KKIK - FSRD

Jurnal Sosioteknologi

Website: https://journals.itb.ac.id/index.php/sostek/index

Social Identity Concept Adjustment in Hate Speech Corpus: A Computational Linguistics Approach

Penyesuaian Konsep Identitas Sosial pada Korpus Ujaran Kebencian: Pendekatan Komputasional Linguistik

Andika Dutha Bachari[1], Fauzan Novaldy Pratama[2], Zainul Muttaqin[3], Heri Heryono[4], Dinda Noor Azizah[5]

Education Indonesia Language and Literature, Universitas Pendidikan Indonesia, Bandung, Indonesia[1]
Linguistics, Universitas Pendidikan Indonesia, Bandung, Indonesia[2, 3]
English Language, Universitas Widyatama, Bandung, Indonesia[4, 5]

fauzan.novaldy92@upi.edu

https://doi.org/10.5614/sostek.itbj.2025.24.2.10
Submitted: October 21, 2024 Accepted: March 11, 2025 Published: July 21, 2025

ARTICLE INFO

Keywords:
hate speech, computational linguistics, semantic domain, natural language processing

ABSTRACT

The identification of hate speech must be accompanied by the identification of social identity concepts. This study aims to provide an alternative corpus with text metadata and social identity based on relevant laws that are designed to be implemented in machine learning. Two key questions are addressed: what social identity semantic domains are realized in the corpus, and what are the accuracy measurement results from the corpus? To achieve these aims, the study adopts a mixed-methods approach: qualitative for the first question and quantitative for the second. This research falls under the broader umbrella of computational linguistics, utilizing semantic domain theory and natural language processing. The first approach shows that the corpus only contributes five out of nine formulated domains, dominated by negative (uncategorized), religion, and ethnicity. The second approach indicates suboptimal conditions in the annotation distribution of the corpus, despite an average accuracy rate of over 80%. This condition limits the model's ability to generalize beyond the information within the corpus, especially regarding social identity categories that are not fully represented. This study differs from previous ones by focusing on data categorization based on more up-to-date legal sources. Future research could elaborate on this work by incorporating new language use concepts aligned with the corpus's original goal to detect hate speech.

INFO ARTIKEL

Kata kunci:
ujaran kebencian, linguistik komputasional, domain semantik, pemrosesan bahasa alami

ABSTRAK

Identifikasi ujaran kebencian harus dilakukan dengan identifikasi konsep identitas sosial. Penelitian ini berupaya memberikan korpus alternatif dengan metadata teks dan identitas sosial berdasarkan hukum terkait yang didesain untuk diimplementasikan pada pembelajaran Mesin. Untuk itu, terdapat dua pertanyaan yang perlu dijawab, yaitu apa saja domain semantik identitas sosial yang terealisasi dalam korpus serta bagaimana hasil akurasi pengujian dari korpus tersebut? Dengan tujuan tersebut, penelitian ini

mengadopsi metode penelitian campuran, yaitu kualitatif untuk pertanyaan pertama dan kuantitatif untuk pertanyaan kedua. Payung besar penelitian ini adalah Linguistik Komputasional yang memanfaatkan teori Domain Semantik dan Pemrosesan Bahasa Alami. Teori tersebut digunakan untuk memproses data korpus ujaran kebencian sebagai luaran dari penelitian sebelumnya yang tersedia secara open-source. Hasil analisis memproyeksikan adanya kondisi yang kurang layak dari segi distribusi anotasi pada korpus, walaupun luaran pengujian akurasi memunculkan angka rata-rata di atas 80%. Kondisi ini mengakibatkan model mesin memiliki kemampuan terbatas hanya pada informasi di dalam korpus saja, sedangkan terdapat kategori identitas sosial yang pengetahuannya tidak termuat dalam korpus. Penelitian ini membedakan dengan penelitian sebelumnya dengan berfokus pada kategorisasi data berdasarkan sumber hukum terkait yang lebih mutakhir. Luaran penelitian dapat dilanjutkan dengan penambahan konsep penggunaan bahasa baru yang sesuai dengan tujuan awal korpus, yaitu mendeteksi ujaran kebencian.

Introduction

Hate speech has become a concern in global societies (Ghasiya et al., 2023, p. 3), especially with the rise of social media (Niemi et al., 2018, p. 5; Perifanos et al., 2021, p. 1), which eliminates spatial and temporal variables (Taradhita and Putra, 2021, p. 225). The rise of hate speech on digital platforms has resulted in various forms of discrimination and verbal violence, often aimed at specific social and personal identities, thereby constituting symbolic violence (Cortés, 2021, p. 8). In response to this phenomenon, the United Nations (2020) has developed guidelines as part of its commitment to addressing this issue. The UN defines social identities as the inclination of people to group themselves according to certain groups (Yang et al., 2025). This serves as the foundation for hate speech acts, along with two other essential elements: pejorative language and communication. This implicates the need for a fundamental and philosophical approach to carefully define a phenomenon as hatred (Kurniasih, 2019, p. 49).

In Indonesia, a country with a high number of social media users (Gumilang et al., 2024), hate speech effects have become more evident, particularly with viral social media incidents that provoke conflicts between different groups based on identity, which is well-known as political polarization (Wasilewski, 2019, p. 176). A notable example is the Jakarta regional election in 2018 and the presidential election in 2019. Both events exposed the dichotomy between two social groups, religious and nationalist, where both related to cultural identity, which often led to societal tensions and partition. During this period, the Internet, particularly social media platforms, served as a battleground for sporadic instances of hate speech. Sazali et al. (2022) reveal that around 52,290 mentions of hate speech using the word 'kadrun' and 25,290 comments using the word 'cebong' emerged as reactive insults, fostering hostility toward specific individuals or groups. This research suggests developing perspectives from both labels to reduce the opposition's reputation during political events.

This exposure resulted from two factors: a lack of legal and social literacy regarding organic speech and the presence of industrialized buzzer products. The first point is that if Indonesia had constructed the Electronic Information and Transactions Law (UU ITE) in 2016 as an amendment of the 2008 version, societal tensions in cyberspace would not have been dammed. Loads of exposure tended to be only statistical reports rather than enforcement, since the exposure itself far exceeds human monitoring capacity. Many researchers have highlighted these phenomena; for instance, although Indonesia has laws addressing electronic transactions, those laws are insufficient to effectively address buzzer activities (Neyasyah, 2020). The phenomenon of political buzzers, often used by campaign teams to influence elections, has led to increased hate speech and attacks on opposing candidates (Kurniawati, 2023). Research on hate speech in Indonesia has been limited in its critical approach, often viewing it as a personal rather than ideological issue (Sirulhaq et al., 2023).

The challenge lies in balancing the need to address hate speech with concerns about potential suppression of political opposition, a dilemma Indonesia has faced in the past (Ahnaf & Suhadi, 2014). To combat this issue, there is a need for clearer regulations on buzzers and their activities (Kurniawati, 2023), as well as social movements to promote tolerance and counter hate speech (Ahnaf & Suhadi, 2014). The spread of verbal hatred remains relevant, as hate speech continues to appear on social media even as this paper is being written. In 2023, the government enacted the third amendment to the Electronic Information and Transactions Law to clarify regulations regarding online actions, including verbal abuse. However, law enforcement needs a comprehensive analysis to determine whether a violation has occurred in a single case. Addressing hate speech in Indonesia requires a multi-faceted approach that includes technological advancements, legal frameworks, educational initiatives, and community engagement strategies (Oktavianus, 2022). Advanced technologies like natural language processing and deep learning enhance detection and mitigation of hate speech on social media. Strengthening digital literacy and promoting dialogue, tolerance, and understanding through community engagement and educational programs are crucial. Monitoring systems and media literacy can empower the youth to counter hate speech effectively.

From this perspective, incremental technology development serves as an alternative solution to measure the social phenomenon. Indeed, technology solutions should not be treated as judgmental results, but this offers a valuable approach for indication measurement. Surfing in some well-known journal publishers such as sciencedirect.com, springer.com, and Google Scholar, the keyword 'hate speech' has emerged as a digital technology trend to detect hate speech. This project draws digital scholars' attention to creating digital solutions within the context of social hate speech across countries, including Indonesia. This is aligned with the UN (2020, p. 38) to enhance the technological skills of researchers who monitor, collect the data, and analyze the trends of hate speech in a national context.

However, as discussions surrounding identity continue to evolve in society, there is an increasing need to update the existing corpus to include a more nuanced categorization that reflects contemporary socio-political contexts in Indonesia. As mentioned above, the third amendment of the Electronic Information and Transactions Law indicates that the corpus also needs to evolve in accordance with this legal adjustment. In terms of social identity, the corpus includes religion, race, physicality, gender, and others. The term 'other' in this context does not define a specific conception. This is still relevant to juxtapose with the second amendment of UU ITE 2016, specifically article 28 paragraph (2) with the categories "…*suku, agama, ras, dan antargolongan* (SARA)." The third amendment, enacted in early 2024, specifically addresses article 28 paragraph (2), which includes the recent statement: "*ras, kebangsaan, etnis, warna kulit, agama, kepercayaan, jenis kelamin, disabilitas mental, atau disabilitas fisik.*" Some specific categorizations detail social identity definitions such as nationalities, skin colors, belief systems, and physical-mental disability. Implementing these classifications in a fair and comprehensive legal framework is a complex challenge, especially in multicultural countries with multidimensional social contexts that value every concept of individual and communal identities.

Computational linguistics, which combines theoretical linguistics with automatic computer programs and testing (Hausser, 2014, p. 8), plays a role as the grand approach. Theoretical linguistics serves as a data describer within a specific framework and conceptual model. In this research, the semantic domain, or semantic field, elaborates on intersecting conceptions between established parameters based on related law. Afterwards, this research requires employing a computational approach, as the second objective needs to measure the accuracy of corpus development.

As the nomenclature itself works with how language or linguistic features are computerized, this requires described or annotated linguistic data based on its purpose. For example, ChatGPT from OpenAI, Gemini from Google, and Copilot from Microsoft are chatbot-like artificial intelligences that enable natural-like interaction between humans and computers. On the computer side, large language models (LLMs) are used as statistical language knowledge with much fine-tuning from each developer. However, building LLMs requires an investment in cost and effort because of the need for reliable infrastructure and

accurate manual measurement from competent individuals. Even if some huge developers like Metaverse have built Lammas as a commercial source, this needs deep consideration with machine development and business processes.

Another key focus of this research is a predictive machine, which can forecast specific information based on the data train (data converted into machine knowledge) and its input. A well-known example of supervised machine learning of natural language processing (NLP) is sentiment analysis, in which the machine is capable of predicting a set of textual data into appropriate sentiment, whether positive, negative, or neutral. The prediction itself is based on data-training quality, unsavory data training results in poor prediction accuracy and vice versa.

As an example, in response to these urgent issues, several studies have been conducted to develop technologies for detecting hate speech using natural language processing (NLP) techniques. A contribution to this field is the hate speech corpus developed by Ibrohim and Budi (2019, 2023), which includes content from social media, especially Twitter, and is annotated with various categories of hate speech. This corpus serves as a crucial resource for creating automated detection models designed to recognize hate speech in Indonesian. Local researchers like Zakariya and Syafrullah (2024), Septiawan and Chairani (2023), and Taradhita and Putra (2021) have provided examples of hate speech detection in local scope. Moreover, international researchers like Raza and Chatrath (2024), Hartvigsen et al. (2022), Fonseca et al. (2024), Mu et al. (2024), Jahan and Oussalah (2023), Khanduja et al. (2024), Pan et al. (2024), Aggarwal and Vishwakarma (2024), and Ibrahim et al. (2024) have produced hate speech detection using an international language, whether it is a single or multiple languages. All of those researchers focused on developing appropriate predictive models. From this point of view, hate speech has come to involve not only linguistics matters but also digital technology, as this issue has appeared to be problematic.

Relating to the explanation of computational linguistics above, rigid conception and consistent annotation are the keys to data-train quality. Again, as previously explained in the introduction section, linguistics provides a specific approach to define a shared conception between word and text, namely semantic domain or semantic field. The semantic domain, which is a more in-depth taxonomy, covers three levels of linguistic features: text level, concept level, and term level (Gliozzo & Strapparava, 2009, p. 29). Taking previous examples 1-4 above, this cannot be enough to share intersecting information at the level of text, but it requires a conception level to see that examples 1 and 2 share a conception of ethnicity and examples 3 and 4 share religion.

Semantic domain or semantic field deals with intersecting conceptions between words and text (Allan, 2001, p. 258; Gliozzo and Strapparava, 2009, p. 13; Saeed, 2016, p. 316). The words 'red', 'white', and 'black', for example, share the same conception of being 'colored'. This conception comes from the same basic characteristic that those three words mostly collocate specific characteristics differentiating the same object. For example, defining a car as a vehicle with four tires and adding red, white, and black will specify the color character of the object. The semantic domain plays a role in justifying whether a statement has the potential to contain social identity information. For instance, Example 1, "*dasar orang Jawa*," or "typical Javanese people," and Example 2, "*ganyang Cina dari Indonesia*," or "eradicate Chinese people from Indonesia," project the identity conception of ethnicity. Other instances, such as Example 3, "*Muslim ga bisa dipercaya*," or "Muslims can't be trusted," and Example 4, "*Kristen agama ga jelas*," or "Christianity is an unclear religion," pertain to the identity of the religion. This shows the necessity of accurately annotating the proper categorization since it may cause different impacts. Ibrohim and Budi (2019, 2023) gave implicit illustrations of how semantic domain plays a role in categorizing contextual searches in the corpus. This approach has the potential to advance by employing specific theories and parameters to appropriately conceptualize the phenomenon. For example, semantic domain research has been done by Zahid (2020), Bakar et al. (2024), and Zahid et al. (2022).

The two examples of computational modeling and semantic domains presented above suggest distinct areas of investigation. However, this study seeks to integrate these foundational approaches. Although this research prioritizes the linguistic perspective by focusing on manual annotation according

to specific parameters and usage (rather than developing built-in computational models), it may have broader implications for related future studies. Thus, by identifying two research problems, which are (1) what semantic domains of social identities appear in the corpus and (2) how is the accuracy measurement realized, this research aims at re-annotating the existing corpus with an emphasis on current social identity categorization in the third amendment of the Electronic Information and Transactions Law utilizing a semantic approach and examining the result using in-built machine learning software. To answer the research questions and achieve its objectives, this study employs a computational linguistics approach that incorporates semantic domain conception and utilizes natural language processing (NLP). This research has potential in two ways: it strengthens the theoretical intersection of social science and computer science, and it serves as an alternative resource for practical needs.

Methods

As the objectives of the research intend to expose the current conception of social identity and its accuracy measurement, this paper applies a mixed qualitative and quantitative method. This is aligned with Huang et al. (2015), arguing that computational technology needs to apply both. A qualitative approach analyzes data to reach specific meaning (Leavy, 2017, p. 9) and explores it in depth (Creswell & Creswell, 2018, p. 303). This approach addresses the requirements for data description by utilizing semantic fields or semantic domains to annotate and analyze specific concepts within a domain. Then, a quantitative approach relates to proving a hypothesis using calculation (Leavy, 2017, p. 9), including probabilities. This approach is needed for accuracy measurement.

Figure 1 below shows the procedures for conducting this research. First is data collection. The data, or corpus, utilizes an existing open-source license created by developers Ibrohim and Budi (2019, 2023), provided that academic use includes proper citation. Corpora are available at https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection. This corpus is designed to be implemented in machine learning completed by metadata flagging such as abusive, hate speech, identity category, grouping or individual, and strength. It is mentioned before that this research attempts to adjust identity categories based on current related law. The corpus itself is a collection of natural language data sourced from Twitter (currently X), which has undergone preprocessing steps that include replacing URLs (for example, transforming "https://www.google.com" into "URL") and usernames (for example, transforming "@abcd" into "USERNAME"). Consequently, data should be recognized as coming from a secondary source, not a primary one.

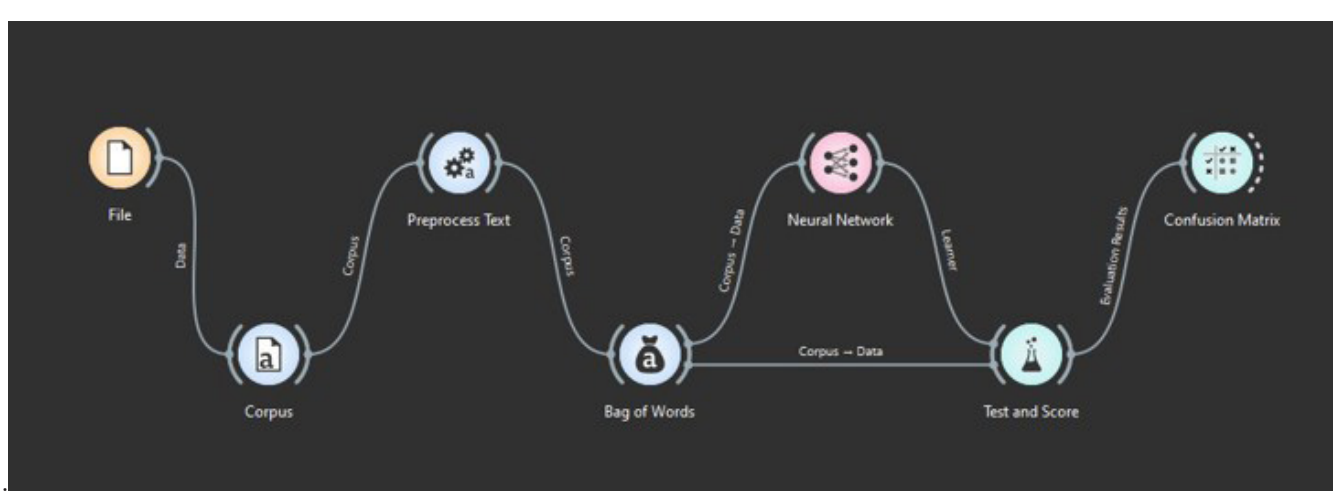

Figure 1 Machine learning flow using orange data mining
Source: Personal documentation, 2024

The second is semantic domain analysis. As mentioned before, this annotation process requires a rigid definition and a fixed conception free from bias; it necessitates referring to a strong and well-established parameter, specifically the Electronic Information and Transactions Law, paragraph 28, article (2), which addresses social identities. Those nine domains are race, nationality, ethnicity, skin color, religion, belief system, gender, mental disability, and physical disability. The analysis focuses on concepts contained in each data set by relating word choices to explicit content and corpus context.

The third step involves applying NLP to measure the accuracy of the corpus in machine learning. The annotation result in the second step is examined by utilizing NLP to see the effectiveness of corpus development. In other words, machine learning learns the corpus and tests its capability by undergoing accuracy measurement. Achieving the score utilizes in-built, stand-alone machine learning software, namely Orange Data Mining. Choosing this stand-alone software is based on its features that facilitate text processing, including built-in tools and a user-friendly interface.

Figure 2 illustrates the flow of widgets from left to right, beginning with File (which inputs a corpus file), followed by corpus (which converts the file into the required format), preprocess text (which transforms text based on specific needs), bag of words (which tokenizes words by their frequency), neural network (the chosen algorithm), test and score (which calculates accuracy results), and confusion matrix (which shows true-false predictions). This research applies to a neural network algorithm based on Jahan and Oussalah (2023), which said that the neural network is better than other traditional algorithms like logistic regression and random forest. All tunings use default settings to see how effective the corpus is in a quite common environment. For example, the neural network widget as an algorithm or machine model tuning uses a hyperparameter size of 100 height and 0 weight, using the ReLU activation function and 200 iterations. Figure 4 depicts a factual illustration of when the neural network widget is placed on the canvas. The last objective generates an accuracy measurement. As it is also featured in the software, the accuracy varies into five calculations: CA (classification accuracy), Prec (precision, true positives among predicted positives), recall (true positives among true predicted), F1 (average of precision and recall), and MCC (Matthew's correlation coefficient, appropriate for classes or categories in distinctive sizes). See Figure 3 for details of each function.

Results and Discussion

The data analysis leads to two main findings, as navigated by research objectives, which are semantic domain realization and accuracy measurement. The qualitative semantic analysis extracts information that the negative category dominates the corpus. In this context, 'negative' refers to data that does not reflect any conception of social identity. Religion and ethnicity follow dominance but still do not equal or have null annotation. Then, sexual orientation, gender, and nationality are minority under 1% in sum. Afterwards, only six out of eight categories appear within the corpus. To be clear, ethnicity and tribe are in a cohort, as they share similar conceptions of culture and tradition, different from race, which sees physical characteristics as variable. In total, five out of six domains are present, excluding the negative categories.

For example, the religion domain is represented in the data by the phrase *"Acara yg mendungukan & membiadabkan bangsa, ngaji budaya dari Muhammad Arab buta huruf ngaku nabi, ganti budaya BACA."* or "An event that stupefies and disgraces the nation, religious chanting of cultural teachings from Muhammad, an illiterate Arab claiming to be a prophet, replacing the culture of reading." Remember that the semantic domain used in this research intersects not only specific texts but also contains concepts, and the text exposes the concept of religion, specifically Islam. The expression brings up *'ngaji' or* 'Quranic recitation,' as a religious activity, reciting the Islamic holy book, and 'Muhammad' as a prophet in this religion. This analysis focuses on the communal identity concept contained in it. Therefore, based on the expression and word choices, the speaker (or caption writer) talks about religion. There may be other intentional meanings, such as insults, but those will fall outside the objectives of this research.

Other examples of expressions relating to religion within the corpus include *'agama'* (religion), *'anti-Islam*' (anti-Islam), *'kafir'* (infidel), *'Kristen'* (Christian), *'Budha'* (Buddhist), *'ulama'* (Islamic scholars), *'Quran,' 'solat'* (prayer), and *'ngaji'* (Quranic recitation). Mostly the expression comes from religion since the context of the corpus relates to specific phenomena mentioned in the background.

Within the ethnicity domain, the following expression examples are "*#MataNajwaDebatJakarta lucu banget jawaban* Ahok... *gak nyambung. Ditanya apa, dijawab apa. Kena skak mat d pertanyaan isu agama tadi. Dasar cina tolol hahaha,*" or #MataNajwaDebatJakarta was so funny, Ahok's answer... completely off-track. He asked one question, then responded with another. He was checkmated on the religious issue earlier. Typical stupid Chinese, hahaha. The word *'cina'* (Chinese) in this context intersects with ethnicity. Some other expressions are '*uyghur'* (Uyghur), '*aseng'* (Chinese foreigner), '*arab'* (Arabian), '*bani onta'* (camel group), and *'cebong'* (tadpole). Some terms, such as *'onta'* (camel) *and 'cebong'* (tadpole), use implicit symbolic expressions to convey specific ethnicities with negative sentiment.

The negative domain, as previously mentioned, does not include any concepts of social identity. Instead, it consists of random conceptions. For example, the statement "*KENAPA GUE NANGIS ANJIR, CUPU LO FIK!*" ("Why am I crying, damn it? "You're so weak, Fik!") does not convey any sense of social identity. Moreover, this domain exposure dominantly places the first rank.

After answering the first question, the next step is to use the annotation results to measure the accuracy of the reannotated corpus for addressing the second question. This abnormal and unbalanced distribution affects the accuracy score. The machine learning process produces accurate results, which are detailed in Figure 6, which shows that CA, F1 scores, precision, and recall generate values around and above 90%. Only the MCC metric has a calculation result below the others, but it remains within an acceptable range. Each calculation has its own advantages and disadvantages, which depend on the perspective of the data and the features being analyzed. To provide a simple explanation, this research emphasizes the importance of MCC while still acknowledging other calculations.

At the end, the process generates output in the form of a 'machine model'. This model, in general, is a statistical calculation involving the corpus parsing process and chosen algorithm, which is the neural network. In a noticeably big picture, the algorithm calculates any possibilities appearing when words combine to create a specific meaning, context, and conception. The model does not recognize meaning, context, and conception in the same way that humans do; instead, it produces statistical calculations to analyze word circumstances and generates a confidential score for all categories. In other words, getting closer to a perfect score projects a higher possibility that a word combination should be in a specific category.

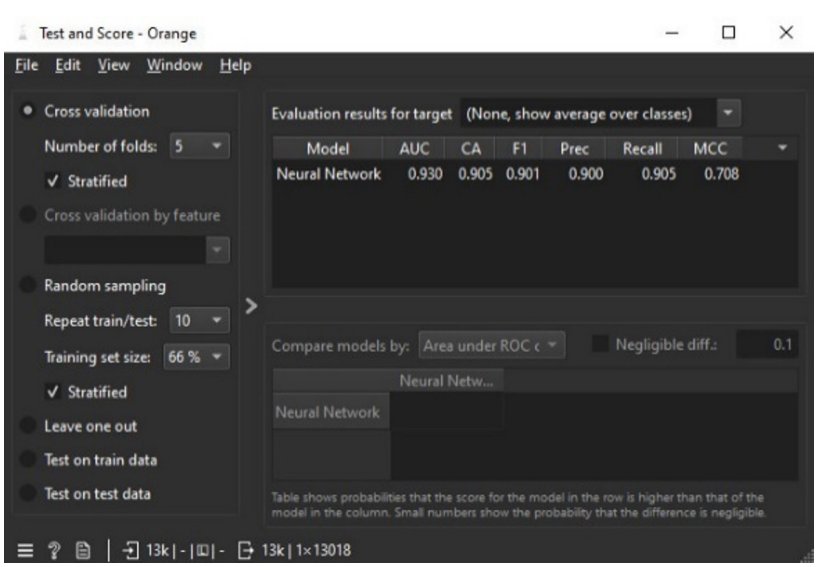

Figure 2 Accuracy measurement result
Source: Personal documentation, 2024

The narrative explanation provided here is based on the realization of domain semantics, which implicates data conditions, accuracy results, and ideal approaches, as the research objectives have been constructed to meet a certain flow. Starting with the first objective, the Electronic Information and Transactions Law in article 28, paragraph 2, still covers social identities within the corpus, so there is no obstacle in referring to the rigid domain.

Two problems lie in distribution. First, the corpus does not contain all social identity categorizations. As previously explained, this corpus is statistically converted into machine basic knowledge; lack of information also implicates a lack of machine capability to learn and predict content outside the corpus. For example, the machine does not recognize "*dasar lu cacat*" or "You're basically disabled" as speech having social identity content, as the corpus initially does not provide this specific information. This implies that word choices affect the domain conception. Using a partial example in the ethnicity domain above, "*Dasar cina tolol*" or "You are stupid Chinese." The word choice *'cina*' or Chinese indicates a social group based on specific ethnicity. The domain concept will be different; for example, when replacing the diction with '*Kristen*' or Christianity, then it will move the domain to religion, as the word brings a conception of it. Thus, diction and its context lead to the projection of specific social identities.

Second is unbalanced distribution. Ideally, machines prefer learning in a balanced distributed domain exposure or in acceptable differences. When the exposure is balanced, the machine also learns to extract knowledge effectively from the training data. It is even more preferred to construct a data train with balanced categorization and conception at the same time to generate proper probability knowledge. As a result, the reverse corpus condition negatively impacts the model, leading to inaccurate predictions. Imagine that a person only learns about certain information while being overwhelmed by a large amount of other information; this person may be able to deeply comprehend it. However, reality is not constructed by a single piece of information but appears to have immeasurable complexity. This person will survive with such complexity. Machine learning also suffers from a similar situation that results in inappropriate predictions outside its knowledge. For example, in the previous case, the statement *"Dasar cina tolol"* (you are stupid Chinese) uses the keyword *'cina'* (Chinese) to imply ethnicity. However, this keyword can also appear in different contexts, such as in the statement *"Cina negara yang besar,"* (China is a big country). The keywords move from the ethnicity domain into a geographical context (which is not a domain parameter formulated for this research objective). The machine will have broader knowledge as keywords are placed in different data contexts.

An in-depth explanation needs to understand the confusion matrix, which is a two-dimensional mapping between the actual value from manual annotation and the predicted value from the model. Referring to the confusion matrix calculation (Figure 4 on the left and the right), the previous assumption that an unbalanced annotation distribution leads to an inappropriate prediction result is proven by the calculation. True positive (the model correctly predicts annotated data) assembles in a negative value, in which the value strongly dominates the corpus; Figure 4 on the left. As a result, models have stronger consideration knowledge that results in biased predictions. A clearer picture is served in Figure 4 on the right, showing the calculation in percentage distribution based on the actual value. Excluding actual negative and predicted negative mapping, above 50% of wrong predictions to actual negative values dominate the overall distribution. This project models the tendency to negative value as the result of dominant negative annotation. On the other hand, this is quite sophisticated to see the religion and ethnicity having almost a 70% true positive score. This strongly proves that the more data, the better the prediction; the more correct and balanced the data, the better the prediction quality.

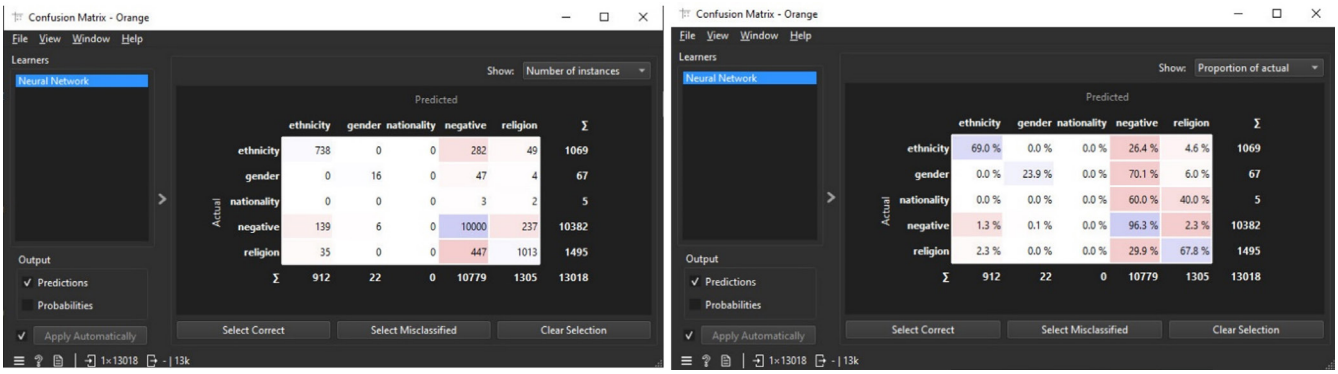

Figure 3 Confusion Matrix, (left) total data and (right) distribution based on actual side
Source: Personal documentation, 2024

Two measurements, the accuracy and the confusion matrix, are not contradictory, but they complement each other with different points of view. The first project, that combination of reannotated corpus and algorithm, works well in building the machine model. To be highlighted, the neural network, advancing machine learning into deep learning, remains an up-to-date algorithm for predictive machines. With an accuracy exceeding 90%, these models exhibit a sophisticated capacity to process and interpret corpus data by leveraging specific word combinations and weighted lexical features—where weighting is derived from the bag of words (tokenizer that processes word frequency). The second point discusses the potential for improving the corpus. Note that improvement on this side covers distribution only, assuming consistent corpus annotation. It is done, at least, in two alternatives: oversampling and augmentation. The first seems more convenient, as the process duplicates data until the distribution is balanced through coding. However, this alternative does not enhance the model in terms of knowledge variety since the process only reproduces existing information. The second has a good point on the possible variety of information since it is done by fabricating data with an existing data basis. Doing this manually using human effort will potentially result in distinctive content but costly for human labor effort.

This corpus has the potential to be redeveloped more than current conditions. For example, adding more open-source corpora in Bahasa Indonesia that are manually annotated using this research approach may lead to a better distribution of categories. Another potential, more focused effort is data augmentation. This alternative step allows researchers to (re)create additional data based on existing data (Hu, Zhang, & Zhang, 2023; Safari & Shamsfard, 2024; Shim, Lowet, Luca, & Vanrumste, 2021; Tareq, Islam, Deb, Rahman, & Mahmud, 2023). For example, a speech with verbal abuse containing the context of religion can be modified into verbal abuse containing physical disability as naturally as possible. This approach aims to balance the distribution of categories, which will improve accuracy and enhance the calculation of the confusion matrix. In the end, this research and similar ones attempt to contribute to predicting hate speech utilizing artificial intelligence.

Conclusion

This research attempts to reformulate (without canceling the existing) alternative corpus annotation based on Indonesia's Law of Electronic Information and Transactions. This research results in a corpus with current formulated social identity domains within the law reference that differ from the original one. Even if it does not cover all domains, this corpus is still applicable for machine learning development. The results of the corpus measurement are applicable for identifying social identity concepts within its knowledge base. However, it should be highlighted that the corpus was based on specific events in the past. While certain conceptual frameworks remain relevant—at least up to the time of this paper's writing—language usage continuously evolves. Moreover, this research focuses on social identity while recognizing the need for different perspectives, including a language usage approach.

Further research may continue utilizing this corpus by integrating additional conceptual frameworks. For instance, as outlined in the introduction, hate speech is characterized by three core elements: communication, verbal abuse, and social identity. This opens the possibility of refining annotation through explicit verbal abuse flagging. However, as this research developed alternative annotation employing academic and social science approaches, abuse annotation required employing scientific language use approaches such as speech acts or face-threatening acts. Furthermore, more sophisticated machine learning techniques, such as directly utilizing Python, may provide greater flexibility in fine-tuning compared to built-in features.

In principle, digital technology, particularly artificial intelligence, serves as an enabler and enhancer of human productivity in specific activities aligned with its intended development purposes. Digital problems can be more effectively addressed through digital solutions, reinforcing efficiency and adaptability. While AI-driven outcomes should be regarded as recommendations rather than definitive judgments, they nonetheless open up vast possibilities and significantly enhance both the quality and quantity of human-driven work.

References

Aggarwal, S., & Vishwakarma, D. K. (2024). Exposing the Achilles' heel of textual hate speech classifiers using indistinguishable adversarial examples. *Expert Systems with Applications*, *254*(October 2023), 124278. https://doi.org/10.1016/j.eswa.2024.124278

Ahnaf, M. I., & Suhadi. (2014). Isu-isu kunci ujaran kebencian (hate speech): Implikasinya terhadap gerakan sosial membangun toleransi. *Jurnal Multikultural & Multireligius*, *13*(3), 153–164. Retrieved from http://www.youtube.com/

Allan, K. (2001). *Natural language semantics*. Oxford: Blackwell.

Bakar, N. A., Zahid, I., Jaafar, M. F., & Ali, W. Z. K. W. (2024). The mapping and classification of shariah's semantic domain based on semantic relations of Arabic loanwords lexical. *Pertanika Journal of Social Sciences and Humanities*, *32*, 1–27. https://doi.org/10.47836/PJSSH.32.S1.01

Cortés, I. (2021). Hate speech, symbolic violence, and racial discrimination. Antigypsyism: What responses for the next decade? *Social Sciences*, *10*(10). https://doi.org/10.3390/socsci10100360

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (Fifth ed). Los Angeles, London, New Delhi, Singapore, Washington DC, Melbourne: SAGE Publications, Inc.

Fonseca, A., Pontes, C., Moro, S., Batista, F., Ribeiro, R., Guerra, R., Carvalho, P., Marques, C., & Silva, C. (2024). Analyzing hate speech dynamics on Twitter/X: Insights from conversational data and the impact of user interaction patterns. *Heliyon*, 10(11), e32246. https://doi.org/10.1016/j.heliyon.2024.e32246

Ghasiya, P., Ahnert, G., & Sasahara, K. (2023). Identifying themes of Right-Wing extremism in Hindutva discourse on Twitter. *Social Media and Society*, 9(3). https://doi.org/10.1177/20563051231199457

Gliozzo, A., & Strapparava, C. (2009). *Semantic domains in computational linguistics*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-68158-8

Gumilang, M. A., Abdillah, F., Amin, M. Y., & Hasan, M. (2024). Sentiment analysis of Indonesian ministries social media: Citizen responses utilizing TextBlob analyser. *Jurnal Sosioteknologi*, *23*(2), 203–216. https://doi.org/10.5614/sostek.itbj.2024.23.2.5

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., & Kamar, E. (2022). ToxiGen: a Large-Scale Machine-Generated dataset for adversarial and implicit hate speech detection. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/2022.acl-long.234

Hausser, R. (2014). *Foundations of computational linguistics, human-computer communication in natural language* (Third Edit). Springer.

Hu, S., Zhang, H., & Zhang, W. (2023). Domain knowledge graph question answering based on semantic analysis and data augmentation. *Applied Sciences (Switzerland)*, *13*(15). https://doi.org/10.3390/app13158838

Huang, H., Lin, D. K. J., Liu, M., & Yang, J. (2015). Computer experiments with both qualitative and quantitative variables. *Technometrics*, *58*(4), 495–507. https://doi.org/10.1080/00401706.2015.1094416

Ibrahim, Y. M., Essameldin, R., & Darwish, S. M. (2024). An adaptive hate speech detection approach using neutrosophic neural networks for social media forensics. *Computers, Materials & Continua/Computers, Materials & Continua (Print)*, *79*(1), 243–262. https://doi.org/10.32604/cmc.2024.047840

Ibrohim, M. O., & Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian twitter. *Proceedings of the Third Workshop on Abusive Language Online*, 46–57. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3506

Ibrohim, M. O., & Budi, I. (2023). Hate speech and abusive language detection in Indonesian social media: Progress and challenges. *Heliyon*, *9*(8), e18647. https://doi.org/10.1016/j.heliyon.2023.e18647

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, *546*(17), 126232. https://doi.org/10.1016/j.neucom.2023.126232

Khanduja, N., Kumar, N., & Chauhan, A. (2024). Telugu language hate speech detection using deep learning transformer models: Corpus generation and evaluation. *Systems and Soft Computing*, *6*(June), 200112. https://doi.org/10.1016/j.sasc.2024.200112

Kurniasih, D. (2019). Ujaran kebencian di ruang publik: Analisis pragmatik pada data pusat studi agama dan perdamaian (PSAP) Solo Raya. *Jurnal Studi Agama Dan Masyarakat*, *15*(1), 49–57. https://doi.org/10.23971/jsam.v15i1.1153

Kurniawati, R. (2023). Buzzer sebagai alat politik ditinjau dari perspektif penegakan hukum di Indonesia. *Justicia Sains: Jurnal Ilmu Hukum*, *8*(2), 260–275. https://doi.org/10.24967/jcs.v8i2.2313

Leavy, P. (2017). *Research design: Quantitative, qualitative, mixed methods, art-based, and community-based participatory research approaches*. New York, London: The Guildford Press.

Mu, Y., Yang, J., Li, T., Li, S., & Liang, W. (2024). HA-GCEN: Hyperedge-abundant graph convolutional enhanced network for hate speech detection. *Knowledge-Based Systems*, *300* (November 2023), 112166. https://doi.org/10.1016/j.knosys.2024.112166

Neyasyah, M. S. (2020). *Legal resilience in the phenomenon of social media political buzzer in Indonesia*. *130*(Iclave 2019), 338–344. https://doi.org/10.2991/aebmr.k.200321.044

Niemi, P. M., Benjamin, S., Kuusisto, A., & Gearon, L. (2018). How and why education counters ideological extremism in Finland. *Religions*, *9*(12), 1–16. https://doi.org/10.3390/REL9120420

Oktavianus, O. (2022). Hate speech and local cultural values in Indonesia. *Proceedings of the International Congress of Indonesian Linguistics Society (KIMLI 2021)*, *622*(Kimli), 151–155. https://doi.org/10.2991/assehr.k.211226.031

Pan, R., García-Díaz, J. A., & Valencia-García, R. (2024). Comparing fine-tuning, zero and few-shot strategies with large language models in hate speech detection in English. *CMES - Computer Modeling in Engineering and Sciences*, *140*(3), 2849–2868. https://doi.org/10.32604/cmes.2024.049631

Perifanos, K., Goutsos, D., Montes-Y-Gómez, M., & Rosso, P. (2021). *Multimodal technologies and interaction multimodal hate speech detection in Greek social media*. Retrieved from https://doi.org/10.3390/mti5070034

Raza, S., & Chatrath, V. (2024). HarmonyNet: Navigating hate speech detection. *Natural Language Processing Journal*, *8*(August), 100098. https://doi.org/10.1016/j.nlp.2024.100098

Saeed, J. I. (2016). *Semantics* (4th ed.). West Sussex: Blackwell.

Safari, P., & Shamsfard, M. (2024). Data augmentation and preparation process of PerInfEx: A Persian chatbot with the ability of information extraction. *IEEE Access*, *12*, 19158–19180. https://doi.org/10.1109/ACCESS.2024.3360863

Sazali, H., Rahim, U. A., Farady Marta, R., & Gatcho, A. R. (2022). Mapping hate speech about religion and state on social media in Indonesia. *Communicatus: Jurnal Ilmu Komunikasi*, *6*(July), 189–208. https://doi.org/10.15575/cjik.v6i2.

Septiawan, Y., & Chairani. (2023). Perbandingan akurasi metode deteksi ujaran kebencian dalam postingan Twitter menggunakan metode SVM dan Decision Trees yang dioptimalkan dengan Adaboost. *Jurnal Teknika*, *17*(2), 297–299.

Shim, H., Lowet, D., Luca, S., & Vanrumste, B. (2021). LETS: A label-efficient training scheme for aspect-based sentiment analysis by using a pre-trained language model. *IEEE Access*, *9*, 115563–115578. https://doi.org/10.1109/ACCESS.2021.3101867

Sirulhaq, A., Yuwono, U., & Muta'ali, A. (2023). Lack of critical approach in the hate speech research as ideological practice in Indonesia. *SHS Web of Conferences*, *173*, 04004. https://doi.org/10.1051/shsconf/202317304004

Taradhita, D. A. N., & Putra, I. K. G. D. (2021). Hate speech classification in Indonesian language tweets by using convolutional neural network. *Journal of ICT Research and Applications*, *14*(3), 225–239. https://doi.org/10.5614/itbj.ict.res.appl.2021.14.3.2

Tareq, M., Islam, M. F., Deb, S., Rahman, S., & Mahmud, A. A. (2023). Data-augmentation for Bangla-English code-mixed sentiment analysis: Enhancing cross linguistic contextual understanding. *IEEE Access*, *11*, 51657–51671. https://doi.org/10.1109/access.2023.3277787

Undang-undang (UU) Nomor 1 Tahun 2024 tentang Perubahan Kedua atas Undang-Undang Nomor 11 Tahun 2008 tentang Informasi dan Transaksi Elektronik. , Pub. L. No. 1 (2024).

Undang-undang (UU) Nomor 19 Tahun 2016 tentang Perubahan atas Undang-Undang Nomor 11 Tahun 2008 Tentang Informasi Dan Transaksi Elektronik. , Pub. L. No. 19 (2016).

United Nation. (2020). United Nations strategy and plan of action on hate speech: Detailed guidance on implementation for United Nations field presences. In *United Nations Report*. Retrieved from https://www.un.org/en/genocideprevention/documents/UN Strategy and PoA on Hate Speech_ Guidance on Addressing in field.pdf

Wasilewski, K. (2019). Hate speech and identity politics: An intercultural communication perspective. *Przegląd Europejski*, *3*(3), 175–187. https://doi.org/10.5604/01.3001.0013.5848

Yang, S., Kong, D., & He, J. (2025). Social identity or social capital : Local CEOs and corporate. *International Review of Economics and Finance*, *98*(August 2024), 103926. https://doi.org/10.1016/j.iref.2025.103926

Zahid, I. (2020). Semantics domain, verbs and collocation in women's beauty product advertisements. *Issues in Language Studies*, *9*(1), 28–50. https://doi.org/10.33736/ils.1797.2020

Zahid, I., Bakar, N. A., Kamaruddin, W. Z., Ali, W., & Jusoff, K. (2022). Pemetaan domain semantik akidah: Penyelesaian kekaburan makna. *Gjat*, *12*(2 1 83), 1–20. Retrieved from www.gjat.my

Zakariya, I., & Syafrullah, M. (2024). Implementasi text mining untuk deteksi ujaran kebencian terhadap Ibu Kota Nusantara menggunakan algorima K-Nearest Neighbors pada platform X. *5th Seminar Nasional Mahasiswa Fakulltas Teknologi Informasi (SENAFTI), 3*(3), 263–270.