インドネシアにおける2025年軍事法案を巡る世論に関するX（旧Twitter）上の感情分析

クリスティアン　ハルジュノ (情報工学分野)
指導教員　本間宏利

1　はじめに

インドネシア国軍法案 (別名 Rancangan Undang-Undang Tentara Nasional Indonesia または RUU TNI) の改正は, 2025 年 3 月 20 日に法として成立した際, 大きな論争を巻き起こした. 主な論点は特に第 47 条にあり, この条文は (武器を携帯する権利を持つ) 現役軍人が非防衛分野の文民職に任命されることを規定しており, これは 1998 年以降の民主化改革を覆すものである. これは, 軍の二重機能 (Dwifungsi ABRI としても知られる)[8] の復活に関する重大な懸念を引き起こす. 本研究はインドネシアのソーシャルメディアのパラダイム, より具体的には, インドネシア全国の市民にとって公平な議論の場として機能した X (旧 Twitter) に焦点を当てる. #TolakRUUTNI などのハッシュタグが, インドネシア政府への抗議の一形態として大規模に使用された [3].

本研究は, X 上での公開された言説の感情分析を行うことにより, RUU TNI 2025 に対する国民感情を定量化し分析することを目的とする. 具体的な目的は, 法案が最初に提案された時から正式に署名された後の数ヶ月間にわたる感情の極性の分布を測定し, 国民の不満の主な要因を特定することである. この分析は, X の検索エンジンから取得した 20 万件以上のツイートからなるコーパスを使用する.

2　先行研究

The sentiment analysis of social media data, particularly regarding this specific legislative issue in Indoensia has been extensively studied with degrees of success and different methodical approaches. One study by Ilham et al (2025) were performed against 400 tweets obtained with orange data mining operation which resulted in mostly negative sentiment of over 41.5%[7]. Another study by Adwin et al (2025) performed sentiment analysis on a time series data between 1st and the 31st of March 2025 scraped from X using the web scraping technique. The sentiment analysis itself was performed using Support Vector Machine (SVM) and evaluated using the 5-Fold Cross Validation method which resulted in an average accuracy of 78.99% and F1-Score of 83%[12] with 395 samples. Throughout previous researches found online, a common problem that was experienced by researches is the lack of quality and quantity which severely affected the accuracy of the used models or algorithms. This “quality” challenge is particularly pronounced in Indonesian social media data. The text retrieved is often extremely dirty which is characterized by informal slang, non-standard abbreviations, and most significantly, pervasive code-switching between Indoensian, English, and a local dialect (Javanese, Sundanese, Bataknese, etc) such as seen in table1. These complex phenomena presents a significant hurdle for standard NLP models.

表 1: コードスイッチングの分析（縦積み版）

セクション 1	
言語	ジャワ語
原文	Rapopo wes tak anggep durung rejekiku
日本語訳	まあいい、まだ運じゃないと思っている
セクション 2	
言語	インドネシア語
原文	duit 700 juta
日本語訳	7 億ルピアのお金
セクション 3	
言語	ジャワ語
原文	nang ngarep moto ilang sak kolo
日本語訳	目の前で一瞬で消えた
セクション 4	
言語	ジャワ語
原文	mergo aku digugah kon tangi
日本語訳	なぜなら突然起こされたから

In this study, we will be implementing data mining technique in a limited environment such as X by utilizing TF-IDF to take full advantage of X’s search engine as well as data sampling methods that enables us to train a model that generalizes better over noisy and low quality data.

3　研究方法

The general overview our established method can be seen in figure 1.

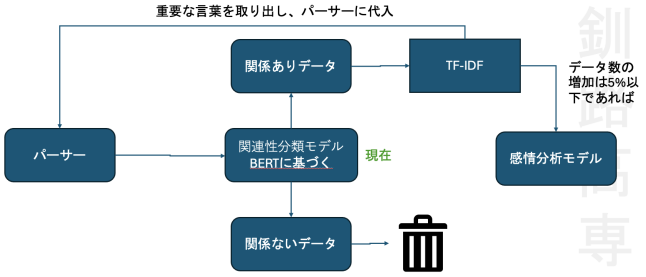

図 1: 研究方法の概念

3.1　データ収集

Data mining through X has been known to become much harder ever since the acquisition in 2022. Along with this acquisition, the free and open access to X's API (Application Programming Interface) became restricted since February 2023 [14]. X used to be a popular source of data amongst researchers due to its open access, ease of use, cheap resources, and most important of all: user generated data.

With the restriction of access to the official API, parsing data from the official database became much more expensive. In order to retrieve data, we will create a web crawler/parser that opens the search page with built search parameter and automatically scroll down the *infinite timeline*[*1] while parsing the page source and uploading it to a self-hosted private database. This parser or crawler will be written in JavaScript (Node.Js) [5] which takes advantage of the Puppeteer[*2] package [2].

3.2　X の検索エンジン

Similar to Google, X's search function implements some search notations that can be used to narrow down and retrieve specific data. Notations such as `"正確なフレーズ"` or `単語１ OR 単語２` can help to retrieve exact posts that contain or not contain those specific words [13].

```
# 特定のユーザーから「気候変動」に関する画像付きの投稿を検索
from:@username ‘‘気候変動’’ filter:media
# 「AI」または「人工知能」を含む投稿を検索
‘‘AI’’ OR ‘‘人工知能’’
# 2025 年 1 月 1 日以降の「地震」に関する投稿を検索
‘‘地震’’ since:2025--01--01
```

By combining multiple search operators, we can retrieve and construct specific datasets which targets specific topics, timeframes, which is essential for reproducible studies in social media analysis.

Despite a well developed search engine, due to the restricted access in the X API itself, utilizing these search operators reveals another significant challenge. As we are parsing through the web-based timeline itself, the recommendation algorithms which allows X to tailor the timeline specific to our habits, searching for a specific topic does not guarantee that the resulting data is clear from unrelated data. This is a massive challenge that may result in spending a lot of time just seperating between tweets (posts) that are related or unrelated to our topic.

For that reason, we will be developing a pipeline to fully saturate our dataset with as much strong-signaled and relevant data as possible.

3.2.1　初回データ収集

As mentioned before in section3.2, X's web-based search engine may often include irrelevant data in our results. To kick off the initial data gathering process, we first contructed a dictionary of words and hashtags that have some relevance to the topic of RUU TNI. This dictionary of words are gathered through some research and self consideration. For example we added the term “demonstration” due to the big demonstration that happened in relation with this topic. We also added popular hashtags such as #TolakRUUTNI (#軍事法案を反対) which is known to be the main tag used in discussions[15]. We then constructed a dictionary of relevant keywords which can be seen in table 2

[*1] a feed that loads content continuously as you scroll, without requiring the user to click to the next page
[*2] A browser automation module

表 2: インドネシア軍法案関連のハッシュタグと言葉の日本語訳

ハッシュタグ・言葉	日本語翻訳
#TolakRUUTNI	#RUUTNI に反対
#RUUTNI	#RUUTNI（国軍法案）
RUU TNI	TNI 法案（国軍法案）
demo mahasiswa	学生デモ
unjuk rasa	抗議デモ／デモ活動
demonstrasi	デモ（demonstration）
#DukungRUUTNI	#RUUTNI を支持
Dukung RUU TNI	TNI 法案を支持する
#RUUTNIPerkuatNKRI	#RUUTNI は NKRI（統一国家）を強化する
#GagalkanRUUTNI	#RUUTNI を阻止しろ
#CabutRUUTNI	#RUUTNI を撤回しろ
#PeringatanDarurat	#緊急警告
#IndonesiaGelap	#暗黒のインドネシア
#TolakRevisiUUTNI	#TNI 法改正に反対
#TolakDwifungsiABRI	#ABRI の二重機能に反対
dwifungsi	二重機能（軍と政治の両立機能）

However, certain keywords such as “demonstration” is very subtle because it can also refer to any other demonstrations that have nothing to do with RUU TNI. Using these X keywords, we then contruct a twitter search string and run an automated browser using Puppeteer to read every single post until the end of timeline (until no more posts are loaded).

As previously explained, this automated browser is programmed in JavaScript through the NodeJS engine which enables us to utilize Puppeteer, a browser automation module to simplify the data extracting process.

Another significant hurdle we expect to see is the issue of rate limiting. Along with the retricted API access, extremely strict rate-limiting prevents a smooth data-mining operation. For that reason, the data miner is tuned to not parse more data than what is limited at a given time (which at the time of writing is ~1000 posts per certain amount of time)

3.3 データ全処理

The parsed posts will be run through several stages of cleaning to normalize noise and code switching.

Unicode Normalization Normalize Unicode characters, ensuring consistency in representation (e.g., combining characters with their base forms).

Width Conversion Convert full-width (zenkaku) digits and ASCII characters to their half-width (hankaku) equivalents.

Newline Removal Replaces literal newline and carriage return strings with a single space.

URL normalization Replaces all HTTP/HTTPS URLs with a special `<url>` token.

Mention Removal Removes Twitter-style mentions.

Retweet Prefix Removal Strips the “RT” (retweet) indicator from the beginning of the text, case-insensitively.

Number-Text Separation Inserts a space between a 4-digit number and an immediately following letter, helping to separate elements like years from text.

Quote Truncation Removes all text from the first occurrence of the word “kutipan” (Indonesian for “quote”), case-insensitively, to the end of the string.

Unavailable URL Removal Removes the specific placeholder string “`<url> ini tidak tersedia`”.

Demojization Converts emoji characters into their textual description.

Character Repetition Reduction Reduces any sequence of three or more identical characters down to just two (e.g., “hellooo” becomes “hello”).

Whitespace & Case Normalization Collapses all sequences of whitespace into a single space, removes any leading or trailing whitespace, and converts the entire string to lowercase.

Final URL Token Removal Removes the `<url>` tokens that were added in the URL replacement step, effectively deleting all URLs from the text.

Other than regular expression cleaning, we will also normalize short form words and code switching through the use of available online dictionaries. Shortened words and expressions such as “BTW” is normalized to “By the way”, “Mager” to “Malas gerak” (Lazy to move), etc. This process of normalizing shortened words is called `lexical normalization task` which utilizes open-source dictionaries (lexicons)[6]. The dictionary that we will be using in this experiment is `nasalsabila/kamus-alay` avalable through GitHub*3[1].

*3 GitHub Link: `https://github.com/nasalsabila/kamus-alay`

3.4 関係性分離モデル

As previously discussed in section3.2, the use of subtle words such as “demonstration” risks including irrelevant posts in the dataset. After hydrating the database with the initial round of data mining, we will be training a model that is able to distuinguish between a post that relates to the RUU TNI paradigm and a post that does not.

A big challenge in seperating between relevant and irrelevant data is the noisiness and dirtiness of Indonesian twitter data. Within our previously created corpora of seemingly related keywords, we have included hashtags which is known to be used for discussion of the RUU TNI topic as seed in table 2. Upon retrieval of text data, we encounter posts that included the hashtag but contain discussions of something completely irrelevant to the main topic. In this case, popular hashtags are being used as an engagement tool to raise the amount of views as can be seen in figure 2.

図 2: タッグ含み無関係データ (通訳版)

Due to amount of unrelated samples within the collected data, we will be training a model based on `indolem/indobertweet` which is a bert-based model trained on a massive corpora of Indonesian twitter data [9]. However, the noisiness of the underlying dataset itself means we will need to prepare a large amount of data to be sampled, cleaned, labeled, and balanced in order to achieve a model that generalizes well over unseen data [4].

3.5 ラベル付け作業の削減

In order to reduce the amount of labeling needs, we will be implementing a density based clustering and subsampling method to remove sample noise that may affect the general performance of the fine-tuned model. The usage of clustering to produce higher quality samples has been known in the machine learning paradigm for a long time. Most notable method utilized UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) [11] and KMeans [10].

However, KMeans works best with clusters that have globular shapes.

3.6 特徴単語を抽出

3.7 第 2 回データ収集

3.8 分析手法

4 期待される結果と貢献

5 行った事

6 結論と今後の課題

参考文献

[1] Nikmatun Aliyah Salsabila, Yosef Ardhito Winatmoko, Ali Akbar Septiandri, and Ade Jamal. Colloquial indonesian lexicon. In *2018 International Conference on Asian Language Processing (IALP)*, pages 226–229, 2018.

[2] Google Chrome. Puppeteer, 2025. Accessed: 2025-10-27.

[3] CNN Indonesia. Koalisi masyarakat sipil serukan #tolakruutni yang ancam demokrasi (civil society coalition calls for #tolakruutni that threatens democracy). CNN Indonesia, 2024.

[4] Mazumder et al. Dataperf: Benchmarks for data-centric ai development. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Track on Datasets and Benchmarks*, 2023.

[5] OpenJS Foundation. Node.js, 2025. Accessed: 2025-10-27.

[6] Bo Han and Timothy Baldwin. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[7] Ilham Hibatullah, Rizky Ichsan Nur Rahman, Hikmal Maulana, David Utomo, Sumanto, and Andi Diah Kuswanto. Sentiment analysis of twitter data related to the ratification of the tni bill using orange data mining. *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, 4(3):2087–2094, Jun. 2025.

[8] Human Rights Watch. Indonesia: Proposed military law threatens rights. Human Rights Watch, 2024.

[9] Fajri Koto, Jey Han Lau, and Timothy Baldwin. Indobertweet: A pretrained language model for indonesian twitter with effective domain-specific vocabulary initialization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, 2021.

[10] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, number 14, pages 281–297, 1967.

[11] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[12] Adwin Nurhasananda and Mutaqin Akbar. Analisis sentimen masyarakat terhadap kebijakan undang-undang tentara nasional indonesia (uu tni) menggunakan support vector machine. *Jurnal Komputer, Informasi dan Teknologi*, 5(1):14, Jun. 2025.

[13] Twitter Developer Platform. Search operators, 2025. Accessed: 2025-10-27.

[14] Twitter Developers. Twitter to end free access to its api in elon musk's latest monetization push, February 2 2023. Accessed: 2025-10-27.

[15] Zuraida. Comparing the effectiveness of hashtags in digital social movements: A case study of #percumalaporpolisi and #polrisesuaiprosedur in indonesia. *CHANNEL: Jurnal Komunikasi*, 11(1):21–32, 2023. Available under CC BY-SA 4.0 license.