N-Gram position is calculated based on the actual position shifting according to the given offset. The offset value depends on the division used in the N-Grams. The N-grams for each string are calculated and then compared one by one. N-grams can be unigrams (n = 1), bigrams (n = 2), trigrams (n = 3), and so on. The N-Gram technique involves two steps, namely dividing the string into overlapping N-Grams (a set of substrings with length n) and performing checks to obtain a substring that has the same structure [37]. N-grams have an advantage in checking a string (string matching), one of which is that they have a low sensitivity to typing errors [38]. As for the characteristics of N-Gram, it is tolerant of text errors, efficient because it has a simple algorithm, and fast in processing. The N-gram word has better accuracy [36]. Before performing N-Gram calculations, it is recommended to perform stemming for the text cleansing process first [37]. The forms of character resolution in N-Gram are as in Table 2.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Table 2. N-Gram Model and Character Resolution

N-gram	Character Resolution
Uni-Gram	'H', 'A', 'L', 'O', ' ', 'N', 'A', 'M', 'A',
Char	'H', 'A', 'L', 'O', ' ', 'N', 'A', 'M', 'A', ' ', 'S', 'A', 'Y', 'A'
Bi-Gram Char	'HA','AL', 'LO', 'O', 'N', 'NA', 'AM', 'MA', 'A', 'S', 'SA', 'AY', 'YA'
Tri-Gram Char	'HAL', 'ALO', 'LO', O N', 'NA', 'NAM', 'AMA', 'MA', 'A S', 'SA', 'SAY', 'AYA'
Unigram Word	'HALO', 'NAMA', 'SAYA'
Bi-Gram Word	'HALO NAMA', 'NAMA SAYA'

Another method available and used in the feature extraction process is word embedding. Word embedding recognizes the distribution of similar word meanings that are then recognized on a vector model [38], [39]. Word embedding does not understand the text as humans do, but it maps the statistical structure more than the language used in the corpus. The aim is to map semantic meaning into geometric space [40]. These geometric spaces are often referred to as embedding spaces. In other words, words that have similar meanings or often appear in the same context will have close vector representations in these vector spaces. By capturing the characteristics of a word, whether it is the original word or a similar word, the resemblance of one word with another word can be counted. Word embedding is usually used in the first phase of the deep learning process of an information [41]. Word embedding represents a word in 50–300 fixed dimensions, where each vector is a dense vector, so the set of vectors can represent the relationship between words well, both semantically and syntactically. Word embedding (pre-trained word embedding) is implemented in various languages, including Indonesian [42].

3. What methods can be used to analyze the interconnection between regulatory documents and how are their capacities?

According to [43], K-Means clustering is a suitable method for grouping large amounts of data into several classes according to attributes or characteristics like other data. In data mining, there are two types of data grouping methods, namely non-hierarchical clustering and hierarchical clustering [44]. K-Means clustering is a non-hierarchical algorithm that tries to partition data into the form of one or more clusters, where each object belonging to a group has similar object and correlates one with the other [45]. Data grouped within a cluster has a greater degree of similarity with one to another, with greater degrees of difference to the data in the other group.

K-Means algorithms generally have the following procedures [8], [11]: determining the value of k, or the number of clusters, then performs the activity to determine the initial centroid randomly. The next process is to calculate the distance between

the data and the center point of the cluster (the centroid). In the case of document grouping, the calculation of the distance of data to the centroid can be done using the cosine similarity method. The object allocation based on re-distance will be performed by looking at the nearest distance of the object to the centroid. The process then continues to repeat centroid calculations until the centroid, or the center point of the cluster, has not changed anymore.

Cosine similarity is the most used method of measuring the similarity of sentences. This method performs the calculation by looking at the two sentences in the form of two vector angles. The sentence here is considered a vector, with the cosine value of the angle between the two vectors as the parameter of distance. The closer the distance between the two vectors, the more similar the two sentences are [11]. Cosine similarity is used in positive spaces where the value of the result is between 0 and 1. If the value is 0, then the document is said not to be similar, and if the number is 1, then it is said to have a high similarity, even equal [46]. The formula for calculating cosine similarity using the TF-IDF concept is as shown in Equation 3.

$$cos\,\theta = \frac{A\cdot}{|A|}\ \frac{B}{|B|} = \frac{\sum_{i=1}^{n} A_i \, x \, B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2 \, x \, \sum_{i=1}^{n} (B_i)^2}} \quad (3)$$

Description:
A = vector A, which will compare the similarity.
B = vector B, which will compare the similarity.
A · B = dot product between vector A and vector B.
|A| = length of vector A
|B| = length of vector B
|A| |B| = cross product between |A| |B|
A_i = the i-th element of vector A.
B_i = the i-th element of vector B.
n = number of dimensions of a vector (number of elements in the vector).

4.1. Regulatory Document Processing Experiments

In finding out how the combination works in analyzing document interconnection, experiments are performed using three regulation documents. The documents are "Peraturan Gubernur Bali Nomor 9 Tahun 2024 Tentang Penugasan Kepada PT Penjaminan Kredit Daerah Provinsi Bali Untuk Melakukan Kerja Sama Dalam Pengembangan Pembiayaan Dan Penyelenggaraan Sistem Angkutan Umum Berbasis Kereta" stated as D1, "Peraturan Gubernur Bali Nomor 7 Tahun 2024 Tentang Perubahan Atas Peraturan Gubernur Nomor 112 Tahun 2018 Tentang Tarif Penumpang Angkutan Umum Trans Sarbagita Di Provinsi Bali" stated as D2, and "Peraturan Menteri Kesehatan Republik Indonesia

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Nomor 17 Tahun 2017 Tentang Rencana Aksi Pengembangan Industri Farmasi Dan Alat Kesehatan" stated as D3. Table 3 shows the results of interconnection analysis between regulatory document D1 and regulatory document D2. The highest interconnection is achieved when using a combination of Case Folding and Tokenizing for document preprocessing and Word Embedding as feature representation with interconnection degree of 92,70% and lowest interconnection of 75,10% when using Case Folding, Tokenizing, and Filtering; and Part of Speech, Lemmatization, and Chunking for preprocessing; and TF-IDF and N-Gram for feature extraction.

Table 3. Results of Interconnection Analysis Between D1 and D2

Combination of Methods	Case Folding, Tokenizing	Case Folding, Tokenizing, Stemming	Case Folding, Tokenizing, Filtering	Case Folding, Tokenizing, Filtering, Stemming	Part of Speech, Lemmatization, Chunking
TF-IDF	79.69%	79.82%	78.29%	78.46%	78.29%
TF-IDF (N-Gram)	76.06%	76.22%	75.10%	75.33%	75.10%
N-Gram	80.92%	81.07%	79.62%	79.88%	79.62%
WordCount	84.48%	84.62%	82.83%	83.06%	82.83%
WordCount (TF-IDF)	79.69%	79.82%	78.29%	78.46%	78.29%
WordCount (N-Gram)	80.92%	81.07%	79.62%	79.88%	79.62%
WordCount (TF-IDF + N-Gram)	76.06%	76.22%	75.10%	75.33%	75.10%
Bag of Words	84.48%	84.62%	82.83%	83.06%	82.83%
Bag of Words (TF-IDF)	79.69%	79.82%	78.29%	78.46%	78.29%
Bag of Words (N-Gram)	80.92%	81.07%	79.62%	79.88%	79.62%
Bag of Words (TF-IDF + N-Gram)	76.06%	76.22%	75.10%	75.33%	75.10%
Word Embedding	92.79%	91.72%	92.47%	89.89%	92.47%
Word Embedding (TF-IDF)	92.79%	91.72%	92.47%	89.89%	92.47%
Word Embedding (N-Gram)	92.79%	91.72%	92.47%	89.89%	92.47%
Word Embedding (TF-IDF + N-Gram)	92.79%	91.72%	92.47%	89.89%	92.47%

Table 4 shows the results of interconnection analysis between regulatory document D1 and regulatory document D3. The highest interconnection level is

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

67,37% when using a combination of Case Folding and Tokenizing for preprocessing and Word Embedding for feature representation and lowest interconnection of 26,69% when using Case Folding, Tokenizing, and Filtering; and Part of Speech, Lemmatization, and Chunking for preprocessing; and TF-IDF and N-Gram for feature extraction.

Table 4. Results of the Averange Level of Linkage Between Regulatory documents

Combination of Methods	Case Folding, Tokenizing	Case Folding, Tokenizing, Stemming	Case Folding, Tokenizing, Filtering	Case Folding, Tokenizing, Filtering, Stemming	Part of Speech, Lemmatization, Chunking
TF-IDF	32.91%	33.21%	29.63%	30.01%	29.63%
TF-IDF (N-Gram)	28.74%	28.89%	26.69%	27.02%	26.69%
N-Gram	34.67%	35.09%	31.66%	32.48%	31.66%
WordCount	39.66%	40.36%	35.28%	36.33%	35.28%
WordCount (TF-IDF)	32.91%	33.21%	29.63%	30.01%	29.63%
WordCount (N-Gram)	34.67%	35.09%	31.66%	32.48%	31.66%
WordCount (TF-IDF + N-Gram)	28.74%	28.89%	26.69%	27.02%	26.69%
Bag of Words	39.66%	40.36%	35.28%	36.33%	35.28%
Bag of Words (TF-IDF)	32.91%	33.21%	29.63%	30.01%	29.63%
Bag of Words (N-Gram)	34.67%	35.09%	31.66%	32.48%	31.66%
Bag of Words (TF-IDF + N-Gram)	28.74%	28.89%	26.69%	27.02%	26.69%
Word Embedding	67.37%	60.66%	65.25%	57.70%	65.25%
Word Embedding (TF-IDF)	67.37%	60.66%	65.25%	57.70%	65.25%
Word Embedding (N-Gram)	67.37%	60.66%	65.25%	57.70%	65.25%
Word Embedding (TF-IDF + N-Gram)	67.37%	60.66%	65.25%	57.70%	65.25%

Table 5 shows the results of interconnection analysis between regulatory document D2 and regulatory document D3. The highest interconnection level is 65,76% when using a combination of Case Folding and Tokenizing for preprocessing and Word Embedding for feature representation and lowest interconnection of 25,78% when using Case Folding, Tokenizing, and Filtering;

and Part of Speech, Lemmatization, and Chunking for preprocessing; and TF-IDF and N-Gram for feature extraction.

Table 5. Results of Interconnection Analysis Between D2 and D3

Combination of Methods	Case Folding, Tokenizing	Case Folding, Tokenizing, Stemming	Case Folding, Tokenizing, Filtering	Case Folding, Tokenizing, Filtering, Stemming	Part of Speech, Lemmatization, Chunking
TF-IDF	32.01%	32.79%	28.50%	29.44%	28.50%
TF-IDF (N-Gram)	28.04%	28.50%	25.78%	26.35%	25.78%
N-Gram	34.27%	35.03%	30.59%	31.67%	30.59%
WordCount	39.01%	40.15%	33.88%	35.59%	33.88%
WordCount (TF-IDF)	32.01%	32.79%	28.50%	29.44%	28.50%
WordCount (N-Gram)	34.27%	35.03%	30.59%	31.67%	30.59%
WordCount (TF-IDF + N-Gram)	28.04%	28.50%	25.78%	26.35%	25.78%
Bag of Words	39.01%	40.15%	33.88%	35.59%	33.88%
Bag of Words (TF-IDF)	32.01%	32.79%	28.50%	29.44%	28.50%
Bag of Words (N-Gram)	34.27%	35.03%	30.59%	31.67%	30.59%
Bag of Words (TF-IDF + N-Gram)	28.04%	28.50%	25.78%	26.35%	25.78%
Word Embedding	65.76%	62.41%	62.21%	58.88%	62.21%
Word Embedding (TF-IDF)	65.76%	62.41%	62.21%	58.88%	62.21%
Word Embedding (N-Gram)	65.76%	62.41%	62.21%	58.88%	62.21%
Word Embedding (TF-IDF + N-Gram)	65.76%	62.41%	62.21%	58.88%	62.21%

Based on the three results, the combination of Case Folding and Tokenizing for preprocessing and Word Embedding for feature representation resulted in highest level of interconnection values and the combination of either Case Folding, Tokenizing, and Filtering; or Part of Speech, Lemmatization, and Chunking for preprocessing; and TF-IDF and N-Gram for feature extraction resulted in lowest level of interconnection values. Furthermore, Table 6 shows the detail interconnection level per component of regulatory documents between regulatory document D1 and regulatory document D2 when using Word Embedding with a combination of TF-IDF and N-Gram as feature representation.

Table 6. Results of Interconnection Analysis Between Regulatory Documents Using the Word Embedding Method (TF-IDF + N-Gram)

Items	Case Folding, Tokenizing	Case Folding, Tokenizing, Stemming	Case Folding, Tokenizing, Filtering	Case Folding, Tokenizing, Filtering, Stemming	Part of Speech, Lemmatization, Chunking
Initiator	100.00%	100.00%	100.00%	100.00%	100.00%
Regulatory Level	100.00%	100.00%	100.00%	100.00%	100.00%
Considerations	81.82%	82.47%	77.23%	73.72%	77.23%
Related Regulations	90.75%	96.14%	97.84%	97.05%	97.84%
Regulatory Content	79.49%	70.51%	74.09%	63.75%	74.09%
Categories of Regulations	100.00%	100.00%	100.00%	100.00%	100.00%
Regulation Topics	100.00%	100.00%	100.00%	100.00%	100.00%
Regulatory Structure	90.27%	84.62%	90.62%	84.58%	90.62%
Average	92.79%	91.72%	92.47%	89.89%	92.47%

Table 7 shows the detail interconnection level per component of regulatory documents between regulatory document D1 and regulatory document D3 when using Word Embedding with a combination of TF-IDF and N-Gram as feature representation.

Table 7. Results of Interconnection Analysis Between Regulatory Documents Using the Word Embedding Method (TF-IDF + N-Gram)

Items	Case Folding, Tokenizing	Case Folding, Tokenizing, Stemming	Case Folding, Tokenizing, Filtering	Case Folding, Tokenizing, Filtering, Stemming	Part of Speech, Lemmatization, Chunking
Initiator	46.06%	37.86%	46.06%	37.86%	46.06%
Regulatory Level	68.25%	50.70%	68.25%	50.70%	68.25%
Considerations	78.79%	70.95%	64.53%	57.37%	64.53%
Related Regulations	86.33%	88.60%	94.03%	92.75%	94.03%
Regulatory Content	78.54%	60.50%	72.09%	49.54%	72.09%
Categories of Regulations	100.00%	100.00%	100.00%	100.00%	100.00%
Regulation Topics	0.67%	13.34%	0.67%	13.34%	0.67%
Regulatory Structure	80.31%	63.35%	76.34%	60.08%	76.34%
Average	67.37%	60.66%	65.25%	57.70%	65.25%

Table 8 shows the detail interconnection level per component of regulatory documents between regulatory document D2 and regulatory document D3 when using Word Embedding with a combination of TF-IDF and N-Gram as feature representation.

Table 8. Results of Interconnection Analysis Between Regulatory Documents Using the Word Embedding Method (TF-IDF + N-Gram)

Items	Case Folding, Tokenizing	Case Folding, Tokenizing, Stemming	Case Folding, Tokenizing, Filtering	Case Folding, Tokenizing, Filtering, Stemming	Part of Speech, Lemmatization, Chunking
Initiator	46.06%	37.86%	46.06%	37.86%	46.06%
Regulatory Level	68.25%	50.70%	68.25%	50.70%	68.25%
Consideration	76.79%	77.20%	62.85%	65.98%	62.85%
Related Regulations	92.57%	89.87%	94.45%	92.68%	94.45%
Regulatory Content	64.31%	61.21%	51.87%	48.00%	51.87%
Categories of Regulations	100.00%	100.00%	100.00%	100.00%	100.00%
Regulation Topics	0.67%	13.34%	0.67%	13.34%	0.67%
Regulatory Structure	77.46%	69.11%	73.51%	62.52%	73.51%
Average	65.76%	62.41%	62.21%	58.88%	62.21%

When applying the K-Means clustering method to further group the three documents into two clusters based on the similarity measures calculated above, it is found that D3 is separated from the other two regulatory documents D1 and D2 with the silhouette coefficient 0.587.

Table 9. Clustering Results using K-Means Clustering

Cluster 1	Cluster 2
D1	D3
D2	

3.6 Discussion

A few combinations of preprocessing can be used in analyzing the degree of correlation between the regulatory documents. They are case folding, tokenizing, and filtering. The advantages of this combination are that case folding ensures that all words are in the same format, avoids differences caused by capital letters, removes general words that are not informative and helps focus on more important words, while tokenizing facilitates the analysis and manipulation of text.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

On the other hand, the use of filtering can remove important words in a particular context, especially in regulatory manuscripts where each word may have a specific meaning.

Then there is a combination of case folding, tokenizing, and stemming. The advantage of this combination is that stemming can help reduce word variations with the same basic shape so that analysis is more focused; case-finding ensures all words are in the same format; and tokenizing facilitates text manipulation. The weakness of this combination is in stemming, where this mechanism can turn words into incomplete forms. It can cause the sentence in the rule script to lose its important meaning.

Another experiment involves a combination of case folding, tokenizing, filtering, and stemming. This combination gives a text that is more consistent and focused on relevant words. These steps reduce the number of words to be analyzed as well as make the process faster and more efficient. However, a combination of filtering and stemming can remove or change words that are important in a particular context, especially in the digestion of sentences in a regulatory document.

There is also a combination of case folding and tokenizing that can be used to ensure textual consistency and break the text into easily analyzed parts where all words can be retained in their original form, as well as provide full context. However, this combination does not include the deletion or simplification of words. This makes analysis slower and less efficient, as well as making variations in word forms impossible to overcome, which in turn can reduce the accuracy of the analysis.

The last combination is part-of-speech tagging, lemmatization, and chunking. These combinations have advantages, including part-of-speech tagging, which provides information about the function of words in sentences and is very useful in understanding the structure and meaning of rule manuscripts. Lemmatization ensures all words are in their basic form and can improve the accuracy of analysis, and chunking can help in understanding phrase structure and relationships between words, as well as providing a better context. The disadvantages of this combination are in terms of resource requirements and processing time. Besides, part-of-speech tagging and chunking can also make mistakes, especially if the training data is inadequate or the text has a complex structure.

After the text preprocessing phase, the next step is feature extraction process, where the process can combine the features of Bag of Words, WordCount, N-Gram, Words Embedding, and TF-IDF. Feature extraction can also be combined with analyzing the degree of correlation between rule manuscripts. Some of these combinations are (1) TF-IDF, (2) TF-IDF and N-Gram, (3) N-Gram, (4)

WordCount, (5) WordCount and TF–IDF, (6) WordCount and N-Gram, (7) TF-IDF dan N-Grams, (8) Bag of Words, (9) Bag of Words and TF–IDF, (10) Bag of Words and N-Grams, (11) TF-IDF and N-Gram, (12) Word Embedding, (13) Word Embedding and TF- IDF, (14) Word Embeddings and N–Grams, (15) TF-IDF and N - Grams.

The combination of preprocessing methods and specific feature extraction results in a higher level of interconnection because they complement each other in maintaining a balance between text consistency, contextual meaning, and focus on relevant information. Simpler and faster methods (such as case folding and tokenizing) assist in basic processing, while more complex methods (such as TF-IDF and Word Embeddings) enable deeper and context-rich analysis, which overall enhances accuracy in identifying relationships among regulatory documents.

3.7 Recommendation

From the results of the review, several items or components in the rules and methods can be used to analyze the degree of correlation between regulatory documents. As explained in the discussion section of the results, eight items or elements in the regulations can serve as a basis for analyzing the level of interconnection between regulatory documents. Items include rule initiator, regulatory level, considerations, related regulations, regulatory content, categories of regulations, regulation topics, and regulatory structure. In analyzing items in regulatory documents, it is necessary to measure whether these items have high similarities with each other or not. If many items have a high degree of similarity, then the degree of interconnection between the regulation documents is high. On the contrary, if many items are compared as having a low degree of similarity, the rate of interconnection between those documents is deemed to be low.

To measure the level of similarity of an item or component of a rule, there are several processes to be performed. The first phase is a text processing phase that consists of case folding to turn letters into small letters, tokenizing to break down text into tokens, filtering to remove reading marks, numbers, or other irrelevant elements, and stemming to turn words into basic words. Then there is also part-of-speech tagging (POS-tagging) to determine the word class, lemmatization to turn the word into a basic word by considering the class of words and chunking to group the tokens into larger phrases based on the class.

Based on the experiments conducted using the three regulatory documents with various combinations of document preprocessing and feature extraction, the combination of Case Folding and Tokenizing for preprocessing and Word Embedding for feature representation resulted in highest level of interconnection

values and the combination of either Case Folding, Tokenizing, and Filtering; or Part of Speech, Lemmatization, and Chunking for preprocessing; and TF-IDF and N-Gram for feature extraction resulted in lowest level of interconnection values. Each combination of preprocessing and feature extraction methods has their own characteristics. Preference will appear whether the analysis needs a simpler and faster method or requires a complex one. Further study might also be needed to confirm a combination that commonly occurs when regulatory document users analyze the interconnection between regulatory documents in daily activities.

4. CONCLUSION

Regulatory documents play a crucial role in shaping the frameworks that govern social, economic, and organizational life. However, these documents are often perceived as complex and disconnected by the public. An automated mechanism that can analyze the interconnections between regulatory texts is essential to address this issue, particularly as a form of unstructured data analysis within the realm of big data. Research Question 1 (RQ1) revealed that eight key components of regulatory documents—such as the initiator, regulatory level, considerations, related regulations, and regulatory structure—can be used to assess the degree of interconnection between these documents. This allows for a more systematic and comprehensive understanding of how different regulations relate to one another, thus making them more accessible and functional for public use and organizational compliance.

Further analysis, as discussed in Research Questions 2 and 3 (RQ2 and RQ3), highlights the various text processing methods that can be used to extract relevant features from regulatory documents. Methods like tokenizing, stemming, and lemmatization can be combined with feature extraction techniques such as TF-IDF, N-Gram, and Word Embedding to measure the interconnectivity of documents. Experimental results showed that combinations like Case Folding with Tokenizing and Word Embedding yielded the highest interconnection values, while others, like TF-IDF and N-Gram, resulted in lower values. These findings have practical implications in fields such as digital governance and regulatory compliance, where automated systems can identify connections between existing and new regulations, enabling organizations to promptly adjust their internal policies and remain compliant with evolving legal standards. By employing similarity measures like cosine similarity and clustering methods such as K-Means, the degree of interconnections between regulatory documents can be effectively analyzed, offering a faster and more accurate way to ensure compliance across different sectors.

REFERENCES

[1] H. Sartika, E. Purnama, and I. Ismail, "Standard Patterns of Considerations in Law, District Regulation and Qanun Based on Legal Rules in Indonesia," *Pancasila and Law Review*, vol. 2, no. 2, pp. 121–132, 2021.

[2] S. J. Spiegel, "Governance institutions, resource rights regimes, and the informal mining sector: Regulatory complexities in Indonesia," *World Dev.*, vol. 40, no. 1, pp. 189–205, 2012.

[3] S. C. Fanni, M. Febi, G. Aghakhanyan, and E. Neri, "Natural language processing," in *Introduction to Artificial Intelligence*, Springer, 2023, pp. 87–99.

[4] E. Rahmi, E. Yumami, and N. Hidayasari, "Analisis Metode Pengembangan Sistem Informasi Berbasis Website: Systematic Literature Review," *Remik: Riset dan E-Jurnal Manajemen Informatika Komputer*, vol. 7, no. 1, pp. 821–834, 2023.

[5] V. R. Prasetyo, N. Benarkah, and V. J. Chrisintha, "Implementasi natural language processing dalam pembuatan chatbot pada program information technology universitas surabaya," *J. TEKNIKA*, vol. 10, no. 2, pp. 114–121, 2021.

[6] D. Apriliani, S. F. Handayani, and I. T. Saputra, "Implementasi Natural Language Processing (NLP) Dalam Pengembangan Aplikasi Chatbot Pada SMK YPE Nusantara Slawi," *Techno. com*, vol. 22, no. 4, 2023.

[7] M. N. Zhafar, K. Usman, and F. Akhyar, "Penerapan Metode Clustering Dengan Algoritma K-Means Untuk Analisa Persebaran Varian Covid-19 (Studi Kasus Kelurahan Antapani Kidul)," *eProceedings Eng.*, vol. 10, no. 5, 2023.

[8] N. W. Utami and I. G. J. E. Putra, "Text Minig Clustering Untuk Pengelompokan Topik Dokumen Penelitian Menggunakan Algoritma K-Means Dengan Cosine Similarity," *J. Inform. Teknologi dan Sains (Jinteks)*, vol. 4, no. 3, pp. 255–259, 2022.

[9] J. Nurvania, J. Jondri, and K. M. Lhaksamana, "Analisis Sentimen Pada Ulasan di TripAdvisor Menggunakan Metode Long Short-Term Memory (LSTM)," *eProceedings Eng.*, vol. 8, no. 4, 2021.

[10] P. M. Prihatini, "Implementasi Ekstraksi Fitur Pada Pengolahan Dokumen Berbahasa Indonesia," *J. Manajemen Teknol. dan Inform. (MATRIX)*, vol. 6, no. 3, pp. 174–178, 2016.

[11] F. B. Sejati, P. Hendradi, and B. Pujiarto, "Deteksi Plagiarisme Karya Ilmiah Dengan Pemanfaatan Daftar Pustaka Dalam Pencarian Kemiripan Tema Menggunakan Metode Cosine Similarity (Studi Kasus: Di Universitas Muhammadiyah Magelang)," *J. Komtika (Komputasi dan Informatika)*, vol. 2, no. 2, pp. 85–94, 2018.

[12] S. Yusuf, M. A. Fauzi, and K. C. Brata, "Sistem Temu Kembali Informasi Pasal-Pasal KUHP (Kitab Undang-Undang Hukum Pidana) Berbasis Android Menggunakan Metode Synonym Recognition dan Cosine Similarity," *J. Pengembangan Teknol. Inform. dan Ilmu Komputer*, vol. 2, no. 2, pp. 838–847, 2018.

[13] P. Widyantari and A. Sulistiyono, "Pelaksanaan Harmonisasi Rancangan Undang-Undang Perlindungan Data Pribadi (RUU PDP)," *J. Privat Law*, vol. 8, no. 1, pp. 117–123, 2020.

[14] A. O. R. Ritz, "Tugas Dan Peran Kepala Bagian Hukum Sekretariat Daerah Dalam Penyusunan Rancangan Peraturan Daerah Di Kabupaten Tapin Provisi Kalimantan Selatan," *Dinamika*, vol. 29, no. 2, pp. 8186–8197, 2023.

[15] A. Fitryantica, "Harmonisasi Peraturan Perundang-Undangan Indonesia melalui Konsep Omnibus Law," *Gema Keadilan*, vol. 6, no. 3, pp. 300–316, 2019.

[16] A. R. Dewi and S. Hadi, "Konstitusionalitas Permenkumham Nomor 02 Tahun 2019 Penyelesaian Konflik Norma Melalui Mediasi," *Bureaucracy J.: Indones. J. Law and Soc.-Political Gov.*, vol. 2, no. 2, pp. 693–702, 2022.

[17] S. W. Laia and S. Daliwu, "Urgensi landasan filosofis, sosiologis, dan yuridis dalam pembentukan undang-undang yang bersifat demokratis di indonesia," *J. Educ. Dev.*, vol. 10, no. 1, pp. 546–552, 2022.

[18] O. I. Khair, "Analisis Landasan Filosofis, Sosiologis Dan Yuridis Pada Pembentukan Undang-Undang Ibukota Negara," *Academia: J. Inovasi Riset Akad.*, vol. 2, no. 1, pp. 1–10, 2022.

[19] J. P. Pratama, L. T. ALW, and S. A. G. Pinilih, "Eksistensi Kedudukan Peraturan Menteri terhadap Peraturan Daerah dalam Hierarki Peraturan Perundang-Undangan," 2022.

[20] S. M. Mahmudah and M. Rahayu, "Pengelolaan konten media sosial korporat pada instagram sebuah pusat perbelanjaan," *J. Komun. Nusantara*, vol. 2, no. 1, pp. 1–9, 2020.

[21] R. Anggraeni, "Memaknakan Fungsi Undang-Undang Dasar Secara Ideal Dalam Pembentukan Undang-Undang," *Masalah-Masalah Hukum*, vol. 48, no. 3, pp. 283–293, 2019.

[22] B. E. D. Tamin, "Tinjauan Yuridis Terhadap Kedudukan Peraturan Mahkamah Agung (Perma) Dalam Hierarki Peraturan Perundang-Undangan Di Indonesia," *Lex Administratum*, vol. 6, no. 3, 2019.

[23] R. I. Amin and A. Achmad, "Mengurai permasalahan peraturan perundang-undangan di indonesia," *Res Publica: J. Hukum Kebijakan Publik*, vol. 4, no. 2, pp. 205–220, 2020.

[24] Y. Prasetyo, "Urgensi Pembentukan Peraturan Perundang-Undangan Yang Berkeadilan," *J. Legislasi Indones.*, vol. 20, no. 2, 2023.

[25] Z. Afif, "Pembentukan Peraturan Perundang-Undangan Berdasarkan Pancasila Dan Undang-Undang Dasar Negara Kesatuan Republik Indonesia," *J. Dialog*, vol. 7, no. 1, 2018.

[26] S. Parendo and Y. F. AW, "Analisis Dan Implementasi Algoritma Active Fuzzy Constrained Clustering Untuk Pengelompokan Dokumen," *JURIKOM (J. Riset Komputasi)*, vol. 9, no. 2, pp. 194–201, 2022.
[27] E. Dinata and H. Syaputra, "Penerapan Metode Agglomerativ Hirarchical Clusturing Untuk Klasifikasi Dokumen Skripsi," in *Bina Darma Conf. Comput. Sci. (BDCCS)*, 2020, pp. 412–422.
[28] E. Setiyowati, "Hidden Markov Model Bigram Untuk Part Of Speech Tagging Bahasa Lampung Dialek A," *J. Teknologi Pintar*, vol. 2, no. 11, 2022.
[29] M. Astiningrum, P. Y. Saputra, and M. S. Rohmah, "Implementasi nlp dengan konversi kata pada sistem chatbot konsultasi laktasi," *J. Inform. Polinema*, vol. 5, no. 1, pp. 46–52, 2018.
[30] H. P. Fitrian, I. Ruslianto, and R. Hidayati, "Implementasi Metode Naive Bayes Classifier Untuk Aplikasi Filtering Email Spam Dengan Lemmatization Berbasis Web," *Coding J. Komput. dan Apl.*, vol. 6, no. 2, 2018.
[31] P. R. Togatorop, R. P. Simanjuntak, S. B. Manurung, and M. C. Silalahi, "Pembangkit Entity Relationship Diagram Dari Spesifikasi Kebutuhan Menggunakan Natural Language Processing Untuk Bahasa Indonesia," *J-Icon: J. Komput. dan Inform.*, vol. 9, no. 2, pp. 196–206, 2021.
[32] M. S. Negara and A. Z. Mardiansyah, "Implementasi Machine Learning dengan Metode Collaborative Filtering dan Content-Based Filtering pada Aplikasi Mobile Travel (Bangkit Academy)," *J. Begawe Teknol. Inform. (JBegaTI)*, vol. 5, no. 1, pp. 126–136, 2024.
[33] R. P. Kawiswara and F. Thalib, "Implementasi Algoritma Convolutional Neural Network Pada Algoritma K-Means Untuk Kategorisasi Data Teks," *J. Teknol.*, vol. 7, no. 2, pp. 149–160, 2020.
[34] T. M. Sari et al., "Penerapan Sorted Wordcount Dengan Mapreduce Hadoop," *J. Network and Comput. Appl.*, vol. 2, no. 1, pp. 1–12, 2023.
[35] M. D. Marieska, A. S. Utami, and E. Oktaviani, "Perbandingan Metode Mapreduce Berbasis Single Node Hadoop Pada Aplikasi Word Count," *JUPITER: J. Penelitian Ilmu dan Teknol. Komput.*, vol. 16, no. 1, pp. 347–356, 2024.
[36] Z. Pratama, E. Utami, and M. R. Arief, "Analisa Perbandingan Jenis N-GRAM Dalam Penentuan Similarity Pada Deteksi Plagiat," *Creative Inform. Technol. J.*, vol. 4, no. 4, pp. 254–263, 2019.
[37] E. A. Lisangan, "Implementasi n-gram technique dalam deteksi plagiarisme pada tugas mahasiswa," *TEMATIKA: J. Penelitian Teknik Inform. dan Syst. Inform.*, pp. 71–77, 2013.
[38] I. G. Anugrah, "Penerapan Metode N-Gram dan Cosine Similarity Dalam Pencarian Pada Repositori Artikel Jurnal Publikasi," *Building of Informatics, Technol. and Sci. (BITS)*, vol. 3, no. 3, pp. 275–284, 2021.
[39] L. K. Şenel, I. Utlu, V. Yücesoy, A. Koc, and T. Cukur, "Semantic structure and interpretability of word embeddings," *IEEE/ACM Trans Audio Speech Lang Process.*, vol. 26, no. 10, pp. 1769–1779, 2018.

[40] F. P. Rachman and H. Santoso, "Perbandingan Model Deep Learning untuk Klasifikasi Sentiment Analysis dengan Teknik Natural Languange Processing," *J. Teknologi dan Manajemen Inform.*, vol. 7, no. 2, pp. 103–112, 2021.

[41] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Comput Intell Mag.*, vol. 13, no. 3, pp. 55–75, 2018.

[42] Y. Yuliska and K. U. Syaliman, "Literatur Review Terhadap Metode, Aplikasi dan Dataset Peringkasan Dokumen Teks Otomatis untuk Teks Berbahasa Indonesia," *IT J. Res. and Dev.*, vol. 5, no. 1, pp. 19–31, 2020.

[43] D. A. C. Rachman, R. Goejantoro, and F. D. T. Amijaya, "Implementasi Text Mining Pengelompokkan Dokumen Skripsi Menggunakan Metode K-Means Clustering," *EKSPONENSIAL*, vol. 11, no. 2, pp. 167–174, 2021.

[44] M. A. Haq, W. Purnomo, and N. Y. Setiawan, "Analisis Clustering Topik Survey menggunakan Algoritme K-Means (Studi Kasus: Kudata)," *J. Pengembangan Teknol. Inform. dan Ilmu Komputer*, vol. 7, no. 7, pp. 3498–3506, 2023.

[45] G. E. I. Kambey, R. Sengkey, and A. Jacobus, "Penerapan Clustering pada Aplikasi Pendeteksi Kemiripan Dokumen Teks Bahasa Indonesia," *J. Teknik Informatika*, vol. 15, no. 2, pp. 75–82, 2020.

[46] M. Z. Naf'an, A. Burhanuddin, and A. Riyani, "Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen," *J. Linguistik Komputasional*, vol. 2, no. 1, pp. 23–27, 2019.