SENTIMENT ANALYSIS AND CLUSTERING OF ISP SERVICE USERS BASED ON SOCIAL MEDIA PLATFORM X IN INDONESIA USING K-MEANS METHOD

Marcel Kurniawan[1*], Hendra Achmadi[2]

[1,2]Faculty of Economics and Business, Universitas Pelita Harapan, Tangerang, Indonesia
E-mail: 01804220014@student.uph.edu.com

Abstract

This study analyzes user sentiment towards Internet Service Providers (ISPs) in Indonesia using the social media platform X (formerly Twitter). Data was collected using web scraping and processed with TF-IDF to convert text into numerical representations. Sentiment was determined using the NLP BERT model. The K-Means Clustering method was used to group negative tweets based on content similarity. The data consists of 6.000 negative tweets. The analysis identified three main clusters: technical issues (network disruptions, high prices, slow connections), customer service and interactions, and communication and customer satisfaction. Inter-cluster distances were: Cluster 1 and Cluster 2 (0.460), Cluster 1 and Cluster 3 (0.349), Cluster 2 and Cluster 3 (0.341). Intra-cluster variations were: Cluster 1 (0.140), Cluster 2 (0.113), Cluster 3 (0.064). Managerial implications include the need to improve technical service quality, customer service responsiveness, and billing transparency. The study's limitations include the limited amount of data and potential bias. Future research is suggested to use regression models to predict customer satisfaction based on complaint patterns and user sentiment.

Keywords: *Sentiment Analysis; K-Means Clustering; Social Media; Web Scraping; NLP*

1. INTRODUCTION

The rapid expansion of internet technology has transformed the global communication landscape, eliminated geographical and temporal boundaries and enabled various electronic activities such as e-commerce and public data services. In Indonesia, the proliferation of internet use has led to a significant increase in network traffic and challenges related to connection quality. Internet Service Providers (ISPs) and network operators compete to offer a wide range of services to meet the growing demand for reliable internet connections (Ruth, 2013).

According to a report by We Are Social in January 2023, there were 167 million active social media users in Indonesia, accounting for 60.4% of the total population. Despite a 12.57% decline from the previous year, which marked the first decrease in a decade, social media remains a critical platform for communication and public engagement (Widi, 2023). Twitter, now rebranded as X, has seen a notable increase in users, with Indonesia ranking fourth globally in terms of user numbers, reflecting the platform's growing popularity despite challenges such as valuation declines following its acquisition by Elon Musk (Annur, 2023).

Sentiment analysis on Twitter has become an essential tool for understanding public opinion, offering insights that can enhance customer service and inform business decisions, Wang and Liu (2022). This study aims to analyze the sentiments of users and potential users

of ISP services in Indonesia by examining tweets related to six major ISPs: First Media, Starlink, Telkomsel, Smartfren, Biznet, and Indihome. Utilizing Natural Language Processing (NLP) and K-Means Clustering, the research seeks to classify user opinions and identify key themes and issues.

This study replicates and modifies the research conducted by Hashfi et al. (2022), which compared Naïve Bayes and Support Vector Machine (SVM) algorithms for sentiment classification. The current research employs NLP techniques and K-Means Clustering to provide a more nuanced analysis of user sentiments expressed on social media. By leveraging these advanced methodologies, the study aims to deliver actionable insights that can help ISPs improve service quality and customer satisfaction, as well as inform policy-making and future research.

2. LITERATURE REVIEW

2.1. Web Scraping

Web scraping is a technique for automatically retrieving information from websites without manual copying. This technique focuses on extracting and collecting semi-structured data from the internet, often in the form of web pages written in markup languages like HTML or XHTML. Specific data from these pages is analyzed and extracted for various purposes. Many studies have utilized scraping tools to collect web data (Yani et al., 2019).

2.2. Node.js

Node.js is a JavaScript runtime built on Chrome's V8 engine, designed for creating network applications that can handle many connections efficiently. Its non-blocking, event-driven architecture makes Node.js ideal for high-performance web applications. The popularity of Node.js is supported by its ease of use and npm, its package manager that aids in rapid prototyping. Node.js enables end-to-end JavaScript development, enhancing productivity (Kyriakou & Tselikas, 2022).

2.3. CRISP-DM

CRISP-DM (Cross-Industry Standard Process for Data Mining) is an industry-standard methodology used for developing data mining and knowledge discovery projects. First proposed in 2000, it has become the de facto standard for data mining projects. CRISP-DM consists of six main phases in the project lifecycle: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Chapman et al., 2000).

2.4. Natural Language Processing (NLP)

NLP involves various techniques such as speech recognition, text understanding, sentiment analysis, and text generation. One common application of NLP is sentiment analysis, which is used to determine the attitude or emotion conveyed by specific texts, such as product reviews or social media posts (Chowdhury & Chowdhury, 2003).

2.5. BERT (Bidirectional Encoder Representations from Transformers)

BERT is a machine learning model developed by Google and introduced in 2018. It is a transformer-based model known for its powerful capabilities in NLP tasks. BERT's uniqueness lies in its ability to understand context bidirectionally, from left to right and right to left, enabling it to comprehend the nuances and meanings of sentences more effectively. BERT has been widely used in various NLP applications, including sentiment analysis and text classification (Devlin, 2018).

2.6. NLPTOWN/bert-base-multilingual-uncased-sentiment

NLPTOWN/bert-base-multilingual-uncased-sentiment is a BERT model specifically trained for sentiment analysis in multiple languages. It is designed to detect positive, negative, and neutral sentiments from given texts. This model is particularly useful for handling sentiment analysis in a multilingual context, making it highly effective for analyzing texts in different languages simultaneously (NLPTOWN, 2021).

2.7. K-Means Clustering

K-Means is a popular clustering method frequently used in data analysis. This technique aims to divide data into a specified number of clusters, where each cluster consists of data points that are similar to one another. The K-Means algorithm starts by randomly selecting k points as initial cluster centres (centroids) (Jain, 2010). Each data point is then assigned to the nearest centroid based on Euclidean distance. This process repeats until the positions of the centroids stabilize. K-Means is employed in this research to cluster tweets based on their thematic similarity, helping identify major themes and issues faced by ISP users in Indonesia (Lloyd, 1982; Macqueen, 1967).

2.8. CRISP-DM Implementation in the Study

Business Understanding: Understanding the business goals and requirements, and determining how data mining can provide solutions to the identified problems. The business goal here is to understand user sentiments towards ISP services in Indonesia, as expressed on social media platforms like Twitter.

a. Data Understanding: Collecting and exploring data to understand its characteristics. Data is collected from tweets containing keywords related to ISPs such as “Indihome”, “Biznet”, “Telkomsel”, etc. The data is gathered using web scraping techniques facilitated by Node.js
b. Data Preparation: Cleaning and transforming the data to make it suitable for analysis. This involves removing duplicate tweets, cleaning text from noise, removing stopwords, and tokenizing the text.
c. Modeling: Applying appropriate algorithms to the prepared data. In this project, K-Means Clustering is used to group tweets based on thematic similarity, and sentiment analysis is conducted using the BERT model.
d. Evaluation: Ensuring that the model meets the business objectives by measuring its accuracy and interpreting the clustering results. The elbow method is used to determine the optimal number of clusters.
e. Deployment: Presenting the extracted knowledge in a useful manner, which could range from generating reports to implementing a repeatable data mining process.

3. RESEARCH METHOD

This study adopts a quantitative approach, focusing on sentiment analysis and tweet classification related to ISP services in Indonesia. The Cross-Industry Standard Process for Data Mining (CRISP-DM) framework is utilized to ensure systematic and structured research, encompassing stages from data collection to data analysis.

Figure 1. The CRISP-DM Process Model for Data Mining Projects

3.1. Business Understanding

The initial phase aims to comprehend the business context and objectives of the research. The primary goals are:

a) Identifying and categorizing user sentiments towards ISP services (positive, negative, neutral).
b) Analyzing sentiment distribution across various ISP providers in Indonesia.
c) Uncovering key negative themes and topics discussed by users regarding ISP services.
d) Identifying main issues and user expectations.
e) Clustering tweets based on textual similarity using the K-Means algorithm to identify patterns and trends in ISP service usage.

3.2. Data Understanding

Data was gathered from the social media platform X (formerly known as Twitter) using web scraping techniques. The dataset comprises 6,000 tweets, with 1,000 tweets sampled from six major ISP brands: First Media, Starlink, Telkomsel, Smartfren, Biznet, and Indihome.

3.3. Data Preparation

a) Data Cleansing: Removal of punctuation, retweets, links, and special characters from the tweets.
b) Stopwords Removal: Eliminating common words that do not contribute significant meaning to the analysis.
c) Labeling Sentiments: Using the NLPTOWN/bert-base-multilingual-uncased-sentiment model to assign sentiment labels (positive, negative, neutral) to each tweet.

3.4. Modeling

The modeling phase applies the K-Means clustering algorithm to group tweets based on text similarity. This method helps in identifying clusters that represent various user sentiments and issues related to ISP services. The BERT model is utilized for sentiment analysis, providing a nuanced understanding of user opinions.

3.5. Evaluation

Evaluation involves assessing the quality and effectiveness of the clustering model. Key evaluation metrics include intra-cluster similarity (homogeneity) and inter-cluster

dissimilarity (heterogeneity). The elbow method is used to determine the optimal number of clusters, ensuring that the model accurately captures the main themes and sentiments expressed by users.

4. RESULTS AND DISCUSSION

4.1. Data Collection

The data for this study was gathered from the social media platform X (formerly known as Twitter). Data collection was performed using web scraping techniques, which allow for automatic data retrieval without manual copying. The NodeJS Twitter Scraper, an open-source library written in JavaScript, was used to scrape tweets from users based on specific hashtags and topics related to ISP services in Indonesia, such as "starlink," "indihome," "biznet," "telkomsel," "smartfren," and "firstmedia." Only original tweets without links were considered to ensure the authenticity of user opinions. A total of 6,000 tweets in Indonesian were collected, encompassing various complaints, praises, and user opinions about ISP services in Indonesia.

4.2. Data Processing

Once the data was collected, the next step was data processing. The initial step in data processing involved cleaning the tweets by removing punctuation, retweets, links, and special characters. The following table illustrates an example of text cleaning:

Table 1. Example of Text Cleaning

Before	After
@ponhpohas @ddockingstation @Telkomsel *Sama* please *kuota gua sekarat mana kaga bisa beli kuota jaringan sinyal lemah* https://t.co/0NUbkfOF5H	*sama please kuota gua sekarat mana kaga bisa beli kuota jaringan sinyal lemah*

The second step was removing common words using stopwords to eliminate insignificant words such as "dan," "yang," "di," etc.

Table 2. Example of Text Stopwords

Before	After
haram sumpah anjing biznet service *hoyo aje kaga konek mau diapain sih bang aplikasi aplikasi populer ini bang* targeted *bener kayaknya* down *nya monyet*	*haram sumpah anjing biznet* service *hoyo aje kaga konek diapain sih bang aplikasi aplikasi populer bang* targeted *bener kayaknya* down *nya monyet*

The third step was sentiment labelling for each tweet using the BERT model, which classified the sentiment into "positive," "negative," and "neutral." The results were as follows:

Tabel 3. Sentiment Results

Sentiment Brand	Indihome	Starlink	First Media	Biznet	Smartfren	Telkomsel
Positive	103	289	231	219	388	193
Neutral	207	159	123	159	172	258
Negative	690	552	646	622	440	549

The sentiment results were visualized, and only tweets labelled as “negative” (3,499 tweets) were selected for further analysis. This selection was made to focus on understanding the primary issues and complaints expressed by users, as negative feedback provides critical insights into areas that require improvement for ISPs.

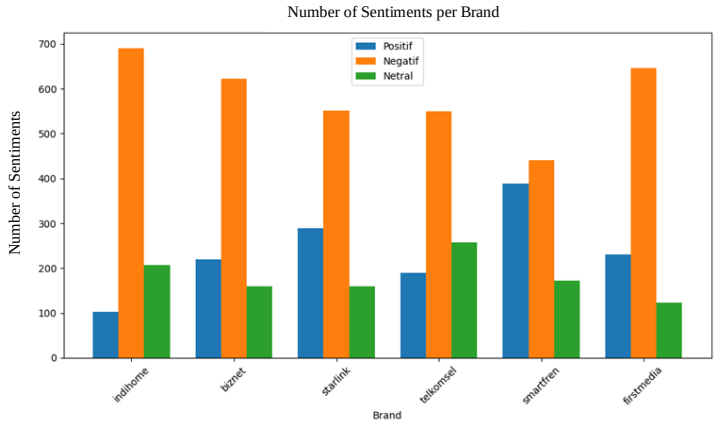

Figure 2. Visualization of Sentiment Results

4.2. Clustering Data

Clustering was performed on the negative tweets to group them based on content similarity. The K-Means Clustering algorithm was used, with TF-IDF (Term Frequency-Inverse Document Frequency) employed to convert text into numerical representations. The Elbow Method was used to determine the optimal number of clusters, which was identified as three based on the SSE (Sum of Squared Errors) plot.

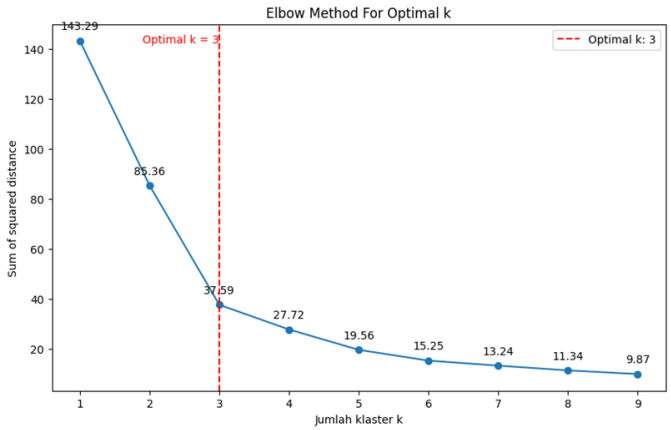

Figure 3. Elbow Method for Determining Optimal k

4.3. Evaluation Of Clustering

The evaluation phase of the clustering model included assessing the quality and effectiveness of the model by measuring the inter-cluster distances and intra-cluster variations.

4.4. Inter-Cluster Distances

Inter-cluster distances measure how far apart the centroids of different clusters are from each other. Larger distances indicate that the clusters are well-separated and distinct. The distances between the centroids of the clusters were calculated and visualized using a heatmap.

a) Cluster 1 to Cluster 2: 0.460
b) Cluster 1 to Cluster 3: 0.349
c) Cluster 2 to Cluster 3: 0.341

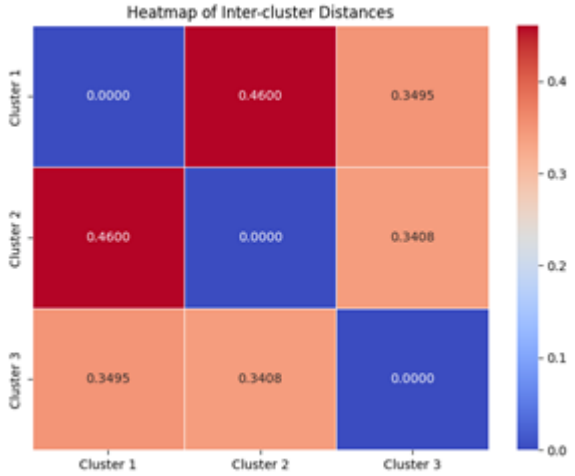

Figure 4. Heatmap of Inter-Cluster Distances

The larger distances between the clusters suggest that the clustering model effectively segregated tweets into distinct groups based on content similarity.

4.4. Discussion

The results of the clustering and sentiment analysis revealed several key insights:

a. Cluster 1 - Technical Issues and High Costs:

Common words such as “biznet,” “starlink,” “indihome,” “smartfren,” “telkomsel,” “internet,” “gangguan,” “mahal,” “lemot,” “sinyal,” “kuota.” This cluster highlighted user complaints about technical issues such as network disruptions, slow internet speeds, and high costs. Users frequently expressed frustration with connectivity problems and the perceived high prices of ISP services.

b. Cluster 2 - Service and Customer Support:

Common words such as “media,” “people,” “bantu,” “mohon,” “layanan,” “tagihan,” “wifi,” “maaf”. This cluster focused on requests for assistance and complaints related to customer service and billing issues. Users often sought help for service disruptions and

expressed dissatisfaction with the responsiveness of customer support.

c. Cluster 3 - Customer Interaction and Technical Support:

Common words such as “smartfren,” “terima,” “nomor,” “teman,” “kasih,” “infoin,” “alamat,” “layanan”. This cluster was dominated by interactions between users and ISP customer support. Users frequently requested additional information, help with technical issues, and updates on service status. This cluster also included expressions of gratitude for resolved issues.

5. CONCLUSION

This study aimed to analyze user sentiments and cluster their opinions regarding ISP services in Indonesia based on tweets from the social media platform X (formerly known as Twitter). By leveraging the CRISP-DM framework, the research systematically collected, processed, and analyzed 6,000 tweets, focusing on understanding negative sentiments to identify key issues and areas for improvement. The sentiment analysis revealed that negative sentiments were predominant, highlighting technical issues like network disruptions and slow internet speeds, high service costs, and poor customer support as major user concerns.

The K-Means clustering algorithm identified three primary clusters: Cluster 1 focused on technical issues and high costs, Cluster 2 on service and customer support issues, and Cluster 3 on customer interactions and technical support. The evaluation of these clusters, using the Elbow Method to determine the optimal number of clusters and calculating inter-cluster distances and intra-cluster variations.

Based on these findings, several managerial implications were identified. ISPs need to invest in improving their infrastructure to reduce network disruptions and enhance internet speeds, which can significantly improve user satisfaction and reduce negative sentiments. Enhancing customer support services with timely and effective responses to complaints, potentially through advanced support systems like AI-driven chatbots, can further improve user experiences. Reviewing pricing strategies to offer more competitive and value-for-money plans, along with proactive communication about service status, upcoming maintenance, and issue resolutions, can help manage user expectations and build trust and loyalty among users.

REFERENCES

Annur, C. M. (2023). *Jumlah Pengguna Twitter Indonesia Duduki Peringkat ke-4 Dunia per Juli 2023*. Katadata. https://databoks.katadata.co.id/media/statistik/5cb357372e82c2d/jumlah-pengguna-twitter-indonesia-duduki-peringkat-ke-4-dunia-per-juli-2023

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS Inc*, *9*(13), 1–73.

Chowdhury, G. G., & Chowdhury, S. (2003). *Introduction to digital libraries*. Facet publishing.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*.

Hashfi, F., Sugiarto, D., & Mardianto, I. (2022). Sentiment Analysis of An Internet Provider Company Based on Twitter Using Support Vector Machine and Naïve Bayes Method. *Ultimatics: Jurnal Teknik Informatika*, *14*(1), 1–6.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666.

Kyriakou, K.-I. D., & Tselikas, N. D. (2022). Complementing JavaScript in High-Performance Node. js and Web Applications with Rust and WebAssembly. *Electronics*, *11*(19), 3217.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137.

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5-Th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*.

NLPTOWN. (2021). *bert-base-multilingual-uncased-sentiment*. Hugging Face. https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment/blob/main/README.md

Ruth, E. (2013). Deskripsi kualitas layanan jasa akses internet di Indonesia dari sudut pandang penyelenggara. *Buletin Pos Dan Telekomunikasi*, *11*(2), 137–146.

Wang, W., & Liu, Y. (2022). Distributed Optimization of Social Welfare and Regulation in Industrial Economy. *Mathematical Problems in Engineering*, *2022*(1), 3232321.

Widi, S. (2023). *Pengguna Media Sosial di Indonesia Sebanyak 167 Juta pada 2023*. DataIndonesia.ID. https://dataindonesia.id/internet/detail/pengguna-media-sosial-di-indonesia-sebanyak-167-juta-pada-2023

Yani, D. D. A., Pratiwi, H. S., & Muhardi, H. (2019). Implementasi web scraping untuk pengambilan data pada situs marketplace. *JUSTIN (Jurnal Sistem Dan Teknologi Informasi)*, *7*(4), 257–262.

Copyrights