Identifying Gender Bias in Online Crime News Indonesia Using Word Embedding

Miftakhul Janah Sulastri
Department of Information Systems
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
janah1448@gmail.com

Nur Aini Rakhmawati
Department of Information Systems
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
nur.aini@is.its.ac.id

Rarasmaya Indraswari
Department of Information Systems
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
raras@its.ac.id

2023 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA) | 979-8-3503-0922-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAMIMIA60881.2023.10427911

Abstract— **In the digital era, news portals have become a primary source of information for millions of individuals. This study investigated the potential influence of word choice and gender representation in news on the public's perception of gender, emphasizing its implications for gender equality and human rights. Research has shown that the language used in news reporting can reflect gender bias, highlighting the significance of analyzing gender representation in crime-related news. A word-embedding model was employed to identify and mitigate bias in word representation and ensure fairness in the data analysis. This study aims to enhance our understanding of gender representation in Indonesian crime-related news and to apply word-embedding techniques to identify biases in word representation. The results indicate a potential bias in word embeddings, emphasizing the importance of addressing and mitigating biases in language models to avoid reinforcing unfair stereotypes.**

Keywords—Word2vec, PCA, Bias Gender, Word Embedding, Crime Online News

I. Introduction

In the digital era, access to information has become increasingly convenient through swiftly updated news portals. News websites have emerged as the primary source of information for most individuals. Indonesia has five popular news portals, including Detik.com, Tribunnews.com, Kompas.com, CNNIndonesia.com, and Suara.com, with millions of visitors [1]. It is crucial to emphasize that word choice and gender representation in news have the potential to influence the public's perception of gender. Gender discrimination is a global issue closely associated with gender equality, human rights, and equal opportunities for all individuals [2]. Gender equality issues encompass various aspects of life and require a holistic approach to address biases that may affect gender stereotypes [3].

Research has indicated that the language used in news reporting can reflect gender bias. In a study conducted by Bolukbasi et al. [4]., gender bias was identified in word representations of the word2vec model on Google News. Through PCA visualization, it was found that specific word representations exhibited significant tendencies towards certain genders. One notable finding was the significant difference in the representations of words associated with various professions. For example, the word "homemaker" tends to be more associated with women, while the word "maestro" tends to be more associated with men [4]. This reflects the gender stereotypes embedded in the training data used by the Word2Vec model, which in turn creates bias in word vector representations.

The issue of gender in the crime context requires significant attention. Globally, it is estimated that there will be approximately 81,100 intentional murders targeting women and girls in 2021, with Asia having the highest number of such homicides. In Indonesia, data show that despite the majority of crime victims being male, there is a yearly fluctuation in the percentage of female victims [5]. Furthermore, statistics from the Ministry of Women's Empowerment and Child Protection [6] indicate that a significant number of violence cases involve female victims, whereas the majority of perpetrators are male [5]–[7].

To analyze the representation of gender in crime-related news, a word-embedding model is employed in Natural Language Processing (NLP), which serves to represent gender in news, particularly in Indonesia. NLP is a branch of computer science and artificial intelligence that comprehends and analyzes human text and is utilized in various tasks, such as sentiment analysis, text classification, and natural language understanding. In this study, word embedding aids in identifying gender bias in news and associating words with crime types. This differs from other methods, such as TF-IDF, which merely calculates word frequency in text without understanding their meaning. The use of word embedding in text analysis is essential to avoid bias in word representation and to ensure fairness in data analysis and public policy [8]. It is important to note that bias in data or machine-learning models can result in inequality and have negative impacts on policies and public perceptions. Therefore, identifying and mitigating bias in data and machine-learning models is crucial for achieving social justice and ethical technology [9].

This study aimed to identify biases in word representation in datasets and mitigate gender stereotypes and discrimination that can influence public perception and fair policymaking. This study used news text data and a word-to-vector model. The evaluation involved PCA visualization. Word-to-vector (word2vec) was used to represent words as numeric vectors based on their textual context [10]. This study assesses the extent to which the model adheres to the principle of equality and avoids bias or discrimination. The main contribution of this research is to increase our understanding of gender representation in crime-related news in Indonesia and apply word-embedding techniques to identify bias in word representation in Indonesian language news.

There are five sections in this study. Section II elaborates on the methodology, discussing the dataset used, data preprocessing steps, word embedding model employed, bias identification method, and evaluation metrics utilized. Section III outlines the results and Section IV presents the conclusions. Finally, Section V provides details of the references used in this study.

979-8-3503-0922-5/23/$31.00 ©2023 IEEE

II. Methodology

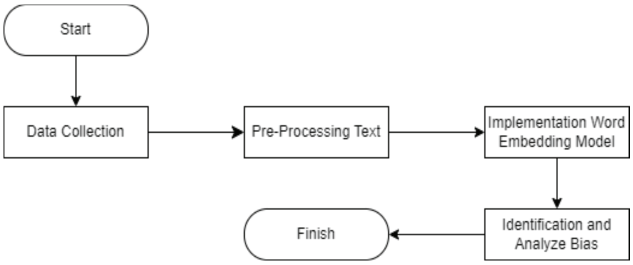
Start
Data Collection
Pre-Processing Text
Implementation Word Embedding Model
Identification and Analyze Bias
Finish

Fig. 1. Research work flow diagram

Fig. 1 shows a flowchart depicting the workflow used in this study. It begins with data collection, followed by data cleaning during the text preprocessing stage. The cleaned data will then go through the word-embedding model implementation phase. The word representation results are identified using PCA visualization and gender portion ratio calculation based on the proximity of target words to their surrounding words. A further explanation based on the flowchart above is as follows:

A. Data Collection

The dataset used in this research is a news with tagged as "crime," which were collected from the website https://www.detik.com/ using web scraping techniques with the implementation of the newspaper3k library. The extracted component is a summary of the news text, totaling 1560 articles. An example of input data in the form of a news summary is presented in TABLE I.

TABLE I. Example of a Raw Dataset

News summary
Seorang kakek bejat berinisial U (72) ditangkap karena diduga mencabuli bocah SD di Cipinang Muara, Jatinegara, Jakarta Timur. "Saat ini sedang menerima proses laporan polisi, dan akan mengajukan visum karena anaknya masih di luar kota," ujarnya. "Terus dia jalan arah pulang, tapi nggak ngomong nggak apa. Rekaman itu mulanya menampilkan pelaku yang menaiki sepeda sedang berbicara dengan bocah yang mengenakan seragam SD. Cuma pas itu saya penasaran anak itu diapain, makanya saya ke sini (pos RW) laporan, mau ngelihat cek CCTV, nggak tahunya sudah kejadian kayak gitu," jelas Erin. (A lecherous grandfather with the initials U (72) was arrested for allegedly molesting an elementary school boy in Cipinang Muara, Jatinegara, East Jakarta. "We are currently receiving a police report and will apply for a post-mortem because the child is still out of town," he said. "Then he walked home, but didn't say anything. The recording initially shows the perpetrator on a bicycle talking to a child wearing an elementary school uniform. "Just at that time I was curious about what the child was doing, that's why I came here (RW post) to report, I wanted to check the CCTV, I didn't know something like that had happened," explained Erin.)
Uba Rialin memberikan sanggahan atas sangkaan terhadap kliennya, Nenek MP (76), yang dijadikan tersangka kasus pencurian kemiri di Samosir, Sumatera Utara (Sumut). Uba menegaskan pencurian kemiri yang dilakukan oleh pelaku sudah sering terjadi dan sudah diketahui penduduk setempat. (Uba Rialin refuted the allegations against his client, Granny MP (76), who was named a suspect in the candlenut theft case in Samosir, North Sumatra (North Sumatra). Uba emphasized that the theft of candlenuts by the perpetrator had often occurred and was known to local residents.)

B. Pre-processing

Text processing is a series of stages used to prepare raw text data for more effective processing in Natural Language Processing (NLP), which is useful for obtaining the desired information, such as classification, sentiment analysis, and bias identification. This study consisted of four stages. Initially, all words in the dataset will be converted to lowercase to ensure uniform formatting in a step called "case folding." This was done to ensure consistency in text analysis and natural language processing, allowing for more consistent and accurate word comparisons and text processing [11] . The second stage is "removing punctuation," in which punctuation marks such as commas, question marks, periods, exclamation points, parentheses, and numbers are eliminated [11] . This is done to clean the text from elements that are not needed in text analysis or natural language processing, allowing for simpler and more consistent processing in specific tasks, such as text modeling or information retrieval.

The third stage involves removing stop words, names of individuals considered gender-neutral, and abbreviations that have little informational value or contribution to text analysis. Stop words are common words like "and," "or," "in," "to," "that," and the like that frequently appear in language and don't carry significant meaning when analyzed [12] . The goal of removing stop words is to clean the text and focus attention on more relevant and informative words in text analysis. The final stage is tokenization, which is used to separate text into distinct segments based on specific rules such as separation by spaces or punctuation [13] . The primary purpose of tokenization is to prepare text for further processing in text analysis. An example of a dataset after text pre-processing is presented in TABLE II.

TABLE II. Example of a Final Dataset

News Summary
kakek bejat berinisial ditangkap diduga mencabuli bocah cipinang muara jatinegara jakarta timur advertisement scroll resume contentsaat menerima proses laporan polisi mengajukan visum anaknya kota jalan arah pulang nggak ngomong nggak rekaman menampilkan pelaku sepeda berbicara bocah mengenakan seragam penasaran anak diapain laporan ngelihat cctv nggak tahunya kejadian kayak gitu erin (The depraved grandfather with the initials was arrested, suspected of molesting a Cipinang Muara Jatinegara boy, East Jakarta, advertisement scroll resume content, when he received the police report process to apply for his child's post-mortem, the city on the way home, he didn't talk, he didn't say anything, the recording showed the perpetrator on a bicycle talking, the boy was wearing a uniform, he was curious, what was the report of looking at the CCTV, didn't know that something like that had happened, Erin)
rialin sanggahan sangkaan kliennya nenek dijadikan tersangka pencurian kemiri samosir sumatera utara sumut pencurian kemiri korban penduduk diberitakan kasat reskrim polres samosir natar sibarani peristiwa desa onan runggu kecamatan onan runggu samosir nenek mengklaim lahan ditanami pohon kemiri kerap diambil nenek miliknya korban dulunya menanam kemiri situ tersangka dulunya nggak berdomisili onan runggu (Rialin refutes his client's suspicions that the grandmother was named a suspect in the theft of candlenuts in Samosir, North Sumatra, North Sumatra. The theft of candlenuts, the victim, was reported to the Criminal Investigation Unit of the Samosir Police, Natar Sibarani, the incident in Onan Runggu village, Onan Runggu sub-district, Samosir. The grandmother claimed that the land planted with candlenut trees was often taken by her grandmother. The victim used to plant candlenuts there. The suspect was not there before. domiciled in Onan Runggu)

C. Word Embedding

Word embedding is an NLP technique that can represent words as numeric vectors in a multidimensional space [14] and is used as a method for text feature extraction. Word-to-vector, or Word2Vec, is a word-embedding model trained through a neural network to generate word representations in vector dimensions while considering the semantic relationships between words [15]. The implementation process of Word2Vec begins by using unlabeled data as the input and producing word vectors as the output. The similarity between two word vectors was measured using the cosine similarity value, where a higher value indicates a closer semantic relationship between the two word vectors [16]. The architecture used in this study was a Continuous Bag of Words (CBOW). CBOW is an architecture used to predict the output word when a given input consists of words surrounding it. The input data for the CBOW architecture are in the form of an n-hot encoded vector, and the output data are in the form of a one-hot encoded vector.

In this implementation, it is crucial to choose the value of `**vector_size**` or the vector dimension used to represent each word in the vector space. The `**window**` value is used to control how many words around the target word will be considered during training. For example, if a window size o of n is given, the target word is compared to n “words before and” n words after it. Lastly, `**min_count**` refers to the minimum number of times a word must appear in the training data for it to be included in the vector generation process.

D. Analyze Model

Analyzing word representations using Principal Component Analysis (PCA) visualizations can aid in comprehending semantic relationships between words in the Word2Vec model. PCA is a statistical technique used for dimensionality reduction in data with the goal of identifying patterns of variability [17]. This is achieved by transforming the original data into a new coordinate system, known as the principal component, where the first principal component explains a significant portion of the variability in the data.

This analysis focuses on the types of crimes that appear around target words. Words used to describe gender, such as "laki-laki (male)-perempuan (female)", "pria (man)-wanita (woman)", "kakek (grandfather)-nenek (grandmother)", "mahasiswa (male student)-mahasiswi (female student)" "ayah (father)-ibu (mother)", and "suami (husband)-istri (wife)" will be analyzed to observe how these words are distributed in the vector space and whether any clusters form around them. Words that represent neutral gender, such as "anak (child)", "bocah (kid)", "bayi (baby)", “pelajar (student)”, “keluarga (family)”, and "pasutri (married couple)" will be disregarded.

III. Results and Discussions

In this study, the Word2Vec model was implemented on a pre-processed dataset with a **vector_size** of 100, a window size of 3, and a **min_count** of 10. The results of the model implementation were visualized using PCA with a top value of 200, which meant that PCA would display only 200 words around the target word. This selection was based on the amount of data used, as well as efficiency, to reduce the likelihood of irrelevant or even possibly noisy words.

Fig. 2 shows that word representations are spread horizontally and have clusters of words around x = 0 and y = 0.2. There are two outlier words, "pria" and "wanita," which are far from surrounding words. The target words around "wanita" appear to be fewer compared to the target words around "laki-laki." Words related to harassment, theft, and murder tend to be associated with "Perempuan" (woman), while "laki-laki" (man) appears to be more associated with various types of crimes such as rape, stabbing, hit-and-run, robbery, and others, as clarified in Fig. 3.

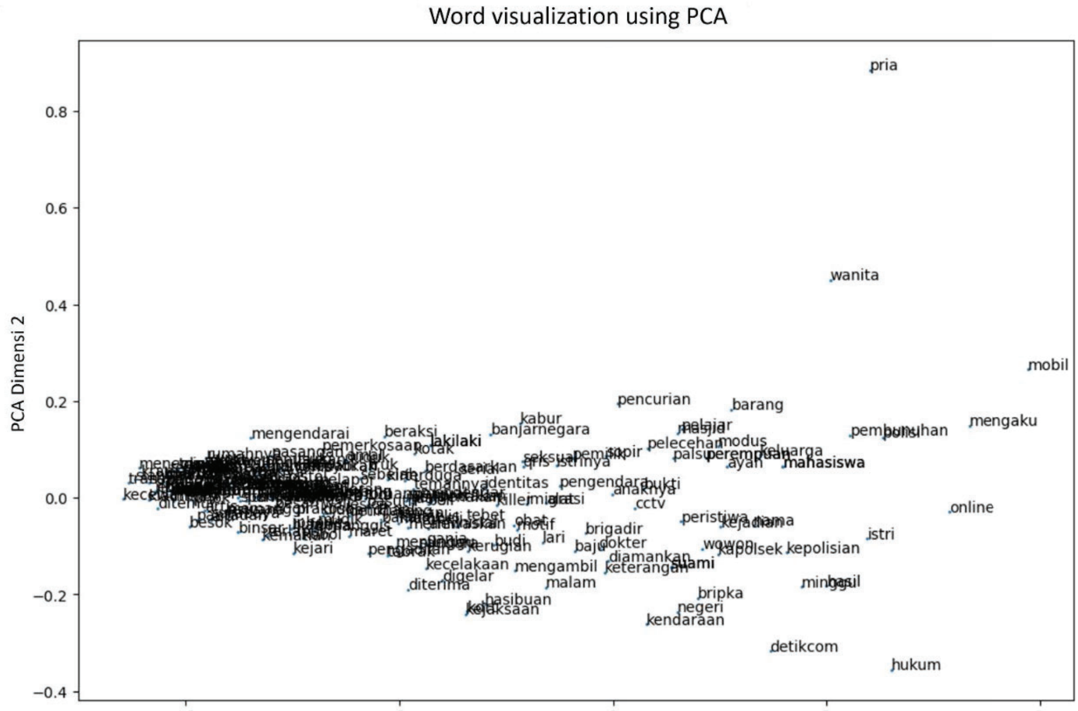

Fig. 2. Word visualization using PCA all part

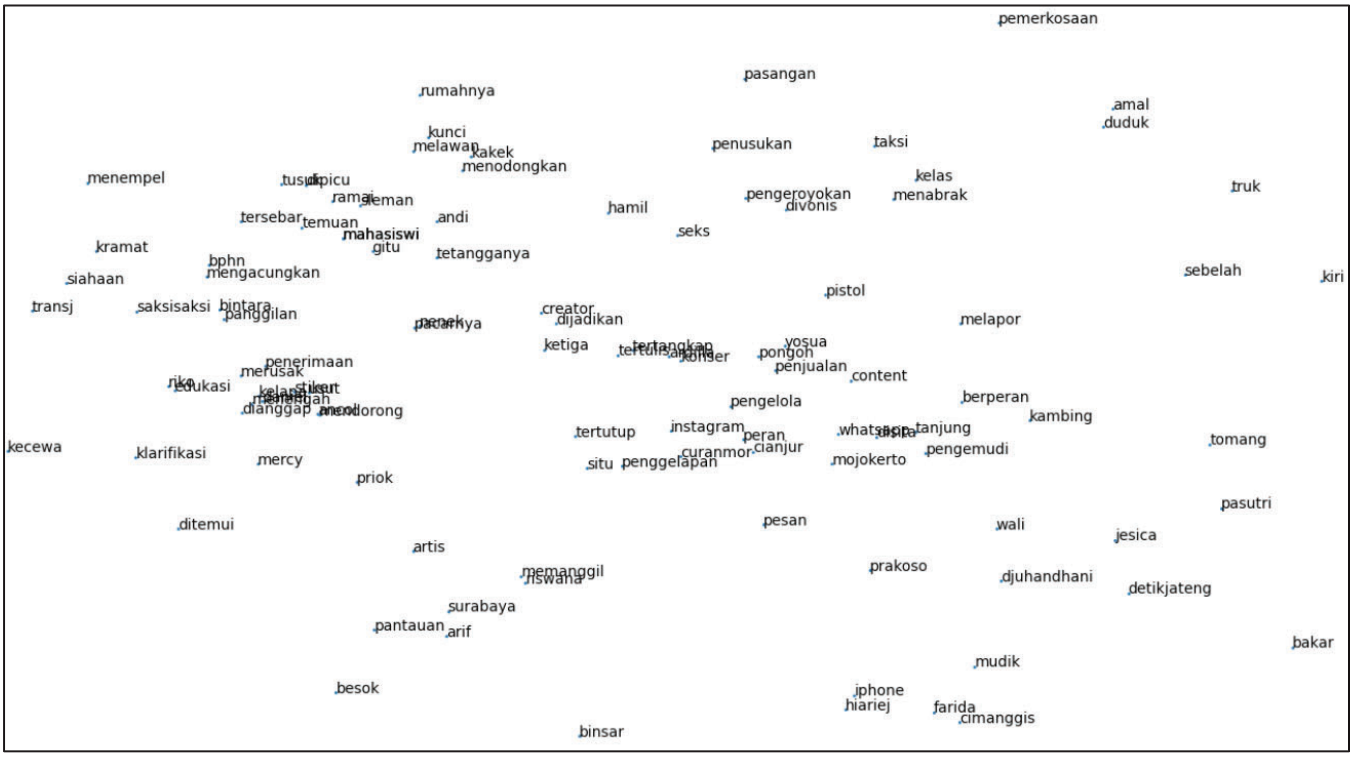

Fig. 3. Part of word visualization using PCA and zooming in overlapping parts

Based on TABLE III. and considering the five gender categories, it is evident that only one category displays a higher gender portion ratio for females: the [Suami – Istri] (Husband-Wife) category. In contrast, the other gender categories indicated a greater association with males in connection with different types of crimes. This observation raises concerns about the potential for bias in the data and highlights the need for closer examination. Specifically, in the [Pria-Wanita] (Man-Woman) category, where 84% of the references involve men in various crimes, there is a risk of reinforcing the stereotype that men are more frequently associated with criminal activities and aggression. This stereotype may not accurately represent the complexities and nuances of criminal behavior and gender dynamics.

TABLE III. GENDER PORTION RASIO AS VIEWED BASED ON GENDER CATEGORY AND CRIME TYPE ASSOCIATION

Gender Category	[Laki-Laki-Perempuan]	[Pria-Wanita]	[Kakek-Nenek]	[Mahasiswa-Mahasiswi]	[Suami-Istri]
Crime Type	*Pelecehan* (abuse)	*Korban* (victim)	*Jenazah* (corpse)	*Membunuh* (kill)	*KDRT* (domestic violence)
	Pengeroyokan (beating)	*Mayat* (corpse)	*Sabu* (methamphetamine)	*Jasad* (corpse)	*Ganja* (marijuana)
	Mutilasi (mutilation)	*Pelaku* (perpetrator)	*Narkotika* (narcotics)	*Pelecehan* (abuse)	*Sabu* (methamphetamine)
	Penusukkan (stabbing)	*Pelecahan* (abuse)	*Terorisme* (terrorism)	*Meninggal* (die)	*Penusukan* (stabbing)
	Mencuri (steal)	*Pencurian* (theft)	*Mutilasi* (mutilation)	*Mutilasi* (mutilation)	*Aborsi* (abortion)
	Pemerkosaan (rape)	*Pemerkosaan* (rape)	*Miras* (liquor)	*Mencuri* (steal)	*Pengeroyokan* (beating)
	Pembunuhan (murder)	*Pembunuhan* (murder)	*Menewaskan/ Tewasnya* (killed)	*Judi* (gambling)	*Tahanan* (prisoner)
Gender Portion Ratio (M = male)	56% *M*	84% *M*	56% *M*	69% *M*	48% *M*

IV. CONCLUSION

Based on the aforementioned explanation, this observation implies a potential bias in word embedding, particularly in the associations between gender and crime-related terms. These findings underscore the vital importance of addressing and mitigating biases within language models to prevent perpetuation or reinforcement of unfair stereotypes in different groups. The emphasis here is on critically analyzing and interpreting data in the gender context, as understanding these linguistic patterns and biases contributes to more balanced and informed discussions concerning crime, gender, and societal perceptions. Furthermore, it highlights the necessity for further research and a careful approach to interpreting and rectifying data bias.

ACKNOWLEDMENT

This study is supported by Institut Technology Sepuluh Nopember Surabaya with contract number: 1378/PKS/ITS/2023

REFERENCES

[1] Similarweb, “News & Media Publishers in Indonesia,” 2023. https://www.similarweb.com/

[2] A. S. Runyan, *Global Gender Issues in the New Millennium*, vol. 5, no. 1. 2015. doi: 10.32873/uno.dc.id.5.1.1116.

[3] N. Tabassum and B. S. Nayak, “Gender Stereotypes and Their Impact on Women’s Career Progressions from a Managerial Perspective,” *IIM Kozhikode Soc. Manag. Rev.*, vol. 10, no. 2, pp. 192–208, 2021, doi: 10.1177/2277975220975513.

[4] T. Bolukbasi, K. W. Chang, J. Zou, V. Saligrama, and A. Kalai, “Man is to computer programmer as woman is to homemaker? Debiasing word embeddings,” *Adv. Neural Inf. Process. Syst.*, pp. 4356–4364, 2016.

[5] BPS, “Statistik Kriminal 2020,” 2020.

[6] KemenPPPA, “Simfoni PPPA - Ringkasan Kekerasan,” 2023. https://kekerasan.kemenpppa.go.id/ringkasan

[7] United Nations Office on Drugs and Crime, “Gender-related killings of women and girls (femicide / feminicide),” *UNODC Search*, pp. 1–47, 2021.

[8] M. Babaeianjelodar, S. Lorenz, J. Gordon, J. Matthews, and E. Freitag, “Quantifying Gender Bias in Different Corpora,” *Web Conf. 2020 - Companion World Wide Web Conf. WWW 2020*, no. September, pp. 752–759, 2020, doi: 10.1145/3366424.3383559.

[9] B. Giovanola and S. Tiribelli, “Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms,” *AI Soc.*, vol. 38, no. 2, pp. 549–563, 2023, doi: 10.1007/s00146-022-01455-6.

[10] S. Al-Saqqa and A. Awajan, “The Use of Word2vec Model in Sentiment Analysis: A Survey,” *ACM Int. Conf. Proceeding Ser.*, no. June 2020, pp. 39–43, 2019, doi: 10.1145/3388218.3388229.

[11] A. Tabassum and R. R. Patil, “A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing,” *Int. Res. J. Eng. Technol.*, no. June, pp. 4864–4867, 2020, [Online]. Available: www.irjet.net

[12] K. Smelyakov, D. Karachevtsev, D. Kulemza, Y. Samoilenko, O. Patlan, and A. Chupryna, “Effectiveness of Preprocessing Algorithms for Natural Language Processing Applications,” in *2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T)*, 2020, pp. 187–191. doi: 10.1109/PICST51311.2020.9467919.

[13] R. Rahutomo, F. Lubis, H. H. Muljo, and B. Pardamean, “Preprocessing Methods and Tools in Modelling Japanese for Text Classification,” in *2019 International Conference on Information Management and Technology (ICIMTech)*, 2019, vol. 1, pp. 472–476. doi: 10.1109/ICIMTech.2019.8843796.

[14] L. Ding *et al.*, “Word Embeddings via Causal Inference: Gender Bias Reducing and Semantic Information Preserving,” *Proc. 36th AAAI Conf. Artif. Intell. AAAI 2022*, vol. 36, no. 2018, pp. 11864–11872, 2022, doi: 10.1609/aaai.v36i11.21443.

[15] Z. Adhari, F. Informatika, and U. Telkom, “Identifikasi Ujaran Kebencian pada Twitter Menggunakan Metode Convolutional Neural Network (CNN),” vol. 10, no. 3, pp. 3464–3474, 2023.

[16] D. Rahmawati and M. L. Khodra, “Word2vec semantic representation in multilabel classification for Indonesian news article,” *4th IGNITE Conf. 2016 Int. Conf. Adv. Informatics Concepts, Theory Appl. ICAICTA 2016*, pp. 0–5, 2016, doi: 10.1109/ICAICTA.2016.7803115.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *1st Int. Conf. Learn. Represent. ICLR 2013 - Work. Track Proc.*, pp. 1–12, 2013.