



Analysis of machine learning approaches to determine online shopping ratings using naïve bayes and svm

Muit Sunjaya¹, Zulham Sitorus², Khairul³, Muhammad Iqbal⁴, A.P.U Siahaan⁵

^{1,3} Department of Computer System, Universitas Pembangunan Panca Budi, Indonesia

²Department of Computer System Engineering, Universitas Pembangunan Panca Budi, Indonesia

Article Info

Article history:

Received May, 2024

Revised May, 2024

Accepted May, 2024

Keywords:

Naïve Bayes, Support Vector Machine
sentiment analysis
classification
machine learning
lazada

ABSTRACT

This research aims to identify and compare the effectiveness of Naïve Bayes and Support Vector Machine (SVM) algorithms in classifying ratings based on customer comments on the Lazada online shopping platform. The main issues identified include data uncertainty, model selection and optimization, as well as improving efficiency and scalability. Using a dataset of comments and reviews from Lazada, this study conducts an analysis using both algorithms to determine which is most effective in classifying comments into appropriate ratings. The research methodology includes data collection, text preprocessing, algorithm implementation, and evaluation using a confusion matrix to measure accuracy, precision, recall, and F-measure. This analysis is supported by data visualization using Python, allowing for in-depth interpretation and understanding of the results. The research findings show significant differences in the performance of both algorithms, with each having strengths in certain aspects of classification. The discussion in this study interprets these results to address the research questions formulated and demonstrates the practical application of machine learning theory in real-world data processing. This study concludes that both algorithms have significant potential in sentiment classification but require further adjustment and optimization to improve accuracy and efficiency. Recommendations for further research include the development of hybrid models or new approaches that can address the identified limitations, as well as exploration of more diverse datasets to test the scalability of the proposed solutions.

This is an open access article under the [CC BY-NC](https://creativecommons.org/licenses/by-nc/4.0/) license.



Corresponding Author:

Muit Sunjaya,
Department of Electrical and Computer Engineering,
Universitas Pembangunan Panca Budi,
Jl. Jend. Gatot Subroto Km. 4,5 Sei Sikambing 20122 Medan, Propinsi Sumatera Utara, Indonesia.
Email: muitsunjaya@gmail.com

1. INTRODUCTION

Technology is a means or system that provides comfort and convenience for humans. In this era, technological developments are increasingly experiencing very rapid developments. This is due to the diversity of people's needs which are supported by various types of technology and also the increasingly tight competition between technology providers. With the development of information technology, people can easily fulfill their daily needs. Technological developments have shifted customer behavior from purchasing through offline shops to purchasing through online shops or via e-commerce.

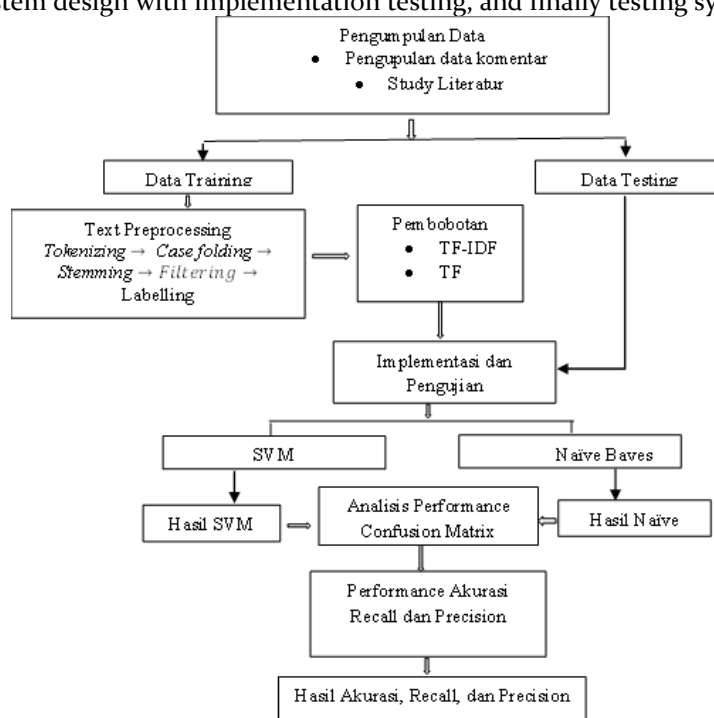
Changes in people's behavior in e-commerce are a challenge for companies to be able to meet the needs and desires of their consumers. With the increasing types of E-Commerce, of course it will greatly influence the competitiveness between E-Commerce. This causes every businessman who joins E-Commerce to have their own way to increase consumer attraction to E-Commerce. Things that E-Commerce business people always pay attention to to increase consumer attraction are paying attention to the quality of service to consumers, using News Letters, always working on product upsells, doing advertising retargeting. However, it is not uncommon for consumers to doubt E-Commerce products because consumers cannot try or see directly the products or services of the E-Commerce Producer in question.

The topic of this research is how to identify text or comments on online shops that will be analyzed and visualized so that they can be used as a rating value for the product being purchased. The process of identifying text in a database is called text mining. Text mining is the process of extracting high quality information from high quality information texts. Text mining aims to obtain useful information from a document

The data set source is from Playstore with the search keyword Shopee using scraping. Accuracy measurements produce a confusion matrix for assessing precision and recall. It is on this basis that this research takes the title: "Analysis of Machine Learning Approaches to Predict Online Shopping Ratings in Online Shopping Applications Using the Naïve Bayes Algorithm and Support Vector Machine"

2. RESEARCH METHOD

In this research, in general, the stages that will be carried out to design an identification system start from collecting data, studying literature such as looking for related references, then carrying out system design with implementation testing, and finally testing system performance.

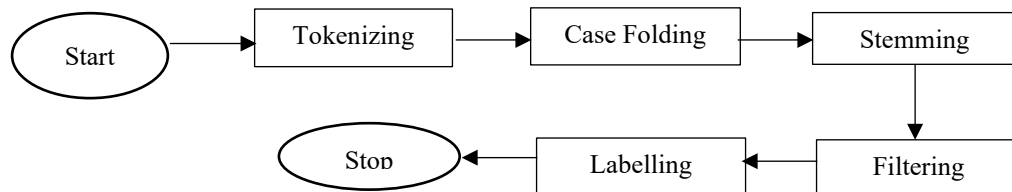


Following are the research stages in Figure

Research Methodology Stages

Preprocessing Analysis

Data preprocessing is an important step in carrying out classification analysis which aims to clean the data from elements that are not needed to speed up the classification process. below is a flowchart of the data preprocessing stages used which can be seen in Figure



Preprocessing Stage Diagram

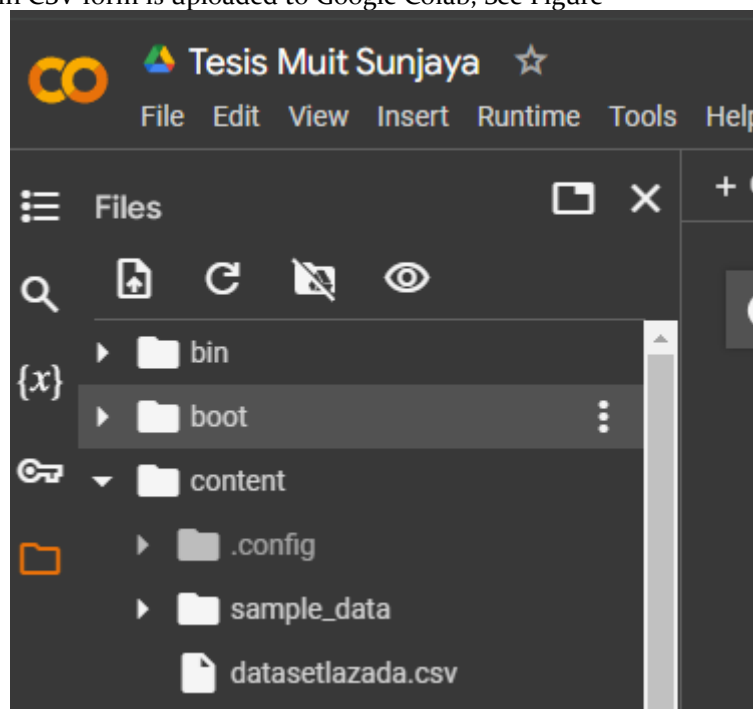
3. RESULTS AND DISCUSSIONS

This chapter explains the results of the research implementation. The initial stage of this research is data preprocessing, then continued with the data normalization stage with TF-IDF weighting, then classification is carried out using Support Vector Machine and Naive Bayes. This chapter will describe how to present data and explain it clearly and regularly.

3.1. Sub section 1

Research data was taken from the Lazada platform site which was taken from Playstore as many as 1044 comment and rating data. Then the data is collected and saved in a CSV file with the notepad++ editor application. This research data can be accessed at the following link: <https://bit.ly/3hgMMwN>

So that the data classification and visualization process can be processed first, the Research Dataset which is in CSV form is uploaded to Google Colab, See Figure



To call a research dataset, type "datasetlazada.csv", use the pseudocode:

#1.2. Import file data.csv

```
data = pd.read_csv('datasetlazada.csv',encoding='latin-1')
```

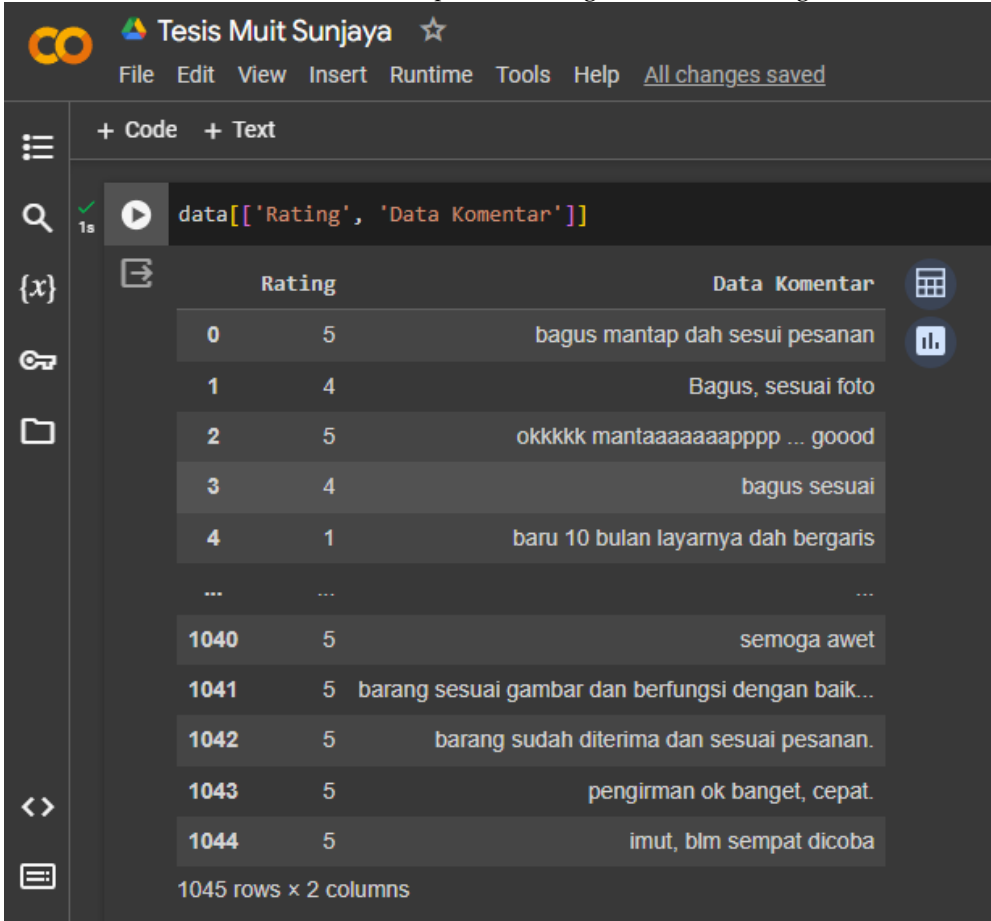
Apart from processing data for classification, this research will also display the results of data visualization in order to get an overview of the data to be processed. What will be displayed in the data visualization is as follows:

1. Displays the dataset histogram data program.

```
data_bersih_01['length']=data_bersih_01['hasil_akhir'].apply(len)
data_bersih_01['length'].plot(bins=50,kind='hist')
data_bersih_01[['Komentar_clean','length']].head()
```
2. Program to display terms/words that frequently appear in the dataset using the WorldCloud library.

```
# grab the sentence and show it
kalimat = data_bersih_01['hasil_akhir'].tolist()
```

Labeling is the process of determining the sentiment in comment data into Positive, Neutral and Negative. In this research, the labeling process uses comment data from consumers along with the ratings that have been given. After labeling, the data processing is then continued to obtain classification results. An example of labeling can be seen in Figure:



	Rating	Data Komentar
0	5	bagus mantap dah sesuai pesanan
1	4	Bagus, sesuai foto
2	5	okkkkk mantaaaaaapppp ... good
3	4	bagus sesuai
4	1	baru 10 bulan layarnya dah bergaris
...
1040	5	semoga awet
1041	5	barang sesuai gambar dan berfungsi dengan baik...
1042	5	barang sudah diterima dan sesuai pesanan.
1043	5	pengiriman ok banget, cepat.
1044	5	imut, blm sempat dicoba

1045 rows x 2 columns

Labels based on Ratings and Comment Data

The program used to obtain labeling results is as follows:

```

data.Rating.value_counts()

def classes_def(x):
    if x == 5:
        return 1
    elif x == 4:
        return 0
    elif x == 3:
        return 0
    elif x == 2:
        return -1
    else:
        return -1

data['class']=data['Rating'].apply(lambda x:classes_def(x))
print("Bagus: ", data[data['class'] == 1].shape)
print("Netral: ", data[data['class'] == 0].shape)
print("Kurang Bagus: ", data[data['class'] == -1].shape)

```

So we get the following output:

Good : (696, 4)

Neutral : (220, 4)

Not Good : (129, 4)

So we get some data that already has labels which can be seen in the table:

ID User	Komentar	Label
1	Bagus mantap dah sesuai pesanan	Bagus
2	Baru 10 bulan layarnya dah bergaris	Tidak Bagus
3	Pengiriman super lama.. tapi datang juga sich	Netral
4	Pengirim barang tidak sesuai janji	Tidak Bagus
5	Imut belum sempat dicoba	Bagus
6	Barang sesuai gambar dan berfungsi dengan baik	Bagus

3.2. Sub section 2

The data that has been taken, which is already in the form of a CSV file, needs to be imported and displayed on the system. In the previous chapter, we discussed how to import the file into the system. Figure 5.1 is a display of a system that successfully displays research data:

```
data[['Rating', 'Data Komentar']]
```

	Rating	Data Komentar
0	5	bagus mantap dah sesuai pesanan
1	4	Bagus, sesuai foto
2	5	okkkkk mantaaaaaaapppp ... good
3	4	bagus sesuai
4	1	baru 10 bulan layarnya dah bergaris
...
1040	5	semoga awet
1041	5	barang sesuai gambar dan berfungsi dengan baik...
1042	5	barang sudah diterima dan sesuai pesanan.
1043	5	pengiriman ok banget, cepat.
1044	5	imut, blm sempat dicoba

1045 rows × 2 columns

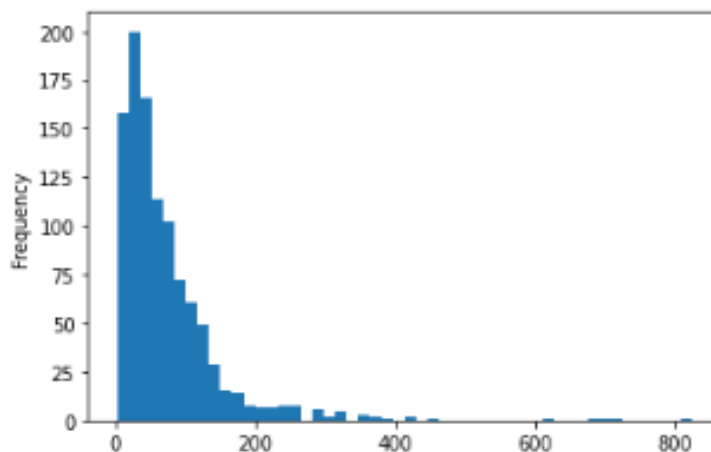
Results of Research Data Display in the System

```
print("Shape: ", id_stopword_dict.shape)
id_stopword_dict.head()
```

```
Shape: (758, 1)
```

	stopword
0	ada
1	adalah
2	adanya
3	adapun
4	agak

Stopword Term Display

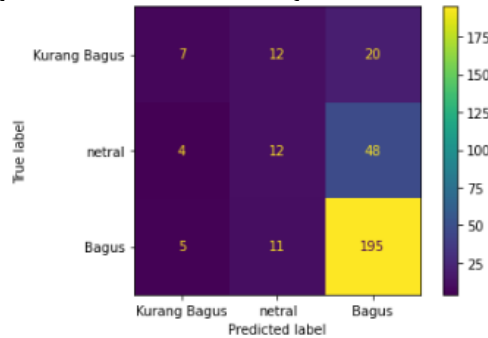


Research Dataset Histogram



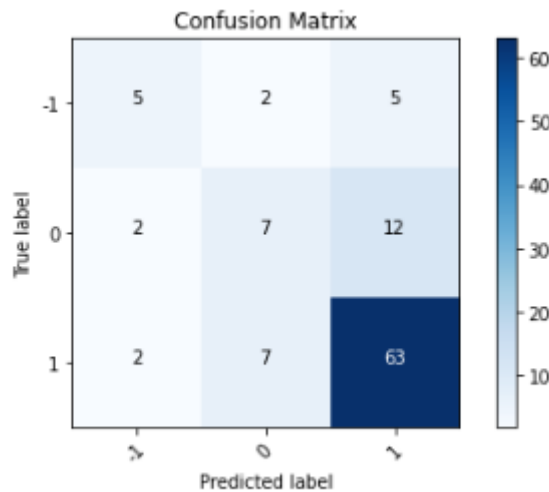
Wordcloud Research Dataset

The results of the Naïve Bayes confusion matrix analysis can be seen in Figure:



Naive Bayes Confusion Matrix Diagram

The results of the SVM confusion matrix analysis can be seen in Figure:



Analisis Cofusion Matrix

System Testing

System testing to check the results of the classification system that has been built using the Naïve Bayes Algorithm and SVM. The classification results are in the form of a report containing the program running time, Precision, Recall and F1-score and Support values. Then a comparison was carried out between the Naïve Bayes Algorithm and Support Vector Machine (SVM) to get the best accuracy value. Table V.1 is the result of system testing.

System Testing Results Using Naive Bayes and SVM

	Precision	Recall	Accuracy	Waktu (detik)
Naive Bayes	0.64	0.69	0.65	0.0274522304
Support Vector Machine	0.69	0.71	0.70	1. 14799261

After classification, the best accuracy value was obtained by SVM, but the speed of the Naive Bayes system was superior to SVM.

Table V.2 displays a comparison between the 3 algorithms.

Comparison of the Accuracy of Each Algorithm

Data	Algoritma	Akurasi
Data Latih	Naïve Bayes	0,81
	SVM	0,91
Data Uji	Naïve Bayes	0,68
	SVM	0,71

The accuracy value obtained is not very large because it is possible that there is still noise in the dataset which results in the process of calculating the percentage accuracy value not being very good. On the training data, the highest level of accuracy was demonstrated by the SVM algorithm.

4. CONCLUSION

The research conclusions based on the results of comment data processing and ratings on the system are as follows:

1. The dataset used in this research is 1044 comments resulting from reviews on the Lazada online shopping application along with ratings. The research only carries out a classification process based on datasets that have been previously input, not to display the classification results of newly input data.
2. Based on the implementation and testing that has been carried out, the Naïve Bayes and SVM algorithms can be applied in carrying out the classification process for comments on online shops that have been tested for classification on Lazada.
3. Classification using the Support Vector Machine (SVM) algorithm is proven to be better than the Naïve Bayes algorithm based on the accuracy value of the SVM algorithm being 71.42% higher than the Naïve Bayes algorithm which only produces 69%.
4. In terms of classification processing time in the system, the Naïve Bayes algorithm provides the best processing speed with a processing time of 0.02745223045349121 seconds compared to the SVM algorithm of 4.1479926109313965

ACKNOWLEDGEMENTS

Alhamdulillah Robbil 'Alamin. Praise be to Allah, God of the Universe, who has given His grace and mercy so that we can complete this thesis with the title "ANALYSIS OF MACHINE LEARNING APPROACHES TO DETERMINE ONLINE SHOPPING RATINGS USING NAÏVE BAYES AND SVM". Writing this thesis is one of the programs to complete the Master's Postgraduate studies at Panca Budi Development University.

This thesis could be completed well because there was a lot of input and support from various parties, especially my parents who I really loved, who always guided me by providing direction and advice. Therefore, I would like to express my gratitude to the father named Nasit, even though he is in a different world, he will always be in the prayers of his children and the mother named Musiem, who is very much loved, and always provides sincere prayers and supports both morally and materially. . Invaluable gratitude is also expressed to several parties who have assisted with knowledge, guidance and moral support in the preparation of this thesis, especially for:

1. Mr. Dr. H. Muhammad Isa Indrawan, SE, M.M as Chancellor of Panca Budi Development University

2. Mrs. Dr. Kiki Farida Ferine, SE., M., as Director of the Master's Postgraduate Program at Panca Budi Development University.
 3. Mr Dr. Zulham Sitorus, S.kom, M.kom As my first supervisor who has guided the author in completing this research, as well as the Head of the Panca Budi Development University Masters Study Program, who has provided invaluable input and motivation to the author
 4. Mr Dr. Khairul, S.kom, M.kom as my II I supervisor who has guided the author in completing this research and provided invaluable input and motivation to the author.
 5. Thank you also to the entire biological family who have helped in terms of prayers, materials, thoughts and so on.
 6. Likewise, I would also like to thank the person who is extraordinarily kind to me who has given me motivation and life lessons and given me the opportunity to be his adopted younger brother and even become part of his family, namely my adopted brother Hendryan Winata, S.Kom, M .Com. I will never forget the kindness and services of my adopted brother because without his presence in my life I would not have been able to achieve what I am today.
 7. Friends and Brothers in the Afterlife, Brother Hendryan Winata, Ustadz Dede Suprayugo, Ustadz Syahrudin, Akbar Idaman, Ari Sandi, Alwi Liyunzira, and other friends in the Afterlife who cannot be mentioned one by one.
 8. All fellow students of the Computer Science Study Program Masters Program Class of 2023-2024 for their cooperation and solidarity during their studies and research.
 9. And now to Mrs. Rizki Fitriani as a companion, hopefully we will achieve what we want, thank you for giving me a sense of support system enthusiasm.
 10. And all parties who cannot be named one by one but have helped and prayed a lot for me.
- This thesis is not free from errors, shortcomings and is far from perfection both in terms of presentation and scientific aspects. For this reason, all forms of feedback, input and suggestions related to the material in this thesis are very necessary. Finally, with all humility, the author realizes that there are still many shortcomings, so the author hopes for suggestions and constructive criticism for the perfection of this thesis.

REFERENCES

- The main references are international journals and proceedings. All references should be to the most pertinent, up-to-date sources and the minimum of references are 25. References are written in IEEE style. Please use a consistent format for references – see examples below (9 pt):
- Adijaya, M.Y. 2020. *Implementasi Reduced Support Vector Machines Dalam Sistem Deteksi Kepribadian Berdasarkan Pola Tanda Tangan*.
- Amelia, R.D., Michael, M. & Mulyandi, R. 2021. Analisis Online Consumer Review Terhadap Keputusan Pembelian pada E-Commerce Kecantikan. *Jurnal Indonesia Sosial Teknologi*, 2(2): 274–280.
- Buntoro, G.A. 2017. Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter. *INTEGER: Journal of Information Technology*, 2(1).
- Cavalin, P. & Oliveira, L. 2018. Confusion matrix-based building of hierarchical classification. *Iberoamerican Congress on Pattern Recognition*. Springer, hal.271–278.
- Chaffey, D., Edmundson-Bird, D. & Hemphill, T. 2019. *Digital business and e-commerce management*. Pearson UK.
- Darmawan, A., Kustian, N. & Rahayu, W. 2018. Implementasi Data Mining Menggunakan Model SVM untuk Prediksi Kepuasan Pengunjung Taman Tabebuya. *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, 2(3): 299–307.
- Darwis, D., Pratiwi, E.S. & Pasaribu, A.F.O. 2020. Penerapan Algoritma Svm Untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia. *Eduitic-Scientific Journal of Informatics Education*, 7(1).
- DHARMAWAN, F.Y. n.d. ANALISIS JAWABAN SOAL ESSAY MENGGUNAKAN TEXT MINING DENGAN METODE NAÏVE BAYES.
- Dougherty, G. 2012. *Pattern recognition and classification: an introduction*. Springer Science & Business Media.
- Farki, A. 2016. *Pengaruh online customer review dan rating terhadap kepercayaan dan minat pembelian pada* *International Journal of Computer Sciences and Mathematics Engineering*

- online marketplace di Indonesia.*
- Firdaus, D. 2017. Penggunaan Data Mining dalam Kegiatan Sistem Pembelajaran Berbantuan Komputer. *Format*, 6(2): 91–97.
- Gunawan, B., Sastypratiwi, H. & Pratama, E.E. 2018. Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes. *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, 4(2): 113–118.
- Handayani, F. & Pribadi, F.S. 2015. Implementasi algoritma naive bayes classifier dalam pengklasifikasian teks otomatis pengaduan dan pelaporan masyarakat melalui layanan call center 110. *Jurnal Teknik Elektro*, 7(1): 19–24.
- Hidayat, A.N. 2015. Analisis Sentimen Terhadap Wacana Politik Pada Media Masa Online Menggunakan Algoritma Support Vector Machine Dan Naive Bayes. *Jurnal Elektronik Sistem Informasi dan Komputer*, 1(1): 12–18.
- HR, R.T., Wardoyo, A.E. & Pratama, M.R. n.d. Analisis Sentimen pada Twitter terhadap Kinerja Komisi Pemberantasan Korupsi (KPK) di Indonesia dengan Metode Naive Bayes.
- Ipmawati, J. 2016. Komparasi teknik klasifikasi teks mining pada analisis sentimen. *IJNS-Indonesian Journal on Networking and Security*, 6(1).
- Kim, W.G. & Park, S.A. 2017. Social media review rating versus traditional customer satisfaction. *International Journal of Contemporary Hospitality Management*.
- Kotler, P., Keller, K.L. & Manceau, D. 2016. Marketing Management, 15e édition. *New Jersey: Pearson Education*.
- Lubis, C.P., Rosnelly, R., Roslina, R., Situmorang, Z. & Wanayumini, W. 2021. Penerapan Metode Naive Bayes dan C4. 5 Pada Penerimaan Pegawai di Universitas Potensi Utama. *CSRID (Computer Science Research and Its Development Journal)*, 12(1): 51–63.
- Mahmoud, M.A., Ahmad, M.S., Yusoff, M.Z.M. & Mustapha, A. 2015. Context identification of scientific papers via agent-based model for text mining (ABM-TM). *New Trends in Computational Collective Intelligence*. Springer, hal.51–61.
- Muhson, A. 2010. Pengembangan media pembelajaran berbasis teknologi informasi. *Jurnal Pendidikan Akuntansi Indonesia*, 8(2).
- Nisah, A. 2021. ANALISIS PERBANDINGAN PENDAPATAN ANTARA PEBISNIS ONLINE SHOP DENGAN UPAH MINIMUM KOTA (STUDI KASUS DI KOTA MAKASSAR). *Economics Bosowa*, 6(004): 38–50.
- Nooraeni, R., Safiruddin, A.B., Afifah, A.F., Agung, K.D. & Rosyad, N.N. 2020. Analisis Sentimen Publik terhadap Sistem Zonasi Sekolah Menggunakan Data Twitter dengan Metode Naive Bayes Classification. *Faktor Exacta*, 12(4): 315–322.
- Norwawi, N.M. 2020. Recognition decision-making model using temporal data mining technique. *Journal of Information and Communication Technology*, 4: 37–56.
- Octaviani, P.A., Wilandari, Y. & Ispriyanti, D. 2014. Penerapan Metode Klasifikasi Support Vector Machine (SVM) Pada Data Akreditasi Sekolah Dasar (SD) Di Kabupaten Magelang. *Jurnal Gaussian*, 3(4): 811–820.
- Rahutomo, F., Saputra, P.Y. & Fidyawan, M.A. 2018. Implementasi Twitter Sentiment Analysis Untuk Review Film Menggunakan Algoritma Support Vector Machine. *Jurnal Informatika Polinema*, 4(2): 93.
- Rasheed, M.M., Faieq, A.K. & Hashim, A.A. 2020. Android Botnet Detection Using Machine Learning. *Ingénierie des Systèmes d'Information*, 25(1).
- Sarifudin, S. & Amelia, R.R. 2020. E-COMMERCE PADA ELLISA BAGS. *JURNAL SINKOM Sistem Informasi, Informatika dan Komputer*, 1(1): 26–34.
- Sitepu, A.C., Wanayumini, W. & Situmorang, Z. 2021. Analisis Kinerja Support Vector Machine dalam Mengidentifikasi Komentar Perundungan pada Jejaring Sosial. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 5(2): 475–484.
- Supradono, B. & Hanum, A.N. 2011. Peran Sosial Media Untuk Manajemen Hubungan dengan Pelanggan Pada Layanan E-Commerce. *Value Added Majalah Ekonomi dan Bisnis*, 7(2).
- Vimala, S. & Sharmili, K.C. 2018. Prediction of loan risk using naive bayes and support vector machine. *Int Conf Adv Comput Technol (ICACT)*. hal.110–113.
- Wu, X. & Kumar, V. 2009. *The top ten algorithms in data mining*. CRC press.