An Integrated Approach for Sentiment Analysis and Topic Modeling of a Digital Bank in Indonesia using Naïve Bayes and Latent Dirichlet Allocation Algorithms on Social Media Data

by Johan Setiawan

Submission date: 03-Feb-2026 05:06PM (UTC+0700)
Submission ID: 2870114555
File name: oach-sentiment-and-topic-modelling-digital-banking-Indonesia.pdf (397.44K)
Word count: 5305
Character count: 29004

An Integrated Approach for Sentiment Analysis and Topic Modeling of a Digital Bank in Indonesia using Naïve Bayes and Latent Dirichlet Allocation Algorithms on Social Media Data

Johan Setiawan
Information Systems Study
Universitas Multimedia Nusantara
Tangerang, Indonesia
johan@umn.ac.id

Anastasia Milenia
Information Systems Study
Universitas Multimedia Nusantara
Tangerang, Indonesia
anastasia.milenia@gmail.com

Ahmad Faza
Information Systems Study
Universitas Multimedia Nusantara
Tangerang, Indonesia
ahmad.faza@umn.ac.id

***Abstract*— Social media provides a public platform for expressing complaints and opinions. Researchers can use text mining techniques such as sentiment analysis and topic modeling on social media data to compare features and gauge public opinion on competing digital banks in Indonesia. The aim of this study is to classify sentiments and identify topics in social media data related to a specific digital bank. To accomplish this, the Naïve Bayes algorithm is used for sentiment analysis, while Latent Dirichlet Allocation is used for topic modeling. The social media data is sourced from Twitter and Instagram for Line Bank digital bank. The study finds that the Naïve Bayes algorithm performs well in classifying sentiments, achieving a maximum F1 score of 0.863. Positive sentiments are more prevalent in Twitter data, while negative sentiments are more prevalent on Instagram. Topic modeling using Latent Dirichlet Allocation algorithm identifies four optimal topics for positive sentiment and five for negative sentiment. The coherence value obtained is 0.426279 for positive sentiment and 0.397232 for negative sentiment.**

Keywords— Digital Bank, Latent Dirichlet Allocation (LDA), Naïve Bayes, Sentiment Analysis, Topic Modelling

I. INTRODUCTION

Over the past few years, there has been a noticeable shift from traditional banking transactions, such as those conducted through ATMs and branch offices, towards the utilization of mobile and internet banking services. Data from the Institute for Development Economy and Finance (INDEF) indicates a consistent decline in the frequency of ATM or debit transactions, from 62% in 2011 to 37% in 2018. Similarly, transactions at branch offices decreased from 17% in 2011 to 4% in 2018, while the frequency of mobile banking as an internet-based service increased from 6% in 2011 to 41% in 2018. Additionally, other internet-based services, such as internet banking, saw an increase from 11% in 2011 to 17% in 2018[1].

One of the internet-based banking services in Indonesia is digital banking. According to a survey conducted by Inventure Indonesia and Alvara Research Center, 43.6% of respondents used internet-based banking services more frequently after the COVID-19 pandemic compared to before[2]. Conversely, only 25.6% of respondents reported an increase in the use of e-wallet electronic payment instruments since the pandemic[2].

The growing number of Indonesians using internet-based banking services has led to an increase in the number of digital banks. Momentum Works reports that over the past five years, at least one digital bank product per year has started operating in Indonesia[3]. One such example is Line Bank, an application owned by KEB Hana Bank Indonesia and Line, a messaging application from Japan[4]. However, the rise in the number of digital banks has also been accompanied by an increase in consumer complaints, with financial services being the most frequently reported category of complaints over the past five years, according to data from the Consumers Association from Indonesia (YLKI)[5].

One of the complaints about financial services received by YLKI complaints related to financial services, is banking [6]. According to the Financial Services Authority (OJK) social media is a source of information that can be obtained quickly to find responses from the public, such as service complaints[7]. Based on this, sentiment analysis and topic modeling of social media data, particularly Twitter and Instagram, are needed, to better understand user comments and their implications. Indonesia's most widely used social media platforms, according to the 2020 report from We Are Social and HootSuite, respectively, are Youtube, WhatsApp, Facebook, Instagram, and Twitter[8]. As of July 2020, Indonesia is ranked sixth in terms of the number of Twitter users[9] and fourth in terms of Instagram users[10], according to data from Statista.

II. LITERATURE REVIEW

A. Sentiment Analysis

Sentiment analysis or opinion mining is a more profound analysis stage from text mining, which helps find opinions or tendencies of things in the text unit, such as positive, negative, or neutral emotions [11]. Sentiment analysis is one of the implementations of a popular classification method because it can find out a person's views, opinions, reactions, and emotions only based on the tone of speech from the written text [12]. The Naïve Bayes algorithm was chosen to perform the automatic sentiment classification from social media because it has proven to positively affect sentiment classification [13]–[16]. Some of these advantages include attributes that are not limited to numeric, speed in modeling until the deployment stage [17], modeling with a small amount

of training data, and better performance with fewer input dimensions [18], assuming all attributes are independent [17].

B. Topic Modelling

Topic modeling can summarize large text documents into a group of topics. The topic in question is a collection of words that often appear together in a document. Each word has its weight in a topic [19]. The process of getting topics from several documents is called topic modeling, based on the unsupervised concept [12]. The topic modeling task uses the Latent Dirichlet Allocation algorithm because it is proven to have advantages [20]–[22], such as the ability to process short, long, and mixed-length documents as well as being able to consider the form of relationships between documents [21].

C. Related Works

Several previous studies conducted sentiment analysis [23] and topic modeling [24] on digital banks and E-Commerce. Cheng and Sharmayne [23] predict sentiment classification and topic modeling based on review ratings of digital banks in the Philippines. Sentiment classification prediction is made without a special algorithm, while topic modeling uses LDA and association methods. This study resulted in seven main topics from application review data, keywords associated with each positive and negative label, and conclusions about the advantages and disadvantages of banks [23]. Yang [24] combines sentiment analysis and modeling by using LDA for topic modeling and a Python library called vader for sentiment analysis. The study compared five brands' product review data from e-commerce about healthy snacks ("new snacks"). The results of topic modeling almost all produce positive words for each store, then used for strategy recommendations for each store owner.

III. METHODOLOGIES

A. Data Collection

The source of research data is public opinion about one of the digital banks in Indonesia, which is published through social media. The social media data used are tweet data and comment data on official Instagram account content from Line Bank, a digital bank. The social media owned by Line Bank, which operates in Indonesia, are @linebankid for Instagram, Twitter, and Tiktok. While "Line Bank by Hana Bank" is the official account on Facebook and Youtube. The data is collected by scraping from social media using Python in Jupyter. The Python library used for scraping tweet data is Tweepy to access the Twitter API [25]. The Python Instagram-Comments- Scraper library collected comment data from Instagram [26]. Twitter and Instagram social media data took over three months, from November 6, 2021, to March 19, 2022. The period is determined based on the release date of the digital bank Line. Based on the provisions of the Twitter API, the period for scraping is also given a limit of seven days before the scraping date so that Twitter data will be retrieved periodically every week. The data collected within one week will then be examined by three students who have proficiency in Bahasa, whose results will determine the labelling of sentiment based on the majority vote results.

B. Research Design

The step of sentiment analysis research will follow a general procedure with some additional steps for modelling the topic of the digital bank. The step of sentiment analysis refers to the research conducted by Pierre [27] with modifications made. Modifications were carried out at the preprocessing stage, where previous studies used only one social media, while in this study, two social media were used. The following change is that the research step does not stop at the sentiment classification stage but continues with the topic modelling task. The stages of the topic modelling task use the reference of the research framework of Asmussen and Møller [28] with modification in the clean document step.

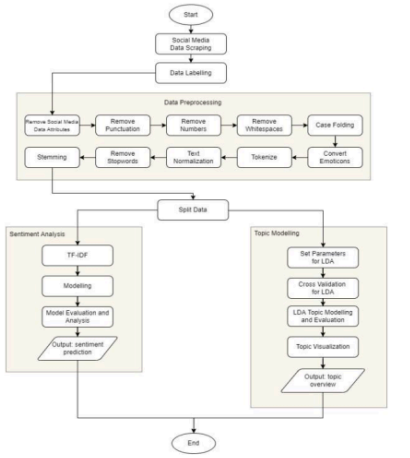

Fig. 1 Sentiment Analysis and Topic Modelling Steps

These stages are combined with the preprocessing step when conducting sentiment analysis for efficiency. Fig. 1. shows the results of improvements for this study.

IV. ANALYSIS AND RESULTS

A. Data Scrapping

Twitter data retrieval is done using the Tweepy library. Tweet data retrieval is carried out every week according to the limits of the Twitter API, with the keyword 'line bank'. The results of scraping tweet data are then stored in a spreadsheet file. The second social media data, Instagram comments data, were obtained using the Instagram-Comments-Scraper library. The results of scraping the comment data obtained are in the form of a spreadsheet file for each Instagram post. The data is combined into a single file consisting of username data and comment content. Based on the scraping process of the two social media, there were 4257 Twitter tweets and 1930 Instagram comments.

B. Data Labelling

Results of scraping Twitter data and subsequent Instagram comments manually labelled sentiment by three students who are digital bank users and have proficiency in Bahasa. Labelling is divided into data with positive sentiment (expressed by 'P') or negative (expressed by 'N'). Each data will then be determined as the final label based on

the most labels obtained. Tweets with positive labels were collected as many as 3894 and negative labels as many as 363. In contrast, positive comments from Instagram were obtained, as many as 276 comments and 1654 negative comments.

C. Data Attributes Punctuation Numbers and Whitespaces Removal

The first stage of data processing is to delete data attributes often found in social media data, such as links, hashtags, usernames, mentions, and the description of 'RT', which represents retweets. The deletion is done with the re.sub() library, which replaces word patterns based on regex with empty strings. Existing punctuation marks other than those contained in the unique attributes of social media data will then be deleted based on the list of punctuation marks in the string.punctuation. The script works the same way as the previous process, which replaces punctuation marks with empty strings. The process of deleting numbers is done by using the script "\d+" to find one or more digits which are then deleted.

D. Case Folding

Case folding is changing letters to non-capital to prevent calculating the same word but having a different capitalization [29]. The data resulting from the next step is then capitalized to be non-capital using the str.lower() function.

E. Convert Emoji

Changing the emoji is done by using an emot library that will convert the emoji into keywords that describe the emoji. Research only divides emoji pool changes into 'sad' and 'happy' words [29].

F. Tokenize

The tokenization stage is a step to break the existing sentences into smaller units. In the case of social media data, it will be broken down into word-for-word [30] using the string library's split() function.

G. Text Normalization

The normalization process works by mapping some words, such as informal words or abbreviations (slang), to become standardized formal word forms [31]. This study uses a dictionary according to the KBBI standard, which is then mapped according to the dictionary.

H. Remove Stop Words

Stop words are a collection of words that are very often found in a sentence and can make other more relevant words uncountable [29]. Some words included in stop words will be deleted using the nltk library in Bahasa to get more relevant data.

I. Stemming

Each existing word will be changed to its basic word based on conformity with the Indonesian language rules [29]. The word then changes using the StemmerFactory function from the literary library.

J. Split Data

Before starting the modelling process, all data that has gone through the preprocessing stage is separated into several parts. In the first step, data from each social media will be separated by 80%. Furthermore, the two data from social media will be combined into training and validation data. The second data split process will take 20% of each social media and then use it as testing data to test the model on the new data. The training and validation data collected was 3406 on Twitter and 1544 on Instagram. The first data split process also generates 20% of data for model testing on each social media, which are 851 on Twitter and 386 on Instagram. The second data split process divides the combined results of the social media data from the previous process by dividing 80% for training data and 20% for validation data. The split process uses the train_test_split function. The data training collected was 3960, and 990 data were collected for validation. The data separation is divided based on negative and positive sentiments for the topic modelling task. The combined data comes from the combined Twitter and Instagram data. There are 4170 data with positive sentiment and 2017 data with negative sentiment.

K. Create TF-IDF (Term Frequency-Inverse Document Frequency)

The Term Frequency-Inverse Document Frequency (TF-IDF) is a word weighting method from the dataset. TF-IDF process is required to convert text data into a word matrix and its weight so that the algorithm can process it [32]. Words that appear less frequently will get a higher score and are considered more relevant, and words that occur too often will be reduced because they are considered less relevant [33]. TF-IDF is divided into two processes: the calculation of the frequency of words in a document divided by the number of words in the document (TF). The IDF stage is when words are given weights based on the logarithmic calculation of the number of documents divided by the number of documents containing the word [34]. After the TF-IDF process, the classification model is also made using the pipeline. Parameters tested for modelling using MultinomialNB() are alpha of 0.01, 0.01, and 0.001.

L. Model Selection

The classification model is also made using the pipeline after the TF-IDF process. Parameters tested for modelling using MultinomialNB() are alpha of 0.01, 0.01, and 0.001. Determination of the best parameters for TF-IDF and Naïve Bayes modelling is done using the GridSearchCV function, with a total of five cross-validations, on the training data. The results of the parameter determination process show that the max_df (TF-IDF) parameter has a value of 0.75, ngram_range (TF-IDF) has a value of (1,2), and alpha (MultinomialNB) has a value of 0.1.

M. Model Evaluation and Analysis

In determining the performance of the classification model, the F1 score will be used to measure performance for an unbalanced amount of data per label [35]. The accuracy, precision, and recall values will be used to analyze its performance based on the confusion matrix. The sentiment classification process stops at this step after the entire process is repeated for each social media data. Table 1 shows the model's performance value on the validation data from the dataset. Fig. 2. show the results of the model's performance value.

Table 1 – Model Performance

Dataset	Precision	Recall	Accuracy	F1-Score
Two social media	0.839	0.841	0.841	0.837
Data testing (Twitter)	0.856	0.872	0.872	0.863
Data testing (Instagram)	0.852	0.733	0.733	0.775

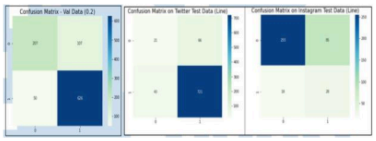

Fig. 2 Confusion Matrix on Validation Data and Test Data

In data validation, there are six hundred seventy-six data with positive sentiment and three hundred fourteen data with negative sentiment. Seven hundred sixty-four data with positive and eighty-seven with negative sentiments were produced from testing data with Twitter sources. In testing data with Instagram sources, forty-six data with positive and three hundred forty with negative sentiments. Those show that Twitter produces data with the most positive sentiment. In contrast, Instagram has data with the most negative sentiments.

N. Set Parameters for LDA

The LDA algorithm requires parameters before starting the modelling stage, namely the optimal number of topics to be generated. A more significant number of issues usually indicates a more complex topic, while a small number of topics indicates a more general topic [28]. In addition, the parameters prepared are data in the form of a dictionary, corpus, and bag of words that consider n-grams. The first step in topic modelling is to prepare data for the LDA algorithm, creating a dictionary, corpus, and bag of words. Function phrases from the gensim library are used to detect the presence of phrases that often appear but consist of more than one word; then converted into a single unit. Making the bag of words format is done with the doc2bow function and the Dictionary function to convert bigram data into a dictionary.

O. Cross Validation for LDA

The cross-validation stage was performed to ensure that the modelling performed with the LDA algorithm resulted in the optimal number of topics [28]. The process of validating the optimal number of topics will use a coherence value, where the level of similarity between words incorporated in one topic is described in this value [36]. The greater the coherence value, the better the resulting topic represents the document [37]. The highest coherence value is used from the number of topics to get the optimal number of topics for topic modelling. The coherence value is calculated using the CoherenceModel function with the coherence value parameter value 'c_v'. A cross- validation process was performed to get the coherence value from each number of topics. The process is repeated on the test on 75% of the data and 100% or all of the data. The topic range is determined from 2 topics to 10 topics [24]. The parameters and coherence values obtained will then be stored in the data frame for visualization purposes.

P. LDA Topic Modeling and Evaluation

Making a topic modelling model with LDA is done using the best parameters obtained from the previous stage, then implemented on social media data that has gone through the preprocessing stage. The topic will be determined based on the data per sentiment label so that the main topics on the data with positive and negative sentiments can be identified. This stage resulted in four topics with positive sentiments with a coherence value of 0.426279. However, on topics with negative sentiments, there are five topics with a coherence value of 0.397232.

Q. Topic Visualization

The results of the topic modelling process will be visualized by displaying a list of topics along with graphs using the Python pyLDAVIs library. The main display on the graph is an inter-topic distance map that can be used interactively or not statically [38]. If the cluster map visualization is focused on one of the topics, the bar chart will change according to the frequency of words in the selected topic. Visualization using pyLDAVis can add to understanding the resulting topic results [38]. The results of the topic modelling show a total of nine topics. There are four topics with positive sentiment and five topics with negative sentiment.

Four topics with positive sentiments are:

Topic 1:	36.7% of Line Bank's social media data discusses administration fees and free transfer fees.
Topic 2:	Integration with digital banking platforms or other payment methods covered 36.6% of social media data about Line Bank
Topic 3:	14.7% of social media data discusses Line Bank physical cards, which have different designs from the usual bank cards in a positive way. Based on the data obtained, the design of the Line Bank's card is considered funny.
Topic 4:	12.1% of the data discusses Line Bank features which are easy to understand and fast application performance.

Five topics with negative sentiments are:

Topic 1:	30% of social media data discusses obstacles when opening a Line Bank account.
Topic 2:	Old physical card creation and problems related to card status change is discussed in as much as 24.9% of the data.
Topic 3:	Promotional activities carried out by Line Bank often fail and are discussed in 21.3% of its social media data.
Topic 4:	as much as 14.2% of social media data with negative sentiments discusses obstacles and errors that occur when transactions.
Topic 5	9.7% of social media data discusses customer service performance that is considered flawed.

Table 2. Comparison of model evaluation results

Authors	Algorithms			
	Naïve Bayes		SVM	
	Accuracies	F1-Scores	Accuracies	F1-Scores
Kristiyanti et al.	0.94	0.94	0.755	0.89
Arora et al	0.7944	0.794	0.7933	0.7907
Milenia et al. (author)	0.872	0.863	-	-

Table 3. Comparison of data amount and preprocessing method

Authors	Accuracies	F1-Scores	Data Training		Data Pre-processing Method
			Positive	Negative	
Kristiyanti et al.	0.94	0.94	100	100	Tokenization, use 2-grams, remove punctuation and numbers, 10-fold cross-validation.
Arora et al	0.7944	0.794	500	500	Tokenization, remove HTML markups, convert number digits to 'NUMBER', and punctuations to 'SYMBOL', and remove less frequent words.
Milenia et al. (author)	0.872	0.863	764	87	Remove punctuations, numbers, whitespaces, case folding, stop words, convert emoji, tokenization, normalization, stemming, 5-fold cross-validation, use 1-3-grams.

Table 4. Comparison of coherence values with previous research

Authors	Coherence Values	Topics	Data Sources	Pre-processing Method
Qomariyah et al	0.1376	4	Twitter	Remove punctuations, stop words, case folding, stemming, tokenization
Cheng & Sharmayne	0.4308	7	Review from Playstore	Remove punctuations, stop words, case folding, and lemmatization.
Milenia et al. (author)	0.426279	4	Twitter and Instagram	Remove punctuations, numbers, whitespaces, case folding, stop words, convert emoji, tokenization, normalization, stemming

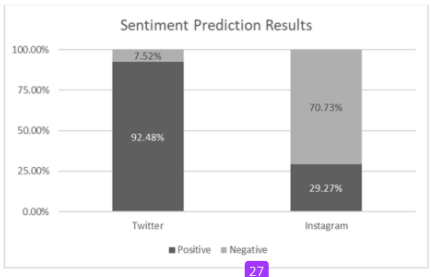

Fig. 3 Amount of Predicted Data based on Sentiment

V. DISCUSSION

Based on the evaluation results in Table 2, the value of the metrics selected for the model will be used to compare with the results of previous research on sentiment analysis [14][39]. Table 2 provides information on comparing the evaluation results of the model for the same task, namely sentiment classification of social media data, by comparing the Naïve Bayes algorithm and the Support Vector Machine (SVM). Both studies concluded that the Naïve Bayes algorithm is superior based on the accuracy value and F1 score. The findings align with this research result, where the Naïve Bayes algorithm also has a high accuracy value and F1 score.

This study resulted in a higher value than Arora et al. [39], with an accuracy value of 0.872 and the F1 score of 0.863. Both studies use a balanced amount of data for each sentiment label. However, this study uses unbalanced data on each sentiment. Then, the research also goes through different preprocessing stages so that it can affect the results of the model's performance. Table 3 compares the training data and preprocessing methods between the three studies [14], [39]. Fig. 3. shows the amount of data based on sentiment prediction. Data from Twitter has more positive sentiment tweets, with a percentage of 92.48%.

On the contrary, the data on Instagram has the highest number of negative sentiments, which is 36%. The positive things the digital bank has carried out can be seen from the results of the topic modelling task. The optimal number of topics has been evaluated, and a coherence value of 0.426279 is obtained on topics with positive sentiment. Table 4 compares coherence values with the research of Qomariyah et al. [20] and Cheng & Sharmayne [23]. This study has a higher coherence value than Qomariyah et al. [20]. However, this study has a lower coherence value than Cheng & Sharmayne study [23].

On the one hand, the results of topic modelling on the positive sentiment dataset show a discussion of the features provided by Line Bank, such as integration with external platforms and the application of free administration and transfer fees. Another topic also discusses cute physical card designs so that Line Bank's image becomes positive on social media. The ease of users' understanding and the speed of response from the features provided also get a positive response. On the other hand, topics with negative sentiments indicate the need for improvement in customer service in handling account openings, transaction settlement performance, and the availability of information regarding certain promotional activities.

VI. CONCLUSION

Sentiment analysis and topic modelling using social media data from a digital bank in Indonesia using the Naïve Bayes algorithm and Latent Dirichlet Allocation has been successfully carried out. On Twitter data prediction results, a lot of positive sentiment data is gathered that contains the general opinion of Twitter users regarding the digital bank. The results of the data prediction from Instagram are more of a negative sentiment because it is a place for digital bank customers who experience service problems or dissatisfaction. The data sentiment classification results show that the Naïve Bayes algorithm obtains a relatively good F1 score of 0.863. The topic modelling process using the Latent Dirichlet Allocation (LDA) algorithm receives the optimal number of topics of four topics in the dataset with positive sentiments and five topics with negative sentiments. The resulting coherence value is 0.426279 with positive sentiment and 0.397232 with negative sentiment.

Limitation

The present study is circumscribed by a limited scope, as its findings are solely derived from an investigation into one specific digital bank operating in the Indonesian context. Consequently, the generalizability of the results is inherently constrained, and they may not be applicable to other digital banking institutions or alternative geographic locations.

Moreover, this study employs unbalanced data on each sentiment, meaning that the sample sizes for positive, negative, and neutral sentiments are not of equal size. Consequently, caution must be exercised in interpreting the findings, as the skewed data may not accurately reflect the sentiments of the broader population.

Future Research

The present study offers an analysis of digital banks using a specific set of algorithms. However, it is important for future scholars to expand the scope of investigation by considering additional digital banks and analyzing alternative algorithms. By doing so, future research will contribute to a more comprehensive understanding of the digital banking industry.

ACKNOWLEDGMENT

We gratefully acknowledge the support from Universitas Multimedia Nusantara for this research.

REFERENCES

[1] H. Widowati, “Transaksi Digital Menggeser Peran Kantor Cabang dan ATM Bank,” 2019. [Online]. Available: https://databoks.katadata.co.id/datapublish/2019/08/08/transaksi-digitalmenggeser-peran-kantor-cabang-dan-atm-bank [Accessed: 01-Oct-2021].

[2] D. . Bayu, “Layanan Perbankan Digital Makin Sering Digunakan saat Pandemi,” 2020. [Online]. Available: https://databoks.katadata.co.id/datapublish/2020/11/18/layanan-perbankandigital-makin-sering-digunakan-saat-pandemi# [Accessed: 01-Oct-2021].

[3] Momentum Works, "Rise of Digital Banks in Indonesia," 2021. [Online]. Available: https://thelowdown.momentum.asia/country_sector/rise-of-digital-banksin-indonesia/?option=2021-indonesia&code=11502. [Accessed: 01-Oct-2021].

[4] Line Corporation, "LINE and PT Bank KEB Hana Indonesia Launch LINE Bank in Indonesia," 2021.

[5] R. Pahlevi, “YLKI Catat 535 Aduan Konsumen Sepanjang 2021 | Databoks,” 2022. [Online]. Available: https://databoks.katadata.co.id/datapublish/2022/01/10/ylki-catat-535-%0Aaduan-konsumen-sepanjang-2021 [Accessed: 05-Jul-2022].

[6] N. Sahara, “YLKI Terima Banyak Pengaduan Soal Lembaga Keuangan,” 2020.

[7] Otoritas Jasa Keuangan, “OJK Goverment PR Forum 2016,” 2016. .

[8] Hootsuite and We Are Social, "Digital 2020: Indonesia — DataReportal – Global Digital Insights," 2020.

[9] Statista, "Twitter: most users by country," 2021.

[10] Statista, "Instagram: users by country," 2021.

[11] D. Sarkar, R. Bali, and T. Sharma, Practical Machine Learning with Python. Apress, 2018.

[12] B. Benjamin, Applied Text Analysis with Python, vol. 53, no. 9. O'Reilly, 2018.

[13] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, "Sentiment Analysis of Social Media Twitter with Case of Anti-LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm," Procedia Comput. Sci, vol. 161, pp. 765–772, 2019.

[14] D. A. Kristiyanti, A. H. Umam, M. Wahyudi, R. Amin, and L. Marlinda, "Comparison of SVM & Naive Bayes Algorithm for Sentiment Analysis Toward West Java Governor Candidate Period 2018-2023 Based of Public Opinion on Twitter," 6th Int. Conf. Cyber IT Serv. Manag. (CITSM 2018), 2018.

[15] A. Nayak and S. Natarajan, "Comparative study of Naïve Bayes, Support Vector Machine and Random Forest Classifiers in Sentiment Analysis of Twitter feeds," Int. J. Adv. Stud. Comput. Sci. Eng., vol. 5, no. 1, 2016.

[16] A. E. Khedr, S. E. Salama, and N. Yaseen, "Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis," MECSJ. Intell. Syst. Appl., vol. 7, pp. 22–30, 2017.

[17] V. Kotu and B. Deshpande, Data Science Concept and Practice. Morgan Kaufmann Publishers, 2019.

[18] M. D. Devika, C. Sunitha, and A. Ganesh, "Sentiment Analysis: A Comparative Study on Different Approaches," Procedia Comput. Sci., vol. 87, pp. 44–49, 2016.

[19] EMC Education Services, Data Science & Big Data Analytics. Wiley, 2015.

[20] S. Qomariyah, N. Iriawan, and K. Fithriasari, "Topic modeling Twitter data using Latent Dirichlet Allocation and Latent Semantic Analysis," in AIP Conference Proceedings, 2019.

[21] R. Albalawi, T. H. Yeap, and M. Benyoucef, "Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis," Front. Artif. Intell., vol. 3, p. 42, 2020.

[22] E. S. Negara, D. Triadi, and R. Andryani, "Topic Modelling Twitter Data with Latent Dirichlet Allocation Method."

[23] L. C. Cheng and L. R. Sharmayne, "Analyzing Digital Banking Reviews Using Text Mining," IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min., 2020.

[24] Q. Yang, "Data Mining of New Snack E-commerce Reviews Based on Text Sentiment Analysis and Latent Dirichlet Allocation Topic Model," 2020.

[25] J. Roesslein, "API — tweepy 4.10.0 documentation," 2022. [Online]. Available: https://docs.tweepy.org/en/stable/api.html. [Accessed: 20-May-2022].

[26] A. Maulana, "Instagram-Comments-Scraper at feature-login," 2021. [Online]. Available: https://github.com/AgiMaulana/Instagram-Comments-Scraper/tree/featurelogin. [Accessed: 20-May-2022].

[27] J. Pierre, "Philippine Twitter Sentiments during Covid-19 Pandemic using Multinomial Naïve-Bayes," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 1, pp. 408–412, 2020.

[28] C. B. Asmussen and C. Møller, "Smart literature review: a practical topic modelling approach to exploratory literature review," J. Big Data, vol. 6, no. 93, pp. 1–18, 2019.

[29] S. Sfenrianto, E. R. Kaburuan, A. Bayhaqy, and K. Nainggolan, "Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes," Int. Conf. Orange Technol., 2018.

[30] A. A. Lutfi, A. E. Permanasari, and S. Fauziati, "Sentiment Analysis in the Sales Review of Indonesian Marketplace by Utilizing Support Vector Machine," J. Inf. Syst. Eng. Bus. Intell., vol. 4, no. 1, 2018.

[31] M. Bollmann, "A Large-Scale Comparison of Historical Text Normalization Systems," 2019. [Online]. Available: https://github.com/clarinsi/csmtiser. [Accessed: 01-Jun-2022].

[32] A. K. Singh and M. Shashi, "Vectorization of Text Documents for Identifying Unifiable News Articles," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 7, 2019.

[33] R. Ali and S. Qaiser, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," Artic. Int. J. Comput. Appl., vol. 181, no. 1, pp. 975–8887, 2018.

[34] R. Lourdusamy and S. Abraham, "A Survey on Text Pre-processing Techniques and Tools," Int. J. Comput. Sci. Eng., vol. 6, no. 3, pp. 148–157, 2018.

[35] A. Priya, S. Garg, and N. P. Tigga, "Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms," Procedia Comput. Sci., vol. 167, pp. 1258–1267, 2020.

[36] N. A. Tresnasari, T. B. Adji, and A. E. Permanasari, "Social-Child-Case Document Clustering based on Topic Modeling using Latent Dirichlet Allocation," Indones. J. Comput. Cybern. Syst., vol. 14, no. 2, p. 179, 2020.

[37] S. George and S. Vasudevan, "Comparison of LDA and NMF Topic Modeling Techniques for Restaurant Reviews," Indian J. Nat. Sci., vol. 10, no. 62, pp. 28210–28216, 2020.

[38] A. F. Hidayatullah and M. R. Ma'Arif, "Road traffic topic modeling on Twitter using latent dirichlet allocation," in International Conference on Sustainable Information Engineering and Technology, 2017, vol. 2018.

[39] A. Arora, P. Patel, S. Shaikh, and A. Hatekar, "Support Vector Machine versus Naive Bayes Classifier: A Juxtaposition of Two Machine Learning Algorithms for Sentiment Analysis," Int. Res. J. Eng. Technol., 2020.

An Integrated Approach for Sentiment Analysis and Topic Modeling of a Digital Bank in Indonesia using Naïve Bayes and Latent Dirichlet Allocation Algorithms on Social Media Data

ORIGINALITY REPORT

11%	8%	8%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to National College of Ireland Student Paper	1%
2	arxiv.org Internet Source	1%
3	www.proceedings.com Internet Source	1%
4	Johan Setiawan, Risanti Galuh Alamsari. "Prediction of Work From Home Post COVID-19 using Classification Model", 2022 Seventh International Conference on Informatics and Computing (ICIC), 2022 Publication	1%
5	Submitted to Murdoch University Student Paper	1%
6	download.bibis.ir Internet Source	1%
7	journal.uinjkt.ac.id Internet Source	<1%
8	jurnal.unimus.ac.id Internet Source	<1%
9	pdfs.semanticscholar.org Internet Source	<1%
10	ioinformatic.org Internet Source	<1%

11	"Pattern Recognition. ICPR International Workshops and Challenges", Springer Science and Business Media LLC, 2021 Publication	<1%
12	Puji Rahayu, Alfredo Julian Saputra, Erika Nurmaida, Brian DM Hutasoit, Kraugusteeliana Kraugusteeliana. "Application of Latent Dirichlet Allocation (LDA) to Identify Research Topics in Journal Publications", 2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS), 2023 Publication	<1%
13	Submitted to Wright State University Student Paper	<1%
14	eprints.umpo.ac.id Internet Source	<1%
15	K. V. Sambasivarao, Anasuya Sesha Roopa Devi Bhima. "Artificial Intelligence, Computational Intelligence, and Inclusive Technologies - Proceedings of International Conference on Artificial Intelligence, Computational Intelligence, and Inclusive Technologies (ICRAIC2IT – 2025)", CRC Press, 2026 Publication	<1%
16	www.researchgate.net Internet Source	<1%
17	Aditya Yulianto, S. Kom, Maria Irmina Prasetiyowati. "BoxLock: Mobile-based Serpent cryptographic algorithm and One-Time Password mechanism implementation for Dropbox files security", 8th International Conference for Internet Technology and Secured Transactions (ICITST-2013), 2013 Publication	<1%

18	archive.org Internet Source	<1%
19	publications.waset.org Internet Source	<1%
20	Nurajijah, Fachri Amsury, Irwansyah Saputra, Frieyadie, Daning Nur Sulistyowati, Bakhtiar Rifai. "Approval of Sharia Cooperative Customer Financing Using PSO-Based SVM Classification Algorithm", Journal of Physics: Conference Series, 2020 Publication	<1%
21	da Silva Vieira, Joana. "A Case Study on the Effect of News on Crude Oil Price", Universidade NOVA de Lisboa (Portugal), 2024 Publication	<1%
22	mail.joiv.org Internet Source	<1%
23	Khandaker Tayef Shahriar, Iqbal H. Sarker. "Exploring a Hybrid Deep Learning Framework to Automatically Discover Topic and Sentiment in COVID-19 Tweets", Computing&AI Connect, 2025 Publication	<1%
24	Leon Y. Xiao. "Lex Loot Boxes: The Regulation of Gambling-like Products in Video Games", Thesis Commons, 2024 Publication	<1%
25	Priyanka Pal, M. Venkateswarlu, Satish Kumar. "Unraveling the Impact of News Analytics on Financial Decisions: A Structured Review", Journal of Economic Surveys, 2025 Publication	<1%
26	Pulung Hendro Prastyo, Igi Ardiyanto, Risanuri Hidayat. "Indonesian Sentiment	<1%

	Analysis: An Experimental Study of Four Kernel Functions on SVM Algorithm with TF-IDF", 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), 2020 Publication	
27	Yessy Asri, Muhamad Fajri. "Sentiment Analysis of PLN Mobile Review Data Using Lexicon Vader and Naive Bayes Classification", 2023 International Conference on Networking, Electrical Engineering, Computer Science, and Technology (IConNECT), 2023 Publication	<1 %
28	Yiming Zhang, Ke Chen, Ying Weng, Zhuo Chen, Juntao Zhang, Richard Hubbard. "An intelligent early warning system of analyzing Twitter data using machine learning on COVID-19 surveillance in the US", Expert Systems with Applications, 2022 Publication	<1 %
29	joiv.org Internet Source	<1 %
30	journal.ugm.ac.id Internet Source	<1 %
31	jurnal.fikom.umi.ac.id Internet Source	<1 %
32	jurnal.iaii.or.id Internet Source	<1 %
33	openaccess.altinbas.edu.tr Internet Source	<1 %
34	radjapublika.com Internet Source	<1 %
35	salford-repository.worktribe.com Internet Source	

		<1%
36	www.inacl.id Internet Source	<1%
37	ejournal.nusamandiri.ac.id Internet Source	<1%
38	kinetik.umm.ac.id Internet Source	<1%

Exclude quotes	On	Exclude matches	< 7 words
Exclude bibliography	On		

An Integrated Approach for Sentiment Analysis and Topic Modeling of a Digital Bank in Indonesia using Naïve Bayes and Latent Dirichlet Allocation Algorithms on Social Media Data

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7