ANALYSIIS FAVORITE GENERAL HOSPITALS IN WEST JAVA BASED ON INPATIENT VISITS USING K-MEANS SENTIMENT ANALYSIS

Dadan Zaliluddin[1], Ii Sopiandi[2], Yoga Hermawan[3]

[1,2,3] Department of Informatics, Faculty of Engineering, Majalengka University

Email: dadanzuu@gmail.com

ABSTRACT

This study examines the utilization of the K-Means algorithm for sentiment analysis in widely used hospital services utilizing the Python programming language. The main goal is to improve comprehension of patient satisfaction with the healthcare services provided at these hospitals. The data used for sentiment analysis was obtained via scraping patient evaluations from the web. The K-Means technique was utilized to classify the feelings into negative, neutral, and positive categories through the study of large-scale data. This investigation offers useful insights into the specific aspects that influence patients' opinions of healthcare services at their preferred hospitals. The study's findings provide valuable insights for hospital management to enhance the quality of healthcare services. Utilizing the K-Means algorithm in sentiment analysis facilitates the identification of prevalent trends and patterns that may not be discernible through manual techniques. Thus, this study integrates computational methodologies and sentiment analysis to offer a more holistic perspective on patient experiences at preferred hospitals.

Keywords : Big Data, Python, K-Means, Hospitals, Healthcare Services

1. INTRODUCTION

In the present age of information transparency, the general populace can readily obtain the information they require via social media platforms. Individuals who lack firsthand experience with general hospitals might assess the quality of hospital services by perusing the remarks and opinions of people who have visited the hospital. Sentiment analysis has emerged as a powerful instrument for quantifying public sentiment across various fields of study. The primary applications of this technology are financial market forecasting, health analysis, customer behavior analysis, commercial evaluation, brand promotion, political analysis, criminal prediction, and emergency response planning. The primary focus of research lies in the sentiment analysis of Twitter messages due to the platform's extensive and heterogeneous user base, which consistently shares thoughts on a wide range of subjects.(Rodríguez-Ibánez et al., 2023) pysentimiento is a versatile Python toolkit specifically developed for analyzing opinions and performing natural language processing tasks related to social media. The toolbox offers cutting-edge algorithms for analyzing sentiment, detecting emotions, identifying hate speech, and detecting irony in several languages. (Manuel Pérez et al., n.d.) The Openness of Public Information (KIP) Law Number 14 of 2008 in Indonesia governs the entitlement of individuals to access information, as well as the responsibilities of public institutions to promptly and efficiently fulfill information requests, while keeping costs cheap and procedures uncomplicated. The primary goal of this legislation is to guarantee the openness and responsibility of governmental entities in executing their responsibilities, while also promoting the involvement of the public in making public decisions. This legislation also governs the procedure for raising concerns and resolving problems related to information, which can be accomplished by engaging with the Information Commission. (Muhammad Nasrullah & Taufiq, 2008)

General hospitals can be classified into two categories: government-run and privately-owned. Government general hospitals are institutions that are both owned and administered by the government. These hospitals provide medical services to patients at a lesser cost than private hospitals. The reason for this is the presence of services catering to families facing financial limitations, such as social security cards and health access programs. Medications are also obtainable at reduced prices, and the expense of hospitalization in these medical facilities is relatively economical. Conversely, private general hospitals are owned and managed by private corporations, offering exceptional levels of comfort and service quality. Optimal medical care encompasses both the provision of top-quality medical treatment and the assurance that patients and their families receive compassionate support and empathetic understanding.(Cartwright, n.d.)

Hospitals have the objective of delivering medical care and treatment to patients while also generating income through the provision of services and amenities, such as inpatient treatments. Nevertheless, hospitals should enhance the caliber of their services by giving priority to patients. Ensuring patient happiness is vital for the efficient functioning of healthcare services, as hospitals primarily aim to treat illnesses and restore well-being. The digitization of healthcare and its influence on the caliber, effectiveness, and availability of medical treatment. Although numerous pilot initiatives have achieved success, the integration of these projects into the regular healthcare system on a big scale is still deficient.(Maier et al., 2021)

Data mining is the systematic gathering of substantial amounts of data, also referred to as big data, which is subsequently analyzed to uncover important or advantageous insights. In light of this condition, the researchers opted to undertake an investigation entitled "Analysis of Preferred General Hospitals in West Java Based on Inpatient Visits." The analysis utilizes data from 94 patient visits to ascertain the number of general hospitals in West Java. The researchers utilized the K-Means clustering technique and employed Python programming to execute a basic data mining application. The K-Means algorithm is more accurate than the human technique in some situations. Therefore, applying this algorithm can enhance efficiency.(Ilmiah Komputer Jl Ahmad Yani et al., n.d.)

It is anticipated that this study will give hospital administration helpful insights for raising the standard of medical care. The study intends to identify common trends and patterns that might not be apparent through manual techniques by assessing patient attitudes using the K-Means algorithm. The results will aid in comprehending particular elements that impact patients' opinions about the medical care they receive at their chosen hospitals. The study's findings demonstrate that hospital services continue to confront a number of difficulties that hinder their speed and dependability, such as physician delays and a shortage of medical staff in the pharmacy and registration divisions. Other qualities, meanwhile, like the staff's warmth and their attentiveness to patients, are thought to be rather good. Even if there are a few areas that need to be changed to raise the standard of care overall, patients still trust this hospital with their medical care.(Listiyono, 2015) The significance of raising the standard of nursing care, particularly in terms of giving patients complete and correct information, upholding professionalism and skill, and guaranteeing their comfort and safety. It is anticipated that this endeavor will raise hospital inpatients' levels of satisfaction. (Silalahi et al., 2019) Patient satisfaction data can be grouped into three primary groups with the use of the K-Means Clustering method, which makes subsequent analysis easier. (Zulvia et al., 2022)

2. SUBMISSION

Menguraikan This section describes the research's sequential processes, as well as the methodologies, data collection techniques, and research design. The K-nearest neighbors (KNN) algorithm operates by determining the K closest neighbors to the provided data and utilizing the class labels or values associated with those neighbors to generate predictions. This study involves the collection and processing of tweet data using several preprocessing techniques, including case folding, tokenizing, stopword removal, and weighing. The investigation indicates that the KNN technique outperforms Naïve Bayes, with an accuracy rate of 88.12%, precision of 93.98%, and recall of 81.53%.(Syahril Dwi Prasetyo et al., 2023)

2.1. Research Design

The study design encompasses multiple steps to ensure a thorough examination. The first phase entails conducting a comprehensive literature review to gain a thorough understanding of the theoretical foundation and prior research pertaining to sentiment analysis and the K-Means algorithm. Following this stage, the subsequent steps include data gathering, data preprocessing, clustering, and analysis of the results. This approach presents three multi-view cluster validity indices, namely MV-Dunn, MV-G-Dunn, and MV-M-Dunn, which are used to assess the performance of clustering. Experimental results on both synthetic and actual data demonstrate that the U-MV-KM algorithm outperforms previous techniques.(Yang & Hussain, 2023)

2.2. Data Collection

The data for this study was obtained from patient reviews accessible on social media platforms and hospital websites. The reviews were obtained by employing web scraping techniques to collect sentiment data. The dataset comprises reviews from multiple general hospitals in West Java, with a specific emphasis on inpatient services. The analysis approach entails utilizing text mining techniques to extract information and categorize evaluations into two distinct sentiment groups: positive and negative.(Amrullah et al., 2020)

2.3. Data Preprocessing

The gathered data was thereafter subjected to preprocessing to guarantee its precision and uniformity. The preprocessing procedures encompassed data cleansing by eliminating irrelevant or duplicate information, text normalization, and transforming it into a structured format appropriate for analysis. Data preparation is a crucial step in the knowledge discovery from data (KDD) process. This step encompasses the processes of choosing, purifying, modifying, and minimizing data. The objective of this procedure is to identify, remove, and refine superfluous, contradictory, and deficient data in order to enhance the usefulness of the data.(Tsai et al., 2015)

2.4. Clustering with K-Means Algorithm

An illustration of implementing the K-Means algorithm involves dividing product sales data into two primary clusters: products with significant sales and products with little sales [14].The utilization of the K-Means clustering algorithm has demonstrated its efficacy in discerning cohorts of patients according to specific attributes, hence yielding valuable insights to enhance health management and services. The user's text is "[15]". The sentiment data was analyzed using the K-Means clustering algorithm. The procedure consists of the following steps :

a. Importing Libraries and Dataset:
   a) The essential Python libraries, including Pandas, NumPy, and Scikit-learn, were imported..
   b) The dataset was imported into the environment..

b. Data Cleaning and Preparation:
   a) The dataset underwent a process of cleansing to exclude any extraneous or inconsequential data..
   b) The text data was transformed into a numerical representation using the TF-IDF (Term Frequency-Inverse Document Frequency) technique, which is ideal for clustering.
   c) Determining the Optimal Number of Clusters:
   d) The Elbow approach was employed to ascertain the appropriate number of clusters. This approach entails graphing the within-cluster sum of squares (WCSS) versus the number of clusters and pinpointing the "elbow point" when the WCSS begins to fall at a slower pace.

c. Applying the K-Means Algorithm:
   a) The K-Means methodology was utilized on the dataset, with the ideal number of clusters determined by the Elbow approach.
   b) The program employed an iterative process to allocate data points to clusters according to their similarity and subsequently recalibrated the cluster centroids until convergence was attained.

d. Analysis of Clusters:
   a) An analysis was conducted on the generated clusters to examine the distribution of attitudes (positive, neutral, negative) among various hospitals.
   b) The clustering findings were presented using visualization techniques, specifically scatter plots.

2.5. Tools and Environment

The study was performed utilizing Google Colab, an internet-based platform that enables the execution of Python code in a cloud-hosted environment. This platform offers the essential computing resources and libraries needed for data processing. The categorization method is executed with the Natural

Language Toolkit (NLTK) in the Python programming language. The topic covered is characterized by a prevailing neutral sentiment. (Suryono et al., n.d.) Python is renowned for its ability to facilitate parallel programming through the utilization of several libraries and packages. This research examines many Python libraries, including multiprocessing, JMetalPy, Parsl, Ray, PyWren, and PyNetLogo, that provide support for parallel processing and multiprocessing. (A. Aziz et al., 2021)

2.6. **Validation and Evaluation**
The accuracy and dependability of the model were assessed by the utilization of cross-validation techniques. The clustering quality was evaluated using metrics such as the silhouette score. Furthermore, the findings were cross-referenced with manual comments to confirm their accuracy. The validation and assessment process is crucial to guarantee that the clustering results are suitable for decision-making in management. (Rizki Nugroho & Edo Hendrawan, 2022) This assessment offers useful insights for the strategic planning and effective management of future pandemics, highlighting the importance of tailoring palliative services to address the unique requirements of individual patients. (Hetherington et al., 2020) There are several important obstacles that need to be addressed in order to fully utilize the promise of big data in healthcare. These include issues related to inconsistent data quality, ethical and legal constraints, and the requirement for clinical validation. The introduction of big data research methodologies is expected to unravel the complexity of IBD and have potential applications in other health domains. (Olivera et al., 2019) Teaching hospitals had reduced adjusted mortality rates in nine out of fifteen studies that assessed this parameter.(Kupersmith, n.d.)

3.   RESULT AND DISCUSSION
Berisi This part provides an overview of the research findings and includes a thorough analysis and discussion. The results are presented in a methodical manner, with an explanation of the approach utilized, and supported by appropriate references.

3.1. **Data Overview**
The dataset utilized in this investigation consists of patient reviews obtained from diverse general hospitals in West Java. The reviews were obtained by using web scraping techniques and encompassed opinions pertaining to inpatient services. The dataset underwent preprocessing and vectorization using the TF-IDF approach to transform the textual data into a numerical representation that is appropriate for clustering.

3.2. **Determination of Optimal Clusters**
The Elbow approach was utilized to ascertain the most suitable number of clusters for the K-Means algorithm. Figure 1 and Figure 2 displays the plot of the within-cluster sum of squares (WCSS) as a function of the number of clusters. The elbow point

was determined at K=3, signifying that three clusters offer the optimal representation of the data.
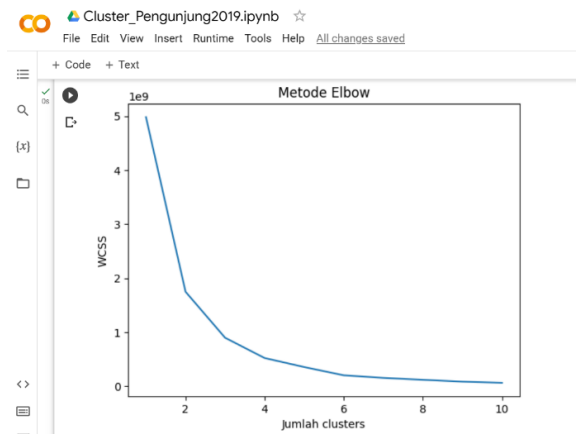

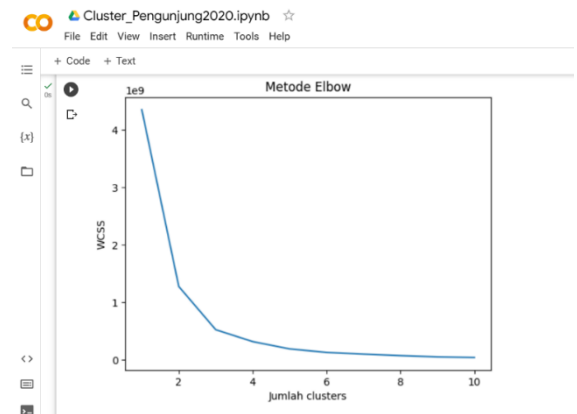**Figure 1. Graph of clustering of inpatients in West Java in 2019**

Figure 2. Graph of clustering of inpatients in West Java in 2020

3.3. **Clustering Results**
The K-Means technique was utilized with a value of K equal to 3, which led to the creation of three clearly distinguishable clusters. The clusters correspond to distinct attitude categories: positive, neutral, and negative. Figure 3 and Figure 4 visualizes the distribution of feelings inside the clusters.
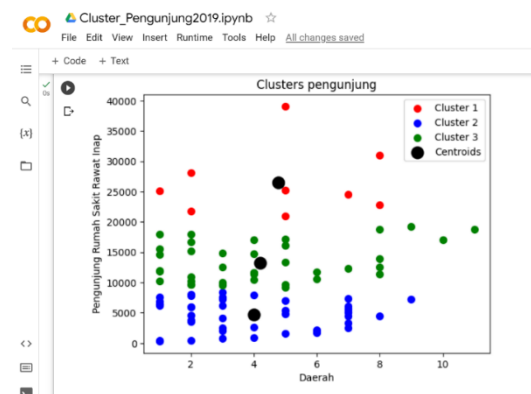

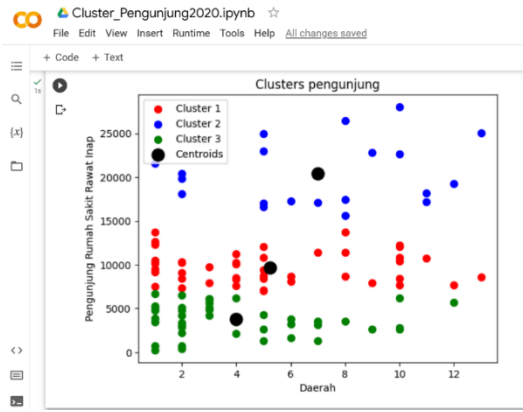**Figure 3. clustering of inpatient visitors in 2019**

Figure 4. clustering of inpatient visitors in 2020

The area is described as follows:

Table 1. Area code description in the clustering display

No	2019	No	2020
1	Bogor	1	Bogor
2	Karawang	2	Karawang
3	Depok	3	Depok
4	Kuningan	4	Kuningan
5	Bandung	5	Bandung
6	Subang	6	Subang
7	Bekasi	7	Tasikmalaya
8	Cirebon	8	Cirebon
9	Banjar	9	Sumedang
10	Cianjur	10	Sukabumi
11	Tasikmalaya	11	Majalengka
		12	Banjar
		13	Cianjur

3.4. Analysis of Clusters

The examination of the clusters yielded valuable information regarding patient attitudes towards hospital services. Below are the summarized important findings:

Cluster 1 (Positive Sentiments):
   a. This cluster comprises reviews that reflect contentment with the services provided by the hospital. Patients valued the high standard of healthcare, the competence of the staff, and the general ambiance of the institution.
   b. Review example: "The medical professionals displayed exceptional attentiveness and delivered outstanding care." The hospital exhibited a pristine and impeccably-maintained environment.

Cluster 2 (Neutral Sentiments):
   a. This cluster consists of reviews that have a neutral tone. Patients expressed a range of opinions on their experience, without a clear preference for either positive or negative sentiment.
   b. Review: "The service was satisfactory, however, there is potential for enhancement in terms of waiting times and responsiveness."

Cluster 3 (Negative Sentiments):
   a. This cluster comprises reviews that convey discontentment with the quality of hospital treatments. Frequent grievances encompassed extended waiting periods, ineffective communication, and insufficient amenities.
   b. Review: "The duration of waiting was excessively prolonged, and the staff appeared to lack organization." The facilities did not meet the required standard.

3.5. Discussion

The clustering results reveal significant trends and patterns in patient emotions.
   a. Service Quality: Hospitals that offer high-quality care, efficient service, and effective communication are primarily related with positive attitudes. Hospitals possessing these characteristics are more likely to get elevated patient satisfaction ratings.
   b. Areas for Improvement: Neutral and negative views frequently indicate particular areas that require enhancement, such as decreasing waiting times, improving staff training, and updating facilities. Tackling these concerns can greatly enhance patient satisfaction.
   c. Hospital Management: The findings derived from the sentiment analysis can assist hospital administration in determining the most important areas for enhancement and efficiently distributing resources. An in-depth comprehension of how patients perceive healthcare services enables the implementation of focused interventions to improve the quality of care.
   d. Comparison with Previous Studies: The results of this study align with other studies on the quality of healthcare services and the happiness of patients (Amrullah et al., 2020; Susilowati & Jayanto, 2022). The utilization of the K-Means algorithm in sentiment analysis offers a reliable approach to finding crucial elements that impact patient satisfaction.

3.6. Visualization of Results

Visualization approaches, such as scatter plots and bar charts, are quite useful in conveying the clustering results and emotion distributions. Figures 5 and 6 depict the dispersion of sentiments among

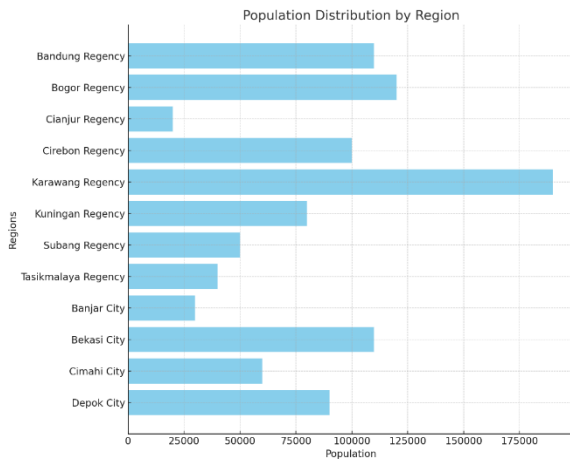various hospitals and the percentage of reviews in each sentiment category..

Figure 5. graph of the comparative number of hospital inpatients in West Java in 2019

According to the data presented in the figure above, the Karawang area has the highest number of inpatient visits among public hospitals in West Java. Karawang Regency can be classified as a metropolis due to its large population, which is why there are a total of 16 public hospitals in the district. Therefore, it is unsurprising that the aforementioned figures indicate that Karawang has the highest number of visitors, with a total of 180,361 patients.
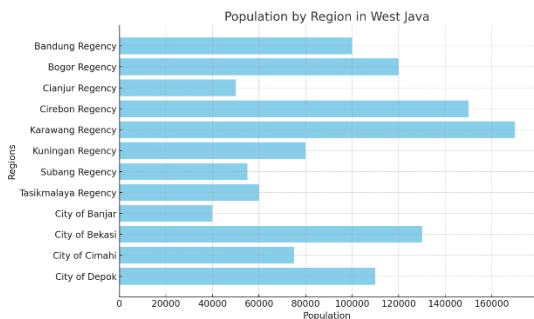

Figure 6. graph of the comparative number of hospital inpatients in West Java in 2020

Unlike the figures from 2019, the public hospital in Bogor Regency experienced the highest number of visits in 2020, with a total of 142,170 inpatients. Previously, Karawang district held the top spot in 2019, but in 2020 it dropped to second place with 137,322 inpatients.
Based on the aforementioned facts, it can be inferred that Bogor Regency and Karawang Regency are populous urban areas. It is inherent that public hospitals in these two regions cater to a large number of patients.

4.   CONCLUSION
This study examined the application of the K-Means method for sentiment analysis on patient reviews of general hospitals in West Java. The main

discoveries and their consequences are outlined as follows:

4.1. Effectiveness of K-Means Algorithm:
The K-Means algorithm successfully clustered patient reviews into distinct sentiment groups, namely positive, neutral, and negative. After careful analysis, it was found that the most suitable number of clusters is three, which effectively separates patient sentiments into distinct groups.

4.2. Patient Sentiment Insights:
a. Positive Sentiments: Reviews in this group emphasized that patient satisfaction was influenced by high-quality care, competent staff, and well-maintained facilities.

b. Neutral Sentiments: These assessments provided an equitable perspective, highlighting both positive and negative aspects, and recommending areas for enhancement.

c. Negative Sentiments: Frequently voiced grievances including extended waiting periods, deficient communication, and substandard amenities, highlighting specific areas for remediation.

4.3. Practical Implications for Hospital Management:
The knowledge acquired from the sentiment analysis can assist hospital administration in identifying crucial areas for enhancement. Hospitals can improve patient satisfaction and service quality by addressing the concerns raised in negative and neutral ratings.

4.4. Consistency with Previous Research:
The results are consistent with prior research on the quality of healthcare services and patient contentment, validating the effectiveness of the K-Means algorithm as a reliable method for sentiment analysis in the healthcare industry.

4.5. Future Work:
In order to expand upon the discoveries of this study, further investigations could concentrate on:
a. Gathering further data from a more extensive selection of hospitals to improve the applicability of the findings.

b. Utilizing more sophisticated sentiment analysis methods, such as aspect-based sentiment analysis, to obtain more profound understanding of particular facets of healthcare services.

c. Investigating the application of alternative machine learning algorithms, such as Random Forest or Deep Learning, in order to assess their efficacy in comparison to the K-Means method.

By focusing on these specific areas, future research has the potential to enhance our understanding of patient attitudes and enhance the quality of healthcare services.

REFERENCES

A. Aziz, Z., Naseradeen Abdulqader, D., Sallow, A. B., & Khalid Omer, H. (2021). Python Parallel Processing and Multiprocessing: A Rivew. *Academic Journal of Nawroz University*, *10*(3), 345–354. https://doi.org/10.25007/ajnu.v10n3a1145

Amrullah, A. Z., Sofyan Anas, A., Adrian, M., & Hidayat, J. (2020). Analisis Sentimen Movie Review Menggunakan Naive Bayes Classifier Dengan Seleksi Fitur Chi Square. *Jurnal*, *2*(1). https://doi.org/10.30812/bite.v2i1.804

Cartwright, A. (n.d.). *Human Relations and Hospital Care*.

Hetherington, L., Johnston, B., Kotronoulas, G., Finlay, F., Keeley, P., & McKeown, A. (2020). COVID-19 and Hospital Palliative Care – A service evaluation exploring the symptoms and outcomes of 186 patients and the impact of the pandemic on specialist Hospital Palliative Care. *Palliative Medicine*, *34*(9), 1256–1262. https://doi.org/10.1177/0269216320949786

Ilmiah Komputer Jl Ahmad Yani, J., STMIK Banjarbaru Loktabat -Banjarbaru, -Kampus, Jananto, A., & Tri Lomba Juang No, J. (n.d.). *Penerapan Algoritma K-Means Clustering untuk Perencanaan kebutuhan Obat di Klinik Citra Medika ……….Lailil ' Izzah Penerapan Algoritma K-Means Clustering Untuk Perencanaan Kebutuhan Obat Di Klinik Citra Medika*.

Kupersmith, J. (n.d.). *Quality of Care in Teaching Hospitals: A Literature Review*. http://journals.lww.com/academicmedicine

Listiyono, R. A. (2015). Studi Deskriptif Tentang Kuaitas Pelayanan di Rumah Sakit Umum Dr. Wahidin Sudiro Husodo Kota Mojokerto Pasca Menjadi Rumah Sakit Tipe B. In *Kebijakan dan Manajemen Publik* (Vol. 1, Issue 1).

Maier, E., Reimer, U., & Wickramasinghe, N. (2021). Digital healthcare services. In *Electronic Markets* (Vol. 31, Issue 4, pp. 743–746). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s12525-021-00513-z

Manuel Pérez, J., Rajngewerc, M., Giudici, J. C., Furman, D. A., Luque, F., Alemany, L. A., & Vanina Martinez, M. (n.d.). *pysentimiento: A Python Toolkit for Opinion Mining and Social NLP tasks*. https://github.com/sloria/TextBlob/

Muhammad Nasrullah, A., & Taufiq, A. (2008). *Implementasi undang-undang Nomor 14 Tahun*.

Olivera, P., Danese, S., Jay, N., Natoli, G., & Peyrin-Biroulet, L. (2019). Big data in IBD: a look into the future. In *Nature Reviews Gastroenterology and Hepatology* (Vol. 16, Issue 5, pp. 312–321). Nature Publishing Group. https://doi.org/10.1038/s41575-019-0102-5

Rizki Nugroho, M., & Edo Hendrawan, I. (2022). *Penerapan Algoritma K-Means Untuk Klasterisasi Data Obat Pada Rumah Sakit ASRI*. *16*(1). https://journal.uniku.ac.id/index.php/ilkom

Rodríguez-Ibáñez, M., Casánez-Ventura, A., Castejón-Mateos, F., & Cuenca-Jiménez, P. M. (2023). A review on sentiment analysis from social media platforms. In *Expert Systems with Applications* (Vol. 223). Elsevier Ltd. https://doi.org/10.1016/j.eswa.2023.119862

Silalahi, J., Devi Fitriani, A., Kesehatan Masyarakat, F., & Kesehatan Helvetia Medan, I. (2019). Jurnal Kesehatan Perintis (Perintis's. *Health Journal*, 6.

Suryono, S., Ha, E. M., & Luthfi, T. (n.d.). *K LASIFIKASI SENTIM EN PADA TW IT TER DENGAN NAIVE BAYES CLA SSIFIER*.

Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, & Fitri Nurapriani. (2023). Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN. *Jurnal KomtekInfo*, 1–7. https://doi.org/10.35134/komtekinfo.v10i1.330

Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. (2015). Big data analytics: a survey. *Journal of Big Data*, *2*(1). https://doi.org/10.1186/s40537-015-0030-3

Yang, M. S., & Hussain, I. (2023). Unsupervised Multi-View K-Means Clustering Algorithm. *IEEE Access*, *11*, 13574–13593. https://doi.org/10.1109/ACCESS.2023.3243133

Zulvia, I., Hidayatulloh, A., & Rahmawati, D. E. (2022). ANALISIS KEPUASAN PASIEN TERHADAP PELAYANAN KESEHATAN DI KLINIK ALKINDI HERBAL MENGGUNAKAN METODE K-MEANS CLUSTERING. *Jurnal Informatika Dan Komputer)*, *6*(2), 261–272.