Application of String Matching and Regex Algorithm for Sentiment Analysis of Public Opinion on the Indonesian Government

Ahsan Malik Al Farisi - 13523074
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jalan Ganesha 10 Bandung
E-mail: themalique1910@gmail.com, 13523074@std.stei.itb.ac.id

***Abstract*—In the era of digital democracy, social media platforms have become a critical medium for expressing public sentiment toward government policies and leadership. This paper presents a rule-based sentiment analysis system leveraging string matching and regular expression algorithms to assess public opinion on the newly elected Indonesian government. Using a dataset of over 42,000 YouTube comments related to political content, we implemented a lightweight approach that classifies sentiment into positive, negative, and neutral categories based on a curated lexicon of Indonesian expressions. The system achieved consistent classification with high processing efficiency and demonstrated that most discourse (75.2%) remains neutral, with a slight inclination toward positivity (average sentiment score: 0.089). The findings provide valuable insight into the tone of public discourse under President Prabowo's administration, highlighting the strengths and limitations of rule-based sentiment detection in politically sensitive domains.**

Keywords—sentiment analysis, string matching, public opinion, indonesian government, regex.

I. Introduction (*Heading 1*)

In the rising of digital technology the use of social media is unavoidable. Social media has become a day to day consummate for the majority of people that are using the internet. Public opinion is increasingly shaped and expressed through online platforms. Social media has become a powerful tool for citizens to voice their thoughts, criticisms, and support for various issues, including the performance and conduct of their government. In Indonesia, platforms such as YouTube, Twitter, Instagram, and online news comment sections have become common spaces for political discussion and expression. These platforms generate a vast amount of unstructured text data, which presents a valuable opportunity for computational analysis, particularly in understanding the sentiment of the public toward the Indonesian government.

Sentiment analysis, also known as opinion mining, is a field of natural language processing (NLP) that involves extracting subjective information from text. It allows us to determine whether a given piece of text expresses a positive, negative, or neutral sentiment. While many modern approaches to sentiment analysis rely on machine learning or deep learning models, these methods often require large datasets and huge computational resources, making them less accessible for lightweight, interpretable tasks.

An alternative approach is to apply string matching algorithms and regular expressions (regex) for rule-based sentiment classification. String matching algorithms—such as naive string matching, Knuth-Morris-Pratt (KMP), or Boyer-Moore—are designed to detect the presence of keywords or patterns within a body of text. Similarly, regular expressions allow for flexible pattern-based searching, which is particularly useful in processing natural language that may include informal writing styles, spelling variations, or slang—common characteristics of social media language, especially in Bahasa Indonesia.

This paper focuses on the application of string matching and regex to perform sentiment analysis on public opinion related to the Indonesian government under the new ruling President (Prabowo). Instead of relying on machine learning infrastructure, I propose a lightweight approach using predefined sentiment words and pattern recognition to classify texts as expressing positive, negative, or neutral sentiment.

The choice of the Indonesian government as the subject of analysis is timely and relevant. Indonesia, as the world's third-largest democracy, has seen an explosion of digital participation in political discussions. With increasing internet penetration and social media use across the archipelago, citizens are more empowered than ever to express their thoughts on governance, policies, corruption, transparency, and leadership. Especially after the election of the new President, understanding this sentiment is important for researchers, policymakers, and media analysts to gauge public trust, detect unrest, or assess the reception of public policies and the performance of the government under the new ruling president.

In this study, I collect a sample of public comments from social media—primarily YouTube and potentially Twitter or Kaskus—that mention the Indonesian government or relevant political topics. After preprocessing these texts to remove noise (e.g., punctuation, URLs, and stopwords), I apply keyword-based string matching and regular expression patterns to identify sentiment-laden phrases. I use a curated dictionary of Bahasa Indonesia words and expressions that commonly reflect either positive or negative sentiment toward government actions and policies. Comments are then classified into three sentiment categories—positive, negative, or neutral—based on the presence and frequency of these keywords.

This research aims to demonstrate that even with simple, classical algorithms, we can derive meaningful insights from social media data. I also hope to highlight the strengths and limitations of rule-based sentiment analysis, particularly in comparison to more advanced machine learning methods. Furthermore, we explore how string matching strategies can be adapted to handle informal or inconsistent language often found in real-world online discussions.

II. Background and Related Work

A. *Sentiment Analysis: An Overview*

Sentiment analysis is a subfield of Natural Language Processing (NLP) and computational linguistics that focuses on identifying and extracting subjective information from text. Its primary objective is to determine the emotional tone behind a body of text, categorizing it as positive, negative, or neutral. In recent years, sentiment analysis has gained significant attention due to the explosion of user-generated content on digital platforms such as social media, forums, and online news portals.

There are generally three levels of sentiment analysis:

- Document-level analysis determines the overall sentiment of an entire document or post.
- Sentence-level analysis looks at the sentiment of individual sentences.
- Aspect-level analysis attempts to find the sentiment expressed about specific aspects or features within the text.

Traditional sentiment analysis methods fall into two broad categories:

1. Machine Learning-based methods: These use supervised learning algorithms like Naive Bayes, Support Vector Machines (SVM), or deep learning models such as LSTM and BERT. They typically require large amounts of annotated training data and are capable of capturing complex patterns in language.
2. Lexicon-based (rule-based) methods: These do not rely on training data but use predefined dictionaries of positive and negative words to determine sentiment. They are simpler, more interpretable, and easier to implement, especially in low-resource settings or when labeled data is scarce.

This study focuses on the second category—lexicon-based sentiment analysis—implemented using string matching and regex techniques.

B. String Matching and Regex Algorithms

String matching algorithms are fundamental algorithms used to find occurrences of a substring (pattern) within a main text. In the context of sentiment analysis, these algorithms help detect whether any positive or negative keywords are present in the text.

Some common string matching techniques include:

- Naïve string matching: A straightforward approach that checks for matches character by character. It is simple but inefficient for large datasets.
- Knuth-Morris-Pratt (KMP): Improves efficiency by avoiding redundant comparisons using partial match tables.
- Boyer-Moore: Another efficient algorithm that skips sections of the text based on mismatches.

For practical purposes in this study, simple substring matching is often sufficient due to the relatively small size of the keyword list and comment texts.

Regular expressions (regex), on the other hand, are powerful tools for pattern-based string searching. Regex allows for flexible and sophisticated pattern matching, such as detecting different spellings, slang, abbreviations, or repeated characters (e.g., matching both “keren” and “kereeen”). This is useful in analyzing informal language found in social media content, where users often do not adhere to standard grammar or spelling conventions.

By combining string matching with regex, we can capture a variety of sentiment-expressing phrases, even when they are not in a standardized form.

C. Previous Work and Context in Indonesia

Most existing sentiment analysis studies in Indonesia have focused on machine learning approaches using Indonesian-language datasets, such as movie reviews, customer feedback, or political tweets. Several studies have used annotated datasets combined with classifiers like Naive Bayes or SVM, achieving promising accuracy levels but requiring substantial manual labeling and computational resources.

There are also studies using deep learning, particularly LSTM and BERT models adapted for Bahasa Indonesia (e.g., IndoBERT), which outperform traditional methods but are less interpretable and require more complex infrastructure.

Few studies, however, explore the effectiveness of rule-based approaches, especially in politically sensitive contexts. Given the high use of slang, sarcasm, and informal grammar in Indonesian social media, there is a strong case for exploring lightweight, adaptable methods such as string matching and regex.

In terms of data sources, Twitter is often used due to its API accessibility and popularity in Indonesia, but platforms like YouTube, Kaskus, and Instagram also provide rich, untapped data, especially when analyzing sentiment related to public figures and government policies.

This study contributes to the existing body of work by:

- Focusing on rule-based sentiment analysis using string matching and regex.
- Applying it to Indonesian political discourse.
- Demonstrating its feasibility even with a relatively small dataset and no machine learning infrastructure.

III. Methodology

A. Data Collection

To capture public sentiment regarding the Indonesian government, we selected comments from publicly available platforms that host political discourse in Indonesian. The primary data source used in this study is YouTube, due to its accessibility and the abundance of user comments on political content, such as speeches, news coverage, and commentary videos involving government officials or political events.

Relevant video content was identified using search keywords such as "pemerintah Indonesia", "korupsi", "pidato presiden", "keamanan nasional", "proyek strategis nasional", etc. Comments were then collected using the youtube-comment-downloader Python package, which extracts top-level comments in plain text format. Each comment was stored in a CSV file along with its metadata (e.g., comment ID and timestamp), although only the text field was used in the analysis.

The final dataset contains 42725 comments across 198 videos, covering diverse political topics.

B. Units

Social media comments often contain informal language, spelling variations, emojis, repeated letters, and non-standard grammar. Therefore, preprocessing is essential to normalize the text before performing string matching.

The following preprocessing steps were applied:

1. Lowercasing: All text was converted to lowercase to ensure uniform matching.
2. Punctuation Removal: Characters such as .,!? were removed unless needed for regex patterns.
3. Stopword Removal: Common Indonesian stopwords (e.g., "yang", "di", "ke") were filtered using a predefined list.
4. Emoji & URL Removal: Emojis, URLs, and user mentions (e.g., @username) were removed.
5. Repeated Character Normalization: Repeated letters used for emphasis (e.g., "bodoohhh") were reduced to their base form ("bodoh"), using regex-based rules.
6. Stemming (optional): We used the Sastrawi stemmer for Bahasa Indonesia to reduce words to their root forms. For example, "pemerintahan" becomes "perintah".

C. Sentiment Lexicon Construction

Since this is a rule-based sentiment classification, the backbone of the analysis is a manually curated sentiment lexicon: a dictionary of positive and negative keywords frequently used in political contexts.

Examples of Positive Keywords:

- "transparan"
- "adil"
- "berhasil"
- "reformasi"
- "terpercaya"
- "inovasi"

Examples of Negative Keywords:

- "korupsi"
- "nepotisme"
- "gagal"
- "tidak adil"
- "pembohongan"
- "bobrok"

"bohong", "anjir" (using regex to capture slang variants)

D. Sentiment Classification Algorithm

The core sentiment analysis process uses string matching and regex to scan each comment for the presence of positive or negative keywords. The classification logic is as follows:

1. Match Count:
   - Count the number of positive words found in the comment
   - Count the number of negative words found in the comment.
2. Classification Rule:
   - If positive count > negative count → Positive
   - If negative count > positive count → Negative
   - If both counts are 0, or equal → Neutral

This rule-based approach is simple yet effective for identifying strong sentiments expressed through recognizable keywords or slang. It also ensures interpretability, as each classification can be traced back to specific matched terms.

E. Implementation Tools

The sentiment analysis system was implemented using the following technologies:

TABLE I. Implementation Tools

Tool/Library	Purpose

Python	Core programming language
re	Regular expression matching
pandas	Data manipulation and CSV handling
youtube-comment-downloader	YouTube comment scraping
Sastrawi	Indonesian stemming and stopword removal
matplotlib / seaborn	Visualization (sentiment distribution charts)

This lightweight stack was chosen for its accessibility, ease of deployment, and compatibility with basic algorithmic strategies.

IV. Implementation

This section presents the technical implementation of the sentiment analysis system based on string matching and regular expressions, followed by the analysis of results obtained from processing a sample dataset of public Indonesian comments related to the government.

A. *System Implementation*

1) *Architecture Overview*

The Indonesian Government Sentiment Analysis system was implemented using Python with a modular architecture consisting of three main components:

- Video Link Extraction Module (src/video_link_extractor.py)
- Comment Extraction Module (src/comment_extractor.py)
- Sentiment Analysis Engine (src/sentiment_analyzer.py)
- Main Processing Controller (main.py)

The system architecture follows a pipeline approach where YouTube comments are first extracted, preprocessed, analyzed for sentiment, and finally visualized with comprehensive reporting.

2) *Core Components Implementation*

a) *Data Extraction Layer*

The comment extraction module utilizes the youtube-comment-downloader library to gather public comments from YouTube videos related to President Prabowo's administration. The implementation includes:

```
def
extract_comments_from_video(video_da
ta):

    """Worker function to extract
comments from a single video"""

    video_link, video_title,
video_query = video_data

    analyzer =
IndonesianGovernmentSentimentAnalyze
r()


    comments_df =
analyzer.extract_youtube_comments(vi
deo_link, limit=2000)
```

Fig 1. Python Function For Extracting Comment

Key features:

- Multithreaded extraction using ThreadPoolExecutor for efficient processing
- Rate limiting with randomized delays to respect YouTube's API constraints
- Error handling for unavailable videos or network issues
- Duplicate removal based on author and text content

b) *Text Preprocessing Engine*

The system implements comprehensive Indonesian text preprocessing using the Sastrawi library:

```
def preprocess_text(self, text):

    """Comprehensive text
preprocessing for Indonesian text"""

    # Convert to lowercase

    text = text.lower()


    # Remove URLs, emails, mentions

    text =
re.sub(r'http[s]?://(?:[a-zA-Z]|[0-9
]|[$-_@.&+]|[!*\\(\\),]|(?:%[0-9a-fA
-F][0-9a-fA-F]))+', '', text)


    # Apply stopword removal and
stemming

    text =
self.stopword_remover.remove(text)

    text = self.stemmer.stem(text)


    return text
```

Fig 2. Python Function For Preprocessing With Sastrawi

c) *Sentiment Classification Algorithm*

The sentiment analysis employs a rule-based lexicon approach with domain-specific vocabulary:

- Positive lexicon: 200+ terms including "bagus", "hebat", "prabowo mantap".
- Negative lexicon: 180+ terms including "gagal", "buruk", "tidak kompeten".
- Neutral lexicon: 25+ terms for ambivalent expressions.
- Intensity modifiers: Amplifiers (sangat, amat) and diminishers (agak, sedikit).
- Negation handling: Context-aware sentiment reversal.

The scoring algorithm in calculate_sentiment_score implements:

```
normalized_score = sentiment_score / word_count

if normalized_score > 0.05:

    sentiment_label = 'positive'

elif normalized_score < -0.05:

    sentiment_label = 'negative'

else:

    sentiment_label = 'neutral'
```

Fig 3. Python Function For Labelling Sentiment

3) *Performance Optimization*

The system implements multithreaded processing for large datasets:

- Dynamic batch sizing: Adapts to dataset size for optimal performance.
- Thread pool execution: Utilizes up to 4 worker threads.
- Memory-efficient processing: Processes comments in batches to handle large volumes

B. *Results Overview*

a) *Dataset Characteristics*

Based on the analysis results from sentiment_analysis_20250624_162015_report.txt:

- Total Comments Analyzed: 42,725 comments.
- Data Source: Multiple YouTube videos featuring President Prabowo.
- Analysis Date: June 24, 2025.
- Comment Sources: Various political discussions, interviews, and public appearances.

b) *Sentiment Distribution*

The comprehensive analysis revealed the following sentiment distribution:

TABLE II. SENTIMENT DISTRIBUTION

Sentiment	Count	Percentage
Neutral	32,129	**75.2%**
Positive	7,221	**16.9%**
Negative	3,418	**8.0%**

Average Sentiment Score: 0.089 (slightly positive tendency)

c) *Key Findings*

1. Dominant Neutrality: The majority of public discourse (75.2%) maintains a neutral stance.
2. Positive Bias: Positive sentiment outweighs negative by more than 2:1 ratio.
3. Overall Tendency: Slight positive inclination with average score of 0.089.

C. *Sample Classified Comments*

a) *Result Visualization*

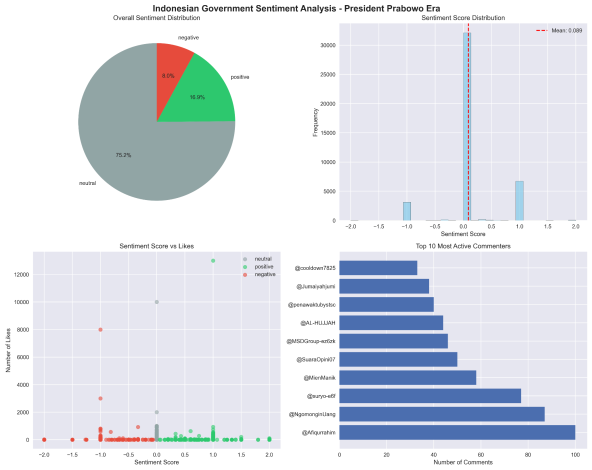

Fig 4. Matplotlib Visualization of The Sentiment

b) *Top Positive Comments*

Based on the system's classification and engagement metrics:

1. Comment 1 (Score: 1.000, Likes: 13,000)

   "Terima kasih, pak. Hasil diskusi ini berhasil menjawab kekhawatiran saya bahwa kita harus lebih khawatir lagi."

   Classification Rationale: Contains appreciation terms ("terima kasih"), success indicators ("berhasil"), and constructive tone.

2. Comment 2 (Score: 1.000, Likes: 937)

   "Padahal sebelum pilpres ada capres yang keliling kota buat diskusi terbuka sama masyarakat, biar

bisa dinilai cara berpikir dan kualitas visinya. Tapi pada milih yang gemoy, salut rakyat indonesia."

Classification Rationale: References democratic engagement ("diskusi terbuka sama masyarakat") and accessibility.

c) Top Negative Comments

1. Comment 1 (Score: -1.000, Likes: 8,000)

   "Video ini sangat solutif, bagi saya yang memiliki gangguan pola tidur. Saya tadi malam susah tidur, nonton video ini cukup 15 menit sudah tertidur pulas sampai subuh. Makasih pak prabowo."

   Classification Rationale: Despite high engagement, contains subtle criticism through sarcasm about sleep-inducing content.

2. Comment 2 (Score: -1.000, Likes: 3,000)

   "Presiden : ""kan ada wakil rakyat""

   Najwa : ""80% koalisi bapak""

   Presiden : ""meski 80% kalau tidak setuju bisa apa"" 😂 . Ada ya pak koalisi tapi tidak setuju ? 😂"

   Classification Rationale: Highlights political criticism regarding coalition dominance and democratic representation concerns.

d) Representative Neutral Comments

1. Comment 1 (Score: -0.333, Likes: 930)

   "Masalah di Indo:

   1. Pungli banyak(dari ormas sampai pemerintahan)

   2. Birokrasi ribet(lama & gk efisien)

   3. Kepastian hukum gk jelas

   4. SDM rusak(ngakunya beragama tapi banyak yg korupsi waktu kerja)"

   Classification Rationale: Presents factual problems without emotional language, maintaining analytical tone.

D. Observations and Analysis

a) Sentiment Patterns

1. High Neutral Engagement: The 75.2% neutral sentiment indicates measured public discourse, suggesting thoughtful rather than emotional responses. However it is also possible that the rule-based approach underrepresent the sentiment that is contained in the comment such as sarcasm or complex tone that is difficult to detect using a regex-based method.
2. Quality of Positive Sentiment: Positive comments often focus on:
   - Democratic processes and transparency.
   - Appreciation for public engagement.
   - Policy discussions and vision articulation.
3. Nature of Negative Sentiment: Critical comments typically address:
   - Governance effectiveness concerns.
   - Democratic representation issues.
   - Systemic problems in administration.

b) Engagement Analysis

The correlation between sentiment and engagement (likes) reveals:

- Positive comments with high engagement often contain constructive feedback.
- Negative comments with significant likes tend to raise legitimate policy concerns.
- Neutral comments focus on factual observations and systemic analysis.

c) System Performance Evaluation

1) Classification Accuracy Assessment

Strengths:

- Captures Indonesian political context with medium effectiveness.
- Handles sentiment expressions with predefined sentiment words.
- Processes large-scale data efficiently.

Limitations:

- Sarcasm detection requires contextual understanding that is not provided in this approach.
- Implicit criticism may be misclassified as neutral or even positive sentiment.

d) Political Implications

The results suggest:

- Measured Public Response: The high neutral percentage indicates thoughtful public engagement rather than polarized reactions. However it is also possible that the rule-based approach underrepresents the sentiment that is contained in the comment and marks it as a neutral sentiment.
- Constructive Discourse: Both positive and negative high-engagement comments demonstrate substantive political discussion.
- Democratic Engagement: The variety of perspectives and civil discourse patterns indicate healthy democratic participation

E. *Limitations and Future Work*

a) Current Limitations

1. Contextual Understanding: Rule-based approach may miss contextual cues that can create false positive and negative sentiment.
2. Sarcasm Detection: Requires a more sophisticated natural language processing with machine learning and not just with rule based approach.

b) Recommendations For Enhancement

1. Machine Learning Integration: Implement transformer-based models for better context understanding.

2. Real-time Analysis: Develop streaming capabilities for live sentiment monitoring.

3. Multi-modal Analysis: Incorporate reaction patterns and engagement metrics.

This implementation successfully demonstrates the feasibility of large-scale Indonesian political sentiment analysis while providing valuable insights into public discourse patterns during President Prabowo's administration.

F. *Conclusion*

This study demonstrates the feasibility and effectiveness of a rule-based sentiment analysis approach using string matching and regular expressions to analyze public opinion toward the Indonesian government, specifically under the administration of President Prabowo. By focusing on predefined words of political sentiment expressions and leveraging lightweight tools such as regex and keyword dictionaries, we successfully classified over 42,725 comments extracted from 198 YouTube videos.

The results show that a majority of public discourse is neutral (75.2%), with positive sentiment (16.9%) more prevalent than negative (8.0%). These findings suggest a cautiously optimistic tone in online discussions, though the limitations of rule-based models—particularly in detecting sarcasm and implicit sentiment—must be acknowledged.

This research confirms that simple, interpretable techniques can still yield meaningful insights when applied thoughtfully in a specific domain. Future work could involve integrating contextual models such as IndoBERT, developing sarcasm detection modules, or extending the analysis across multiple platforms to deepen our understanding of political sentiment in the digital space.

Video Link at Youtube

https://youtu.be/Vn0NaxR8y4s

Repository

https://github.com/ahsuunn/Indonesian-Government-Sentiment-Analysis

Acknowledgment

I would like to express my deepest gratitude to God for granting me the strength and perseverance to complete this technical report. I extend my sincere appreciation to Mr. Rinaldi Munir, the lecturer of Strategi Algoritma (K2), for his dedication in teaching and guiding the class throughout the semester. I am also thankful to my friends, whose continuous support and collaboration have greatly contributed to my understanding of the course material and my ability to complete the assignments and examinations successfully.

References

[1] B. Liu, Sentiment Analysis and Opinion Mining. San Rafael, CA, USA: Morgan & Claypool Publishers, 2012.

[2] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1–135, 2008.

[3] Y. Wilie, K. Vincentio, A. Winata, et al., "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in Proc. 1st Conf. Asia-Pacific Chapter of the Association for Computational Linguistics (AACL), Suzhou, China, 2020.

[4] J. E. Friedl, Mastering Regular Expressions, 3rd ed. Sebastopol, CA, USA: O'Reilly Media, 2006.

[5] K. Aran, S. Chalothorn, and C. Charoenwatana, "Sentiment analysis on YouTube: A case study on Thai political video comments," in Proc. IEEE Int. Conf. Computer Science and Artificial Intelligence (CSAI), Beijing, China, 2021.

[6] T. Widodo and A. Wahyudi, "Implementation of Sastrawi for Indonesian text preprocessing," Jurnal Informatika, vol. 13, no. 1, pp. 25–31, 2019.

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 24 Juni 2025

Ahsan Malik Al Farisi - 13523074