A Comprehensive Literature Review on Preprocessing Techniques for Text Data in Natural Language Processing on the Topic of Renewable Energy Policies in the Indo-Pacific Region

Student name: Tuan Dat Nguyen

Student's ID: 104489467

1. Introduction

The Indo-Pacific region is facing mounting pressure to make the shift to renewable energy in the context of the global push toward sustainable development [1]. The pressing need to address issues with energy security and climate change is what is driving this shift. But attaining this objective is difficult and multidimensional, involving a range of parties, different laws, and shifting public perceptions.

"Garbage in, garbage out" is a saying used in data science to underline that messy input data will result in messy output. This emphasizes a fundamental fact: preparing data is essential before using any model. If this step is skipped, the outcome will be "garbage out." Currently, unstructured data makes up 90% of all data in the world and might be in the form of text, video, audio, or photos [2]. Text can be found in many different formats, such as lists of single words, sentences, paragraphs on the web, HTML, and papers. There is a lot of noise in this data and it is never clean. Additionally, it must be processed before this data is used to apply any models. Any algorithms developed on top of such data are worthless to a corporation if the data is not handled.

This report's goal is to review every research on text data preparation and contrast the approaches taken by the team. Research will be done to identify problems with text data preparation and provide remedies. It is our responsibility to evaluate various approaches and then suggest a shared one for the team to use when managing text data.

2. Basic Terminology and Delimitations

2.1. Definition of Indo-Pacific regions

The Indian Ocean, the western and central Pacific Oceans, and the seas that separate them are all included in the large region known as the Indo-Pacific region [3]. The following are some crucial regions in the Indo-Pacific:

1. South Asia: Includes countries like India, Sri Lanka, and the Maldives.
2. Southeast Asia: Encompasses nations such as Indonesia, Malaysia, the Philippines, Singapore, Thailand, and Vietnam.
3. East Asia: Includes China, Japan, and South Korea.

4. Oceania: Covers Australia, New Zealand, and the Pacific Island nations like Fiji, Papua New Guinea, and Samoa.
5. Western Pacific: Includes the waters around the Philippines, Taiwan, and the South China Sea.
6. Indian Ocean: Encompasses the waters around the eastern coast of Africa, the Arabian Sea, and the Bay of Benga

2.2. Definition of renewable energy

Energy from natural sources that replenishes more quickly than it is used up is referred to as renewable energy [4]. Typical renewable energy sources include the sun, wind, rain, waves, tides, and geothermal heat. Because these sources replenish spontaneously and have less of an impact on the environment than fossil fuels, they are regarded as sustainable.

2.3. Definition of text data

Information written and saved in a text format is referred to as text data. Anything from emails and blog entries to remarks on social media and online forums can be included. In essence, it's any verbally presented data. Since text data frequently lacks structure and adheres to an arbitrary format, preparation procedures may be necessary before analyzing the data.

2.4. Preprocessing text data

Real-world data is usually noisy, inconsistent, and incomplete, with many inaccuracies present. Preprocessing is a useful method for dealing with these issues. It readies unprocessed text data for further examination. In text mining and natural language processing (NLP), preprocessing text data is an essential step. To clean and get raw text ready for examination, it takes a few distinct methods.

3. Part A

3.1. Literature review

3.1.1. Overview

The shift to renewable energy in the Indo-Pacific region is a complicated and multidimensional task that calls for creative methods of policy analysis, stakeholder involvement, and public opinion surveying [5]. The quantity of articles on pre-processing has expanded in tandem with the advancement of using natural language processing (NLP) in text data analysis. Most studies mention that preprocessing includes: Converting Text Data to Lowercase, Removing Punctuation, Removing Stop Words, Standardizing Text, Correcting Spelling, Tokenizing Text, Stemming, Lemmatizing, Dealing with Emojis and Emoticons. The Scopus database has indexed around seven thousand works on sentiment analysis, according to Mäntylä et al.

The article will review the literature of various articles to identify key studies, critique evidence and claims, and highlight knowledge gaps.

3.1.2. Overview of Key Studies

Table 3.1 Main key studies

Article's name	Author	Key studies
Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis [6]	Marco A. Palomino and Farida Aider	Important of Pre-Processing: The paper highlights how important pre-processing is for sanitizing and standardizing text data, particularly in social media contexts where spelling and grammar conventions are frequently disregarded. This covers using slang, emoticons, emojis, and acronyms. Review of existing literature: The authors reviewed the body of knowledge about text pre-processing methods. Although pre-processing is commonly seen as essential, they pointed out that there are no firm guidelines for optimal techniques.

		Order of pre-processing components: One important discovery is that sentiment analysis classifiers, especially naïve Bayes classifiers, can perform significantly differently depending on the sequence in which pre-processing components are applied. This implies that the order of the components matters in addition to the components themselves. Effectiveness ò lemmatization: According to the study, lemmatization can improve an index's performance, but it has no discernible effect on sentiment analysis's overall quality. This emphasizes the necessity of carefully choosing pre-processing methods according to how they really affect the results of analysis.
The impact of text preprocessing on the prediction of review ratings [7]	Muhittin IŞIK and Hasan DAĞ	Importance of text preprocessing: The authors point out that whereas sentiment analysis for review classification has been extensively studied, text preprocessing has received less focus. They contend that efficient preprocessing techniques can greatly improve categorization accuracy—a critical component of opinion mining Experimental analysis of preprocessing techniques: A variety of preprocessing techniques, such as tokenization, stemming, lemmatization, and emoticon substitution, are tested in this study. The findings demonstrate that while certain techniques, like lemmatization and stemming, slightly increase classification accuracy, others, such emoticon

		substitution and punctuation removal, have no effect. Misspelling correction: The study describes the implementation of an autocorrect or to resolve misspelled words in review texts. The findings imply that fixing misspellings can increase the quality of the data provided into the classifier, while the overall influence on accuracy differs Role of emoticons: The study looks into how review ratings are affected by emoticons. The impact of emoticons on sentiment classification is examined by the authors by substituting matching words for them. Nevertheless, the outcomes show that this approach does not considerably enhance the rating star categorization.
Comparative analysis of effect of stopwords removal on sentiment classification [8]	Ghag KV and Shah K.	Effect of stopwords removal: The study looks into how sentiment classification algorithms' accuracy is affected when stopwords (common words like "the," "and," and "is") are removed. It was discovered that while the suggested classifiers handled stopwords well without requiring further preprocessing, standard sentiment classifiers shown improved accuracy when stopwords were eliminated. Sentiment classification models: The Traditional Sentiment Classifier (TSC), Average Relative Term Frequency Sentiment Classifier (ARTFSC), Sentiment Term Frequency Inverse Document Frequency (Senti-TFIDF), and Relative Term Frequency Sentiment Classifier (RTFSC) were the

		four sentiment classification models that were assessed. Even in the absence of stopword removal, the suggested models (ARTFSC, Senti-TFIDF, and RTFSC) outperformed the conventional model in terms of accuracy. Results: The accuracy of the conventional sentiment classifier increased from 50% to 58.6% when stopwords were removed. The findings show that while eliminating stopwords significantly improves the traditional sentiment classifier's classification accuracy, other classifiers, including the sentiment term frequency, inverse document frequency, average relative term frequency sentiment classifier, and sentiment term frequency sentiment classifier, show no discernible changes.
Comparison research on text pre-processing methods on Twitter Sentiment Analysis [9]	Jianqiang Z, Xiaolin G.	Evaluation of pre-processing methods: Six pre-processing techniques are evaluated in the study: enlarging acronyms, reversing repeated letters, eliminating stop words, eliminating numerals, and replacing negation. The goal of the study was to ascertain how these techniques affected the F1-measure and accuracy of sentiment categorization using four classifiers: Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM), and Logistic Regression (LR) across five Twitter datasets. Impact on classifier performance: The study discovered that a few pre-processing techniques, like replacing negation and enlarging acronyms, greatly enhanced classification performance. On the other hand, eliminating integers, stop words, and

		URLs had little effect on classifier performance, indicating that these techniques only function to lower noise rather than improve classification accuracy. Dataset and classifier sensitivity: The study emphasizes how different datasets and classifiers can have different effects on how effective pre-processing techniques are. For example, compared to Logistic Regression and Support Vector Machine classifiers, Naive Bayes and Random Forest classifiers were found to be more sensitive to changes in pre-processing techniques. Conclusions: The study comes to the conclusion that improving Twitter sentiment classification requires careful consideration of the pre-processing techniques chosen. To improve sentiment analysis performance, techniques like expanding acronyms and replacing negation are advised. On the other hand, techniques like eliminating stop words, URLs, and digits can be used to streamline the dataset without appreciably hurting categorization results.
Evaluating preprocessing techniques in text categorization [10]	Srividhya V, Anitha R.	Importance of preprocessing: According to the study, preprocessing can account for 50% to 80% of the overall text categorization process, underscoring its crucial significance in the process. This emphasizes how important it is to perform adequate preprocessing in order to increase the effectiveness and precision of text classification models. Preprocessing techniques:

		• Stop Word Removal: Eliminating frequently used, non-informational words (such "the" and "and") that don't help organize texts into categories. • Stemming is the practice of condensing words to their base or root form (e.g., "running" to "run") in order to improve classification accuracy and group related keywords together. • Term Frequency-Inverse Document Frequency, or TF/IDF, is a weighting technique that helps distinguish between texts by assessing a word's relevance inside a document in relation to a corpus. Experimental results: The Reuters 21578 dataset was used in the study's experiments using these preprocessing methods. The results showed that by lowering the quantity of superfluous features and raising the relevancy of important phrases, preprocessing—specifically, stop word removal and TF/IDF—significantly enhanced the performance of text categorization models. Conclusion: The study comes to the conclusion that efficient text classification requires effective preprocessing, with TF/IDF and stop word removal being especially crucial for enhancing system performance. The results indicate that careful selection of preprocessing procedures is necessary to maximize the efficacy and precision of text categorization models.

Optimizing sentiment classification using preprocessing techniques [11]	Ghag K, Shah K.	Preprocessing techniques for sentiment classification: The significance of preprocessing in improving the accuracy of sentiment categorization is emphasized in the paper. Tokenization, punctuation removal, stopword removal, and other techniques were investigated; special emphasis was paid to the processing of apostrophe-containing terms. The study suggested particular guidelines for handling terms like "I'm" and "isn't" in order to decrease dimensionality and enhance model performance. Impact of preprocessing on different classifiers: The study assessed how different preprocessing methods affected the performance of several sentiment classifiers, such as the Relative Term Frequency Sentiment Classifier (RTFSC), Senti-TFIDF, Delta-TFIDF, Average Relative Term Frequency Sentiment Classifier (ARTFSC), and Traditional Sentiment Classifier (TSC). Preprocessing, particularly managing apostrophes and punctuation, greatly increased categorization accuracy, according to the study. Conclusions: The study found that the suggested sentiment classification models outperformed conventional models in terms of accuracy, particularly when paired with enhanced preprocessing methods. The study recommended that idea adaptability be integrated into sentiment classification models as the primary area of future research.

Public perspective on renewable and other energy resources: Evidence from social media big data and sentiment analysis [12]	Dahye Jeong, Syjung Hwang, Jisu Kim, Hyerim Yu, Eunil Park	Sentiment analysis on social media: By examining data from Reddit, the study utilized sentiment and correlation analyses to explore public attitudes toward various energy resources. This approach provided insights into the public's perception of energy resource relationships, such as viewing renewable and fossil fuels as substitutes. Monthly keyword and correlation analysis: The study conducted detailed monthly analyses of public sentiment towards different energy sources, showing how events and discussions on social media can influence public perception and sentiment over time.

3.1.3. Critique of the literature

a. Limited focus on preprocessing techniques

Though many different approaches have been thoroughly investigated in the literature on sentiment analysis, there is still a clear lack of focus on text preparation methods. Even though sentiment analysis has gotten a lot of attention, preprocessing plays an important but frequently overlooked function. This error could result in less than ideal classification performance because sentiment analysis model accuracy is greatly impacted by preprocessing. For this reason, it is essential to incorporate preprocessing techniques into sentiment analysis studies in order to improve classification results.

b. Diverse and complex preprocessing techniques

Reviewing a variety of preprocessing techniques, including emoticon handling, tokenization, stemming, and lemmatization, the study shows the varied effects of each strategy on classification accuracy. Although certain techniques, such as stemming, can enhance performance, others might have neutral or even detrimental impacts. This result emphasizes how difficult preprocessing may be and shows that a one-size-fits-all strategy might not work.

Personalized preparation techniques are required, based on the particular dataset and classification task in question.

c. Neglect of contextual factors

The possible disregard of contextual elements that could affect how successful preprocessing methods are a major criticism of previous research. For instance, emoticons have different effects depending on the context in which they are used. This variation implies that their impacts are not always favorable or unfavorable, emphasizing the necessity for more complex strategies that take into account the particular circumstances of the text under study.

3.1.4. Knowledge gap and justification

The accuracy of sentiment classification has significantly improved thanks to existing research in sentiment analysis and text preprocessing techniques, especially when applied to structured datasets like movie reviews. However, there is a noticeable lack of application of these methodologies in the context of community engagement strategies and renewable energy policy. The majority of recent research, including that conducted by Ghag and Shah (2015), has concentrated on conventional domains such as customer sentiment in product reviews, rather than applying the same techniques to more intricate, multifaceted domains like policy interpretation and mapping stakeholder interactions in the renewable energy industry.

Furthermore, there are other difficulties that are not covered in the body of current literature because of the distinct sociopolitical and cultural dynamics of the Indo-Pacific area. These difficulties include the complexity of policy papers, the diversity of languages spoken, and the disparities in public participation. These necessitate a more sophisticated approach to sentiment analysis and text preprocessing. The complexity of reading policy documents and mapping stakeholder interactions in such a varied and regionally distinct context is not fully taken into account by the existing methodologies.

By expanding and modifying current sentiment analysis and text preparation methods to better fit the field of renewable energy policy, this research seeks to close this gap. Additionally, it will incorporate SNA to map and evaluate stakeholder interactions—an essential step in promoting an Indo-Pacific region-wide cooperative approach to the uptake of renewable energy. The project aims to improve public sentiment and stakeholder dynamics by creating a specialized insights

platform. This will help to create renewable energy policies and community engagement methods that are more inclusive and effective.

3.2. Methodology

In order to investigate and evaluate the effects of text preprocessing on sentiment analysis in the context of renewable energy policy and community participation in the Indo-Pacific region, this study uses a literature review on preprocessing data

3.2.1. Research Design

In order to methodically examine and contrast several forms of literature reviews on the subject of preprocessing text data in natural language processing (NLP), this study uses a literature review technique. The aim is to ascertain, assess, and amalgamate diverse techniques employed in text data preparation, accentuating their methods, uses, and possible constraints.

3.2.2. Search Strategy

A thorough search was done using a variety of scholarly resources, such as PubMed, IEEE Xplore, Google Scholar, and library catalogs. "Text preprocessing," "NLP preprocessing techniques," "text data cleaning," "text normalization," "tokenization," "stemming," "lemmatization," and "handling emojis and emoticons" were among the search terms that were used. In order to guarantee the inclusion of recent and pertinent studies, the search was restricted to peer-reviewed articles, conference papers, and reputable books published within the last ten years.

3.2.3. Inclusion and Exclusion Criteria

The inclusion criteria for selecting studies were:

- Peer-reviewed articles, conference papers, and authoritative books.
- Studies focused on preprocessing techniques for text data in NLP.
- Publications in English.
- Studies published within the last ten years.

The exclusion criteria were:

- Non-peer-reviewed articles and grey literature.
- Studies not directly related to text preprocessing in NLP.
- Publications in languages other than English.
- Studies published more than ten years ago.

3.2.4. Data Extraction and Synthesis

The selected studies were reviewed and categorized based on the type of literature review they discussed. The main categories included:

Systematic Review: involves using a methodical, systematic approach to find, assess, and compile all pertinent literature on a certain text preprocessing research subject.

Narrative Review: summarizes the results of numerous investigations and offers a thorough review of text preparation methods without using a methodical search approach.

Thematic Analysis: finds and examines themes or trends in qualitative data about text preparation.

Methodologies, benefits, and drawbacks of the various review types were examined for every category. In order to create the synthesis, the results from many studies were compared and contrasted, common themes were found, and any research gaps were highlighted.

3.2.5. Data Preprocessing: Techniques and Justification

a. Tokenization

Tokenization is dividing the text into discrete pieces, called tokens, that are words or sentences. In order to transform unstructured text into a format that can be analyzed, this step is essential [13].

b. Lowercasing

To standardize the data and avoid duplicate tokens (such "Energy" and "energy") that merely differ in case, all text will be changed to lowercase [14].

c. Stopword removal

Words like "the," "and," and "is," which are frequently used, will be eliminated from the text as stopwords. These terms can complicate the analysis because they usually have little meaning [15].

d. Stemming and lemmatization

Words will be lemmatized or stemmed down to their basic or root forms. Lemmatization takes the context into account and returns words to their basic form, whereas stemming strips suffixes in order to identify the root word [16].

e. Handling special characters and punctuation

To lower noise in the text data, special letters and punctuation will be eliminated or standardized. In order to handle apostrophes and minimize dimensionality, contractions (such as "isn't" to "is not") must be converted [17].

f. Emoticon and emoji handling

In order to preserve the meaning that emojis and emoticons express without adding noise to the analysis, they will either be standardized or substituted with equivalent language [18].

g. Misspelling correction

Misspelled words will be automatically corrected by an autocorrect tool, guaranteeing reliable text data and lowering the possibility of errors in the analysis [19].

The rationale behind the adoption of these strategies stems from their demonstrated effectiveness in previous studies, including Ghag and Shah (2015). These methods are well known for their ability to lower data noise and increase sentiment analysis model accuracy.

3.2.6. Quality Assessment

Based on their methods, data analysis, and findings, the chosen studies' quality was evaluated. In the synthesis, studies with strong methodology and distinct, convincing results carried higher weight. The analysis took into account and took note of any potential biases or limitations in the studies.

3.2.7. Reporting

With sections devoted to each kind of literature review, the results of the review are presented in an organized manner. A synopsis of the main conclusions, a discussion of the implications for NLP, and recommendations for further research are included in each section.

3.3. Part B

3.3.1. Background of the Project

There is increasing pressure on the Indo-Pacific region to transition to renewable energy in order to address concerns related to energy security and climate change. But this change is complex, involving many different parties, laws, and public opinions.

Key challenges in this region include:

- Policy fragmentation refers to the uneven and disorganized renewable energy policies that exist in different nations and areas.
- Limited stakeholder engagement refers to the underutilization of key stakeholders in the creation and execution of policies, such as companies, NGOs, and communities.
- Variability in public sentiment: Varying degrees of public support and resistance to efforts involving renewable energy.
- Lack of insights and paucity of data: Inadequate access to the full data and analytical tools required for making well-informed decisions.

The project aims to address these challenges by:

- Providing a comprehensive grasp of the complex relationships that exist between stakeholders, policy, and public opinion in the renewable energy industry.
- Highlighting chances for important actors to work together and cooperate.
- Enabling decision-makers and interested parties to make well-informed choices based on insights supported by evidence.

3.3.2. Project goals and Objective

- Analyze policies pertaining to renewable energy to identify important themes, players, and connections.

- Construct and display networks that illustrate these linkages, including the links between policies, interactions between stakeholders, the impact of public opinion, and regulatory effects.
- Provide a dashboard that is easy to use so that stakeholders and policymakers can examine and understand network data.
- Provide practical insights to assist strategic planning and well-informed decision-making in the renewable energy industry.
- Encourage cooperation and coordination between the region's major players in the Indo-Pacific.

3.3.3. Desired outcome and benefits

a. Outcome

- Improved policy formulation is the process of formulating choices based on insights derived from evidence.
- Increased collaboration and cooperation amongst important participants in the renewable energy sector can be achieved through improved stakeholder involvement.
- Increased public support: increasing the public's knowledge of and encouragement for renewable energy projects.
- Encourage the uptake of renewable energy practices and technology to facilitate a quicker transition to a sustainable energy future.

b. Benefits

- Policymakers: Acquire practical knowledge to create sustainable energy policies.
- Stakeholders: A deeper understanding of the complex relationships and cooperative potential within the renewable energy industry.
- Communities: Better access to renewable energy projects and increased participation in decision-making processes.
- Environment: Improved air quality and reduced greenhouse gas emissions.
- Economy: Creation of jobs, expansion of the economy, and energy independence.

3.3.4. Learning issue

- How can unstructured text data, including policy documents and social media posts, be efficiently mined for important concepts, entities, and relationships using natural language processing techniques?
- How do we build and represent networks that truly capture the complex interrelationships among stakeholders, public opinion, and renewable energy policy?
- How can the quality and dependability of the conclusions drawn from network analysis be guaranteed?

3.3.5. Project scope and exclusions

a. In-scope

- Data collection: Obtaining policy documents, surveys of public opinion, and information on stakeholders from designated sources.
- Data processing is the process of organizing, cleansing, and getting the gathered data ready for analysis.
- Natural language processing (NLP) is the process of taking important information out of unstructured text input using NLP techniques.
- Network construction: Building networks to show the connections between stakeholders, public opinion, and policy is known as network construction.
- Network analysis: Locating isolated clusters, core linkages, and important nodes inside the built networks.
- Dashboard development: The process of developing an interactive dashboard to display network data.

b. Out-of-scope

- Real-time data updates: Constantly adding fresh data to the dashboard.
- Technical support: Providing continuous help with dashboard upkeep and operation.
- Forecasting future policy results is known as predictive modeling.

3.3.6. Project deliverables

Table 3.2 Project deliverables

Tasks and details	Individual or Group	Weighting	Assessment Due Date	Details
Individual research and report	Individual	20%	30th August 23:59 pm	individual Research and Report Each team member is responsible for conducting their own research and producing a comprehensive report on a specific aspect of the project, demonstrating their personal insights and findings.
Team innovation concept	Team	25%	20th September 23:59 pm	Team Innovation Concept The team works together to create a novel concept or solution that addresses the project's objectives, blending

				individual ideas and research into a unified approach.
Team project demonstration/presentation	Individual	25%	18th October 23:59 pm	The team showcases their project through a presentation, demonstrating their developed concept, key insights, and the potential impact of their work in an engaging format.
Individual project report	Individual	25%	28th October 23:59 pm	Individual Project Report Each participant submits a detailed report highlighting their specific contributions, individual research

				efforts, and reflections on the project's overall progress and outcomes.
Peer assessment	Individual	5%	28th October 23:59 pm	Team members assess each other's contributions, providing feedback on teamwork, participation, and effectiveness, ensuring fair recognition of each member's efforts.

3.3.7. Project management plan

a. Timeline

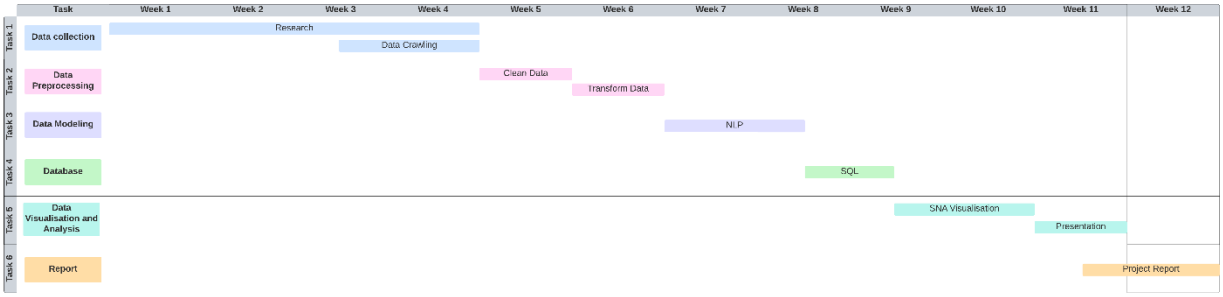

Figure 3.1 Project management timeline

b. Goals and Milestones

Table 3.3 Project management goals and milestones

Milestones	Goals
End of week 1	Start the project Do some research
End of week 2	Do some research Assign the duties of each member
End of week 4	Finish collecting raw data
End of week 6	Completely clean data Start transform data
End of week 8	Finish Natural Language Processing (NLP), including Sentiment Analysis, Text Classification, Network Analysis
End of week 9	Send results data to relational database Start building dashboard
End of week 10	Final dashboard is completed Embedding Dashboard to webpage User Testing completed
End of week 11	Project presentation
End of week 12	Finish and submit report

c. Team breakdown and duties

Table 3.4 Team breakdown and duties

Student name	Title	Responsibilities	Duties
HA TRANG NGUYEN	Project leader	Schedule meetings, track progress, handle communication with stakeholders, and resolve any issues that arise.	Schedule meetings, track progress, handle communication with stakeholders, and resolve any issues that arise.
TUAN DAT NGUYEN	Data analyst	Collect, process, and analyze data.	Gather data from various sources, clean and structure data, apply NLP techniques, and document the data processing steps.
JUNLADIT ROWSATHIEN	Network Analyst	Construct and analyze networks representing relationships between policy, public opinion, and stakeholders.	Develop network models, perform network analysis, identify key nodes and clusters, and interpret the results.
PINSAWAN SRISAI	Dashboard Developer	Design and build the interactive dashboard for visualizing network data.	Develop the dashboard interface, and ensure user-friendliness.
NAPATCHADA ROOKEN	Document and technical writer	Compile the reports and, prepare any documents needed submitting to clients, and prepare the presentation.	Write and edit the reports, create visual aids for the presentation, and ensure all documentation is clear and comprehensive.

Word count: 5027

REFFERENCES

[1] R. Glasser, C. Johnstone, and A. Kapetas, *The geopolitics of climate and security in the Indo-Pacific*. Australian Strategic Policy Institute, 2022.

[2] K. Adnan and R. Akbar, "An analytical study of information extraction from unstructured and multidimensional big data," *Journal of Big Data,* vol. 6, no. 1, pp. 1-38, 2019.

[3] W. Haruko, "The" Indo-Pacific" concept: geographical adjustments and their implications," 2020.

[4] O. Ellabban, H. Abu-Rub, and F. Blaabjerg, "Renewable energy resources: Current status, future prospects and their enabling technology," *Renewable and sustainable energy reviews,* vol. 39, pp. 748-764, 2014.

[5] P. Tangney, C. Nettle, B. Clarke, J. Newman, and C. Star, "Climate security in the Indo-Pacific: A systematic review of governance challenges for enhancing regional climate resilience," *Climatic Change,* vol. 167, no. 3, p. 40, 2021.

[6] M. A. Palomino and F. Aider, "Evaluating the effectiveness of text pre-processing in sentiment analysis," *Applied Sciences,* vol. 12, no. 17, p. 8765, 2022.

[7] M. IŞIK and H. Dağ, "The impact of text preprocessing on the prediction of review ratings," *Turkish Journal of Electrical Engineering and Computer Sciences,* vol. 28, no. 3, pp. 1405-1421, 2020.

[8] K. V. Ghag and K. Shah, "Comparative analysis of effect of stopwords removal on sentiment classification," in *2015 international conference on computer, communication and control (IC4)*, 2015: IEEE, pp. 1-6.

[9] Z. Jianqiang and G. Xiaolin, "Comparison research on text pre-processing methods on twitter sentiment analysis," *IEEE access,* vol. 5, pp. 2870-2879, 2017.

[10] V. Srividhya and R. Anitha, "Evaluating preprocessing techniques in text categorization," *International journal of computer science and application,* vol. 47, no. 11, pp. 49-51, 2010.

[11] K. Ghag and K. Shah, "Optimising sentiment classification using preprocessing techniques," *International Journal of IT & Knowledge Management,* vol. 8, no. 2, pp. 61-70, 2015.

[12] D. Jeong, S. Hwang, J. Kim, H. Yu, and E. Park, "Public perspective on renewable and other energy resources: Evidence from social media big data and sentiment analysis," *Energy Strategy Reviews,* vol. 50, p. 101243, 2023.

[13] S. Vijayarani and R. Janani, "Text mining: open source tokenization tools-an analysis," *Advanced Computational Intelligence: An International Journal (ACII),* vol. 3, no. 1, pp. 37-47, 2016.

[14] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text preprocessing for text mining in organizational research: Review and recommendations," *Organizational Research Methods,* vol. 25, no. 1, pp. 114-146, 2022.

[15] J. K. Raulji and J. R. Saini, "Stop-word removal algorithm and its implementation for Sanskrit language," *International Journal of Computer Applications,* vol. 150, no. 2, pp. 15-17, 2016.

[16] T. Korenius, J. Laurikkala, K. Järvelin, and M. Juhola, "Stemming and lemmatization in the clustering of finnish text documents," in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, 2004, pp. 625-633.

[17] H. Woo, J. Kim, and W. Lee, "Validation of text data preprocessing using a neural network model," *Mathematical Problems in Engineering,* vol. 2020, no. 1, p. 1958149, 2020.

[18] A. A. Arifiyanti and E. D. Wahyuni, "Emoji and emoticon in tweet sentiment classification," in *2020 6th Information Technology International Seminar (ITIS)*, 2020: IEEE, pp. 145-150.

[19] C. P. Chai, "Comparison of text preprocessing methods," *Natural Language Engineering,* vol. 29, no. 3, pp. 509-553, 2023.