International Journal on Advanced Science Engineering Information Technology

Vol.12 (2022) No. 5
ISSN: 2088-5334

Classification of Indonesian Population's Level Happiness on Twitter Data Using N-Gram, Naïve Bayes, and Big Data Technology

I Nyoman Krisna Bayu [a,*], I Made Agus Dwi Suarjaya [a], Putu Wira Buana [a]

[a] *Department of Information Technology, Faculty of Engineering, Udayana University, Bukit Jimbaran, Badung, 80362, Indonesia*
*Corresponding author: *krisnabayu1999@gmail.com*

***Abstract*—The level of happiness is one factor that influences social interaction in the community. Therefore, the population's happiness level within the current year has become an exciting concern to be studied. Since last year, the world has been facing a COVID-19 pandemic. COVID-19 pandemic dramatically affects the happiness level of the population from a social, economic, health, education, and tourism perspective. The various affected sectors cause different levels of emotional happiness in the community in terms of social interactions in opinions and issues on social media. In addition, the number of issues on social media induce a vast data warehouse and high complexity. Big Data is a science that handles large amounts of data, which is unmanageable using traditional data processing methods or techniques. Various companies, organizations, researchers, and academics practice Big Data to extract and analyze the necessary information. Big Data is a general term used for all data collection forms of vast and complex nature. The utilization of Big Data can be valuable for a better decision-making process. This study uses Big Data Technology to evaluate the Indonesian population's happiness level on Twitter data. Method classified and technique using the N-Gram, Naïve Bayes, and Laplacian Smoothing Technique. The emotion in this research is classified into two aspects: happy and unhappy emotions. A total of 4.306.581 tweet data is classified; the obtained results revealed 39,4% happy emotion and 60,6% unhappy emotion.**

***Keywords*— Happiness level; tweet; big data; n-gram; naïve bayes; laplacian smoothing.**

Manuscript received 3 Feb. 2021; revised 21 May 2021; accepted 29 Jun. 2021. Date of publication 31 Oct. 2022.

I. INTRODUCTION

Happiness is an emotion that can reflect the value of self-satisfaction [1]. Circumstances, actions, or speech can influence the level of happiness. The influence level of happiness certainly affects social interactions in society [2]. Differences in happiness levels can occur in situations like the COVID-19 pandemic. The health threats caused by the COVID-19 pandemic lead to feeling anxiety and loneliness [3]. Feelings of anxiety and loneliness will certainly significantly affect the level of happiness. Lockdown regulations in dealing with COVID-19 show that lockdown regulations cause a decrease in the level of happiness [4]. The COVID-19 pandemic significantly impacts various sectors, including the economy, society, health, education, and tourism. Of course, the various affected sectors cause different levels of emotional happiness in the community in terms of social interactions in, opinions, and issues on social media. In addition, the number of issues on social media induce a vast data warehouse and high complexity, so a method or technology is needed to process this to be valuable.

Big Data is a science that handles large amounts of data unmanageable using traditional data processing methods or techniques [5]. Various companies, organizations, researchers, and academics practice Big Data to extract and analyze data to obtain important information from the data [6]. Big Data is a general term used for all data collection forms in a vast and complex nature. Its complexity makes it challenging to handle standard database management or traditional data processing applications. Nevertheless, Big Data can be valuable for a better decision-making process; recently, it has attracted interest from academics and practitioners. Big Data Analytics is a solution for almost every organization to help predict the customers' purchase behavior patterns, detect fraud and abuse [5]. Big Data Analytics is increasingly becoming the latest practice adopted by many organizations to obtain valuable information [6]. Big Data Analytics aims to reveal hidden patterns, incomprehensible relationships and resolve improved decisions [5]. Volume, variety, velocity, and veracity are characteristic of Big Data. Volume means that the processed data is often overwhelming and tends to increase, making it impossible to manage traditional database management. Variety means a varied set of data. Velocity means a very high growth of data and tends to be real-time. Veracity means the accuracy of the data. Big Data was a concept that refers to the inability of traditional data architectures to handle new data sets efficiently. Volume, variety, velocity, and veracity make analytics challenging for traditional data warehouses [7]. One of the implementations of Big Data Analytics is emotion detection. Emotion detection is language processing and an automatic text extraction to determine a person's emotions from textual data such as tweets. Tweet posts are brief texts, and their content varies widely based on user interests. Due to its complexity, Twitter has gained users, organizations, and research scientist's attention in various disciplines [8].

Several studies have conducted emotion detection using tweet data because the text is short and informal, which is then considered a challenging problem. The Tweepy API is an API provided by Twitter for developers to build software integrated with Twitter, as one example of helping an application for companies analyze data from customer responses on Twitter social media. The data provided by the Tweepy API is public. The Tweepy API accessed by registering as a developer account at developer.twitter.com. After registering an account as a Twitter developer. Next, create apps at apps.twitter.com to access the Tweepy API. Then registering the apps on Twitter, the developer account will receive the API access consist of consumer key and secret consumer key. Last, Configure access-token and access-token-secret to access data on Twitter social media.

The previous research related to Big Data Analytics and Sentiment Analysis on the UK-EU Brexit negotiations on Twitter users shows that Sentiment Analysis has a significant contribution to the government's decision-making. Moreover, Big Data Analytics and Sentiment Analysis are proposed as international negotiation tools to fix information gaps between decision-makers [9]. Other research also shows that Big Data Analytics has an essential role in measuring positive and negative sentiment toward Amazon.com's electronic products with competing products; in the implementation of using a suitable algorithm, we can find out the best-selling products based on ratings and by analyzing sentiment on the given reviews by users with an accuracy of 95.16% [10]. Implementing the classification method in sentiment analysis research on GSM service using the Multinomial Naive Bayes method yield the highest accuracy of 73.15% using the entire dataset of 1665 features [11]. Implementing the Naive Bayes method also analyses mobile service providers' performance research to determine profits, increase customer satisfaction, and business growth [12]. Research on implementing the N-gram and Naive Bayes methods to detect Indonesian e-mail spam yields a high accuracy of 94% [13].

The research in this paper aims to evaluate the Indonesian population's level of happiness during the COVID-19 pandemic based on Twitter data in Indonesian and Indonesian territory. The methods and techniques used in this research are the N-Gram method, Naïve Bayes method, Laplacian Smoothing Technique, and Big Data Technology classified into happy emotions and unhappy emotions with visualized data on Indonesian province territory and percentage difference in the form of a line.

II. Materials and Method

The material and method section describes overall research starting from data collection using the API Twepy, data storage in the MongoDB database, data pre-processing, and data processing up to the data visualization process. The pre-processing data phase comprises several processes; the cleaning process aims to remove the unused attributes, like punctuations and even URL links. The lowering case process aims to lower-case the words within the text, then continues to the emoticon convert process. The emoticon conversion process aims to convert emoticons into text. The stemming process aims to eliminate prefixes and suffixes in words so that the resulting output is in the form of basic words. The library used in the stemming process is the Sastrawi library. After going through the stemming process, it will enter the stop words process. The stop words process is used to remove words that are not necessary within the text.

The data processing stage uses 2 data consisting of training and testing data. Training data is the reference data used in the analysis using CSV format, and testing data is tweet data in JSON (JavaScript Object Notation) stored in the MongoDB database. The processing stage uses the N-Gram method, Naïve Bayes, and the Laplacian Smoothing Technique by calculating the probability of each class in the training data. The data visualization stage in the emotion's classification throughout the study used Tableau tools to classify the emotion's data stored in the MongoDB database. The latitude and longitude data in the visualization of classification emotions data using the Tableau tool are essential to add visualized based on Indonesian territory. The visualization in this study is visualized based on the maps of the Indonesian territory to show the distribution of areas with the highest or lowest classification emotions, then visualized based on the percentage difference to determine the percentage of increase or decrease in classification emotions per-day—research overview diagram described in Fig 1.

Fig 1 describes the research overview consisting of data preparation, data classification, and data visualization. This research overview is the development of research A Survey on Cleaning Dirty Data Using Machine Learning Paradigm for Big Data Analytics [14], by adding the emoticon convert process on data pre-processing and implementing methods on data analysis using N-Gram method, Naive Bayes method, and Laplacian Smoothing technique. Data visualization using location-based on provinces in Indonesia and integrated MongoDB database with Tableau makes visualizing analytical data faster even with Big Data. The data structure used in this research is JSON format, which has good query retrieval speed and CPU usage with various data variations. JSON technology shows tremendous potential to become a key database technology for handling substantial yearly data increases [15].

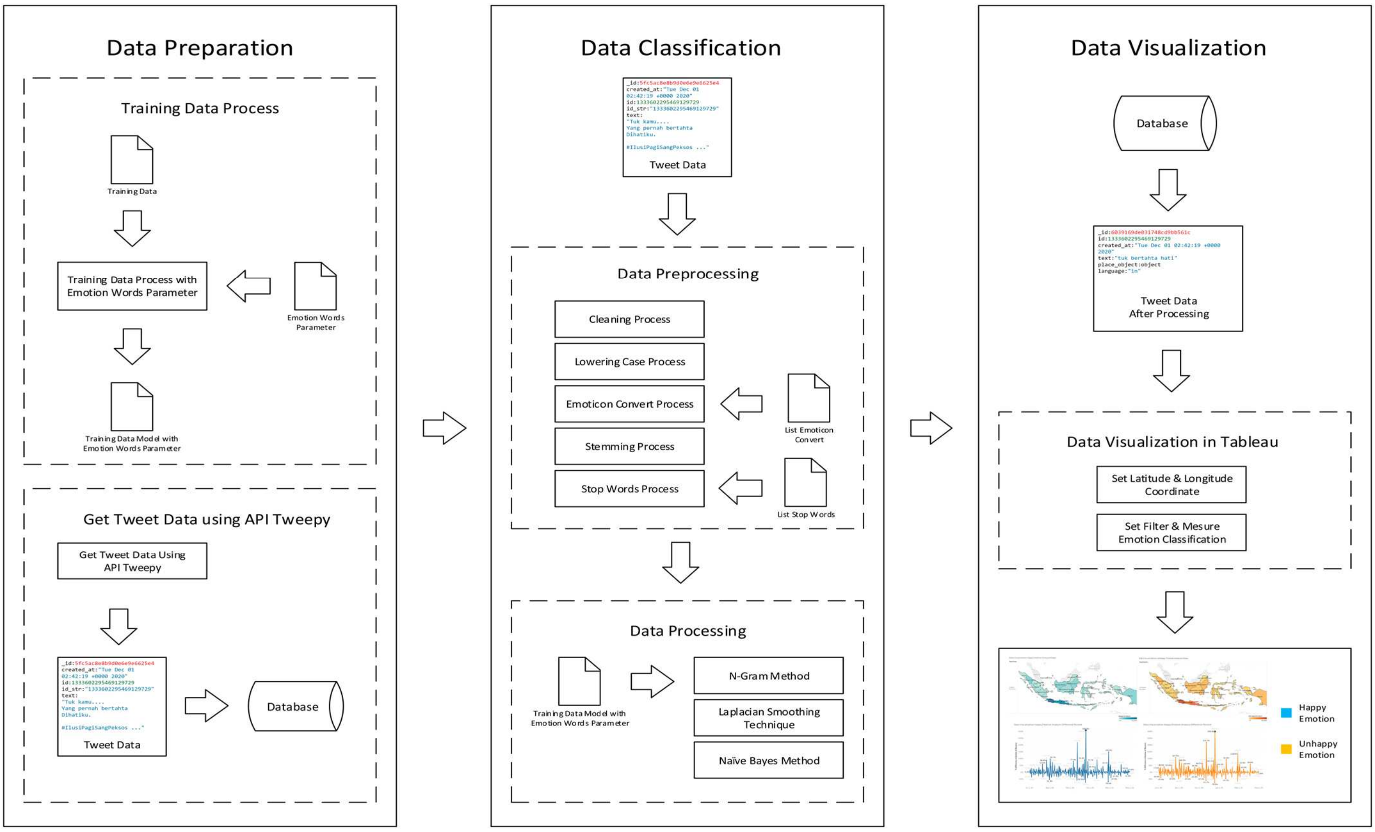

Data Preparation
Training Data Process
Training Data
Training Data Process with Emotion Words Parameter
Emotion Words Parameter
Training Data Model with Emotion Words Parameter
Get Tweet Data using API Tweepy
Get Tweet Data Using API Tweepy
_id:5fc5ac8e8b9d0e6e9e6625e4
created_at:"Tue Dec 01 02:42:19 +0000 2020"
id:1333602295469129729
id_str:"1333602295469129729"
text:
"Tuk kamu....
Yang pernah bertahta
Dihatiku.
#IlusiPagiSangPeksos ..."
Tweet Data
Database
Data Classification
Tweet Data
Data Preprocessing
Cleaning Process
Lowering Case Process
Emoticon Convert Process
Stemming Process
Stop Words Process
List Emoticon Convert
List Stop Words
Data Processing
Training Data Model with Emotion Words Parameter
N-Gram Method
Laplacian Smoothing Technique
Naïve Bayes Method
Data Visualization
Database
_id:6039169de031748cd9bb561c
id:1333602295469129729
created_at:"Tue Dec 01 02:42:19 +0000 2020"
text:"tuk bertahta hati"
place_object:object
language:"in"
Tweet Data After Processing
Data Visualization in Tableau
Set Latitude & Longitude Coordinate
Set Filter & Mesure Emotion Classification
Happy Emotion
Unhappy Emotion

Fig. 1 Research Overview

A. Data Preparation

Data preparation consists of training data, emotion word parameter data, and testing data. Training data is knowledge in the emotion classification analysis. The emotion classification word parameter serves as a parameter in identifying a class into happy emotions and unhappy emotions. Testing data is tweet data used in the classification analysis.

1) *Training data*: dataset used as knowledge in the classification analysis of emotions. The dataset used a sentiment analysis dataset developed based on emotion word parameters. The dataset used IndoNLU dataset: benchmarks and resources for evaluating Indonesian natural language understanding [16]. Testing data was stored in CSV format.

2) *Emotion words parameter*: data serves as a parameter in identifying a class into happy emotions and unhappy emotions. Emotion word parameters data used word parameters to classify positive and negative based on research on Twitter using Hybrid TF-IDF and Cosine Similarity [17], which developed into happy and unhappy emotions.

3) *Testing data*: The tweet data used in the emotion classification analysis was collected from Twitter using Tweepy API based on the Indonesian language and territory. Tweet data streamed using the Tweepy API will be stored in MongoDB database JSON (JavaScript Object Notation) format.

B. Data Pre-processing

Data Pre-processing describes the process before entering the processing stage. The pre-processing stage consists of several processes starting from the cleaning, lower-case, emoticon conversion, stemming, and stop words.

1) *Cleaning process*: remove unnecessary text attributes, users' usernames, and a URL link in the text. The cleaning process on emotion classification is needed because online text such as tweets contains uninformative attributes such as URL links.

2) *Lowering case process*: functions to lower-case the capital letters. The lowering case process is needed because the capital case is very influential in the analysis process. Therefore, it is necessary to equalize the case to the lower case to calculate word frequency using the N-Gram method.

3) *Emoticon conversion process*: converting emoticons into text because emoticons affect the classification of emotions. The emoticon or emoji data used is based on positive and negative Sentiment of emoji research [18], then developed into a happy and unhappy emoticon.

4) *Stemming process*: searching the main of each word by removing prefix and suffix affixes. An example of stemming data is to remove the affix from the word "*membaca*", which its prefix is "*mem*", and the main word is "*baca*". Utilizing the stemming process can solve problems in text mining [19].

5) *Stop words process*: eliminating words not considered necessary to speed up the text analysis process at the processing stage, and then after going through the pre-processing stage, the tweet data is stored in a configured MongoDB database JSON (JavaScript Object Notation) [20] document.

C. Data Processing

Data processing is a series of stages in the classification of emotion analysis using the N-Gram method, Naïve Bayes method, and the Laplacian Smoothing Technique.

1) N-Gram method: a method formed from the processing and calculation of words in a sentence. Unigram is a part of the N-Gram method, which processes single words in a sentence [21]. The N-Gram method converts each word in the text into a numeric variable.

TABLE I
BEFORE TURNING TEXT TO UNIGRAMS

No.	Text
1	good morning baby
2	morning sir
3	good mood today
4	good morning sir, good mood today

Table 1 describes the text before turning to Unigrams. Four texts were examples of turning text into unigrams, and the result of turning text into unigrams is described in Table 2.

TABLE II
AFTER TURNING TEXT TO UNIGRAMS

No.	Word in Text					
Text	good	morning	baby	mood	today	sir
1	1	1	1	0	0	0
2	0	1	0	0	0	1
3	1	0	0	1	1	0
4	2	1	0	1	1	0

Table 2 describes the implementation of a unigram in text and adds up each similar word in the text. Text mining with N-Gram variables is a study that implements the N-Gram method to determine similar words in the text and records how often they appear in the text [21]. The N-gram is a feature applied in textual content classification as a feature extraction method [22]. The implementation of the N-Gram method in this study is to find the frequency value of each word in a sentence, where the frequency is used in calculating the emotion classification using the Naïve Bayes method. The calculated frequency starts from the frequency of the emotion class in the training data and the chance of each word appearing in the training data.

2) Naïve Bayes: algorithm in classification with machine learning techniques [23]. One of the implementations of a supervised learning classifier is the Naive Bayes method [24]. The Naïve Bayes method is a probability method wherein this calculates the probability by using the frequency and value of the dataset. This study uses the Multinomial Naïve Bayes method.

$$\Pr(c|t_\mathrm{i}) = \frac{\Pr(c)\Pr(t_\mathrm{i}|c)}{\Pr(t_\mathrm{i})}, c \in C \qquad (1)$$

It is important to note that the Naïve Bayes theorem is that the classification process requires various clues to determine the class according to the analyzed sample. The Naive Bayes Multinomial Method will give a tweet tested for t_i class, which has the highest-grade probability Pr (c | t_i). Pr (c) is obtained by dividing the number of tweets belonging to c class by the total number of tweets. Pr (t_i | c) is the possible tweet of t_i class if known to c class. Pr (t_i | c) is obtained by calculating the multiplication between word probabilities in t_i's tweet.

$$\Pr(t_\mathrm{i}|c) = \alpha \prod_n \Pr(w_n|c)^{\mathrm{f_{ni,}}} \qquad (2)$$

$\mathrm{f_{ni}}$ is the number of n words in the tested tweet, t_i and Pr (w_n | c) are the probability of the word n if c class is known. Pr (w_n | c) is the number of words in the class, divided by the number of unique words in all tweets, next multiplied by the number of words in that class.

$$c = argmax \Pr(c \mid t_\mathrm{i}) \qquad (3)$$

The t_i tweet (ct_i) is obtained by comparing the respective probabilities and then look for the class with the highest probability.

3) Laplacian Smoothing: a Technique used refines the calculation if there is a zero value in the probability model. These zero values will make the Naïve Bayes Classifier unable to classify the input data. In this case, the Laplacian Smoothing method eliminates zeros in the probability model. Laplacian Smoothing is one of the methods used in the Naïve Bayes Classifier. The Laplacian Smoothing technique adds one to each data calculation in the training data set.

$$\Pr(w_n|c) = \frac{1 + F_{nc}}{N + \sum_{x=1}^{n} F_{XC}} \qquad (4)$$

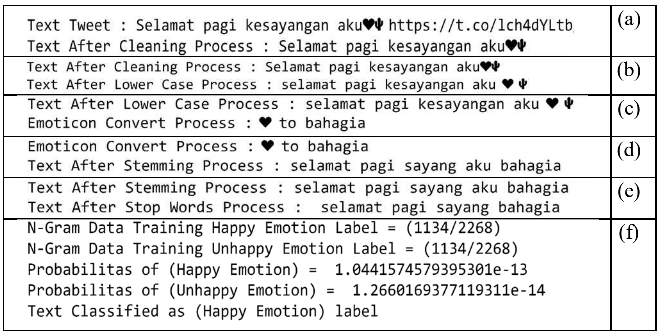
The Laplacian Smoothing is a process to eliminate the zero multiplication in Equation 2, one value-added to the number of wn words in class c. This technique is considered necessary because a word is in one class but not another.

D. Data Visualization

Data visualization is a stage in data visualization using Tableau. Tableau is software designed to instantly create data visualizations, reports, and dashboards. Moreover, it can connect multiple data sources, perform multidimensional data analysis, and create dashboards and reports. The data processing is efficient and has a friendly user interface [25]. The facilities provided by Tableau can connect to various data sources of many types, such as data systems arranged in file formats (CSV, JSON, XML, MS Excel), relational, non-relational data systems (PostgreSQL, MySQL, SQL Server, MongoDB), and cloud systems (AWS, Oracle Cloud, Google BigQuery, Microsoft Azure). The novelty of this study is that Tableau has a particular data blending feature. Another feature is the ability to collaborate in real-time, making it a valuable investment for commercial and non-commercial organizations. Data visualization starts with integrating the MongoDB database with Tableau, next by adding latitude and longitude to determine coordinate points, then selecting filter dimensions in emotion text, place names, and country codes.

III. Result and Discussion

The results and discussion describe result emotion classification of the Indonesian population's happiness level on Twitter data explaining data pre-processing begins cleaning process, lowering case process, emoticon convert process, stemming process, and stop words process. Implementation methods are beginning from the N-Gram method, Naïve Bayes, and the Laplacian Smoothing Technique in emotion classification and visualized by maps of the Indonesian region and line form. Emotion classifications use 4,306,581 tweet data filtered based on Indonesian language and Indonesian territory.



Text Tweet : Selamat pagi kesayangan aku♥🌵 https://t.co/lch4dYLtb Text After Cleaning Process : Selamat pagi kesayangan aku♥🌵	(a)
Text After Cleaning Process : Selamat pagi kesayangan aku♥🌵 Text After Lower Case Process : selamat pagi kesayangan aku ♥ 🌵	(b)
Text After Lower Case Process : selamat pagi kesayangan aku ♥ 🌵 Emoticon Convert Process : ♥ to bahagia	(c)
Emoticon Convert Process : ♥ to bahagia Text After Stemming Process : selamat pagi sayang aku bahagia	(d)
Text After Stemming Process : selamat pagi sayang aku bahagia Text After Stop Words Process : selamat pagi sayang bahagia	(e)
N-Gram Data Training Happy Emotion Label = (1134/2268) N-Gram Data Training Unhappy Emotion Label = (1134/2268) Probabilitas of (Happy Emotion) = 1.0441574579395301e-13 Probabilitas of (Unhappy Emotion) = 1.2660169377119311e-14 Text Classified as (Happy Emotion) label	(f)

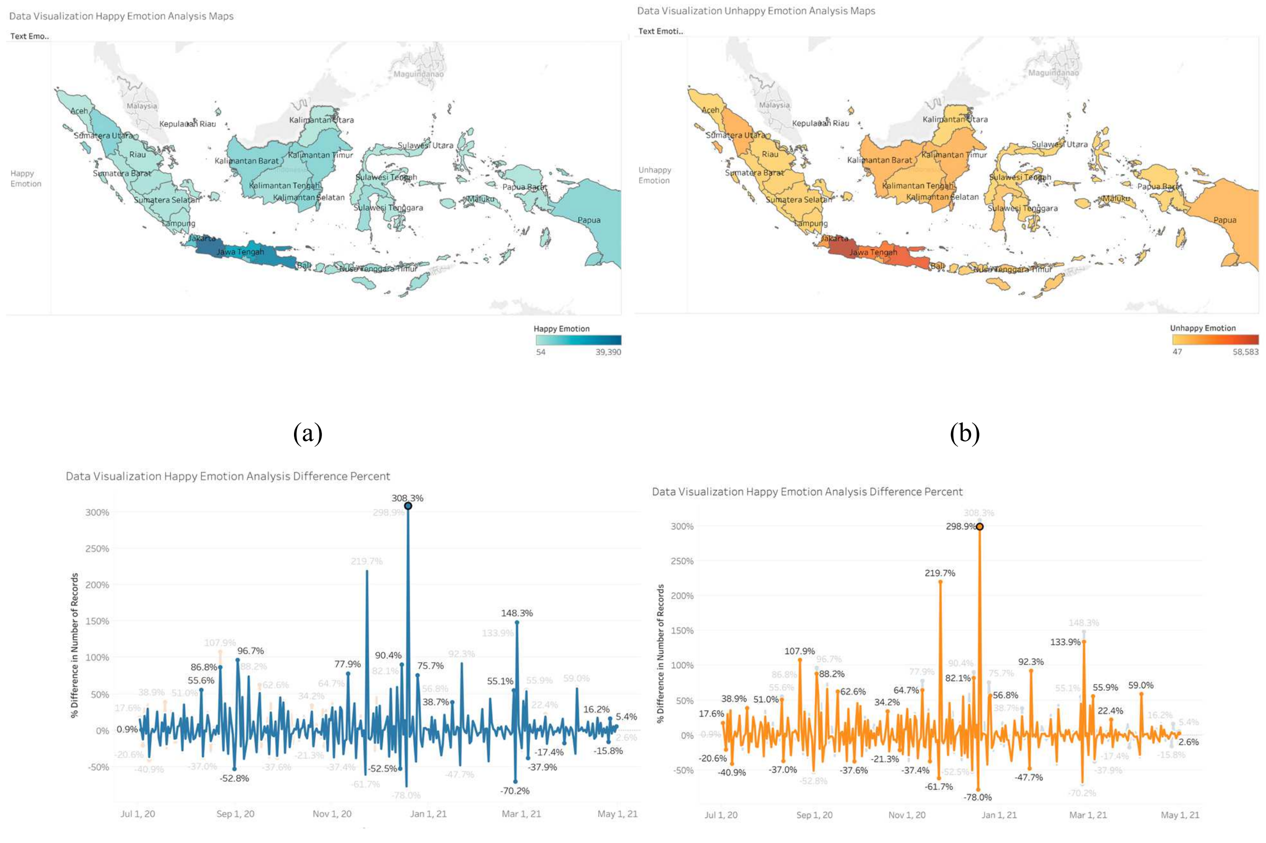
Fig. 2 Implementation Data Preprocessing and Data Processing, (a) Cleaning Process, (b) Lowering Case Process, (c) Emoticon Convert Process, (d) Stemming Process, (e) Stop Words Process, (f) Implementation Method

Fig 2 (a) describes the implementation cleaning process to remove usernames, hashtags, and URLs. Fig 2 (b) describes the lowering case process to change capital words in a tweet. Fig 2 (c) describes the implementation of converting emoticons into text. Fig 2 (d) describes the implementation stemming process to eliminate word prefixes and suffixes. Fig 2 (e) describes the implementation stop words process to eliminate words that are not considered necessary in tweets. Fig 2 (f) describes the research method's implementation. Firstly, group and count how many similar labels are in the training data using the N-Gram method. Next, give the default value using the Laplacian Smoothing technique in training data, then calculate the probability value using the Naïve Bayes method and look for the highest probability value.

TABLE III
EVALUATE TRAINING DATA MODEL

Train Data	Test Data	Accuracy	Precision	Recall	Fscore
55%	45%	89,5%	86,6%	89,5%	89,5%
60%	40%	89,6%	87,4%	89,6%	89,7%
65%	35%	89,7%	87,8%	89,7%	89,8%
70%	30%	89,8%	87,4%	89,8%	89,8
75%	25%	88,3%	86,8%	88,3%	88,4%
80%	20%	88,9%	87,1%	88,9%	89,08%
85%	15%	89,7%	88,8%	89,7%	90,1%
90%	10%	90,7%	90,6%	90,7%	91,06%

Table 3 describes the evaluation training data model using the N-Gram, Naïve Bayes, and Laplacian Smoothing Technique from 2.268 data. Evaluate scenario using 90% training data and 10% data testing yields a high accuracy of 90,7%, a precision of 90,6%, a recall of 90,7%, and a fscore of 91,06%.



(a)

(b)

(c)

(d)

Fig. 3 Data Visualization Emotion Classification, (a) Happy Emotion Classification Using Indonesian Maps, (b) Unhhapy Emotion Classification Using Indonesian Maps, (c) Happy Emotion Classification Using Difference Percent Graphic, (d) Unhappy Emotion Classification Using Difference Percent Graphic

Fig 3 (a) describes the classification of happy emotions from July 2020 to April 2021 was mostly founded in West Java province, with as many as 39.390. The discussed topic was 1.407 tweets discussing activities in Bandung city, 1.374 tweets discussing eating, 230 tweets discussing hotels, and 316 tweets discussing coffee. Fig 3 (b) describes the classification of unhappy emotions from July 2020 to April 2021 was mostly founded in West Java province, with as many as 58.583. The discussed topic 399 tweets discussed COVID, 279 tweets discussed appeal, 253 tweets discussed protocol, 46 tweets discussed corruption, and 219 tweets discussed bombs. Fig 3 (c) describes a significant increase in happy emotion tweets on December 19, 2020. The result of the significant increase in happy emotion was 308,3%, with 6.014 tweets. Topics discussed were 520 tweets discussed Indonesia shopping for TV shows from Tokopedia, 53 tweets discussed activities on the weekend, and 48 tweets discussed the activity of drinking coffee. There were 25 tweets discussed activities ahead of Christmas celebrations, 13 tweets discussed activities ahead of the New Year celebration, 25 tweets discussed holiday activities. There were 25 tweets discussed technology, two tweets discussed products on Tokopedia to free delivery Tokopedia promos, and seven tweets discussed Shopee product. Finally, there were eight tweets discussed discounts on products jackets and coffee ahead of Christmas celebrations, as well as ten tweets discussed promos on handicraft products, baby diapers, home, milk, and tourist destinations. Fig 3 (d) describes a significant increase in happy emotion tweets on December 19, 2020. The result of the significant increase in happy emotion was 298,9%, with 8.733 tweets. Topics discussed were 146 tweets discussed the law in Indonesia, 102 tweets discussed terrorism, 26 tweets discussed negative sentiment of swabs cost, 39 tweets discussed negative sentiment of rapid antigen cost, 37 tweets discussed COVID-19 vaccination. Meanwhile, there were nine tweets discussed economic conditions, five tweets discussed corruptors, five tweets discussed the impact of coronavirus, and six tweets discussed the impact of tourism object.

IV. Conclusion

This paper shows that the level of happiness emotions using 4.306.581 tweet data resulted in 39.4% happy emotions and 60.6% unhappy emotions based on Indonesian territory. Adding the emotion word parameter to training data improves the training data model, then adding the emoticon conversion process to the pre-processing data increases knowledge of emotion classification. The implemented method with a combination N-Gram method, Naive Bayes method, and Laplacian Smoothing Technique resulted in high accuracy of 90,7% using a 2.268 training data model, which is better than previous research [11].

References

[1] A. H. Goldman, “Happiness is an emotion,” *J. Ethics*, vol. 21, no. 1, pp. 1–16, Mar. 2017, DOI: 10.1007/s10892-016-9240-y.

[2] M. Holmes and J. McKenzie, “Relational happiness through recognition and redistribution: Emotion and inequality,” *Eur. J. Soc. Theory*, vol. 22, no. 4, pp. 439–457, Sep. 2018, DOI: 10.1177/1368431018799257.

[3] V. Cauberghe *et al.,* “How adolescents use social media to cope with feelings of loneliness and anxiety during covid-19 lockdown,” *Cyberpsychology, Behav. Soc. Netw.*, vol. 24, no. 4, pp. 250–257, Apr. 2021, DOI: 10.1089/cyber.2020.0478.

[4] T. Greyling, S. Rossouw, and T. Adhikari, “A tale of three countries: what is the relationship between covid-19, lockdown and happiness?,” *South African J. Econ.*, vol. 89, no. 1, pp. 25–43, Feb. 2021, DOI: 10.1111/saje.12284.

[5] M. A. Memon *et al.,* “Big data analytics and its applications,” *Ann. Emerg. Technol. Comput.*, vol. 1, no. 1, pp. 45–54, Oct. 2017, DOI: 10.33166/AETiC.2017.01.006.

[6] U. Sivarajah *et al.,* “Critical analysis of big data challenges and analytical methods,” *J. Bus. Res.*, vol. 70, pp. 263–286, Jan. 2017, DOI: 10.1016/j.jbusres.2016.08.001.

[7] S. A. El-Seoud *et al.,* “Big data and cloud computing: Trends and challenges,” *Int. J. Interact. Mob. Technol.*, vol. 11, no. 2, pp. 34–52, 2017, DOI: 10.3991/ijim.v11i2.6561.

[8] F. Atefeh and W. Khreich, “A survey of techniques for event detection in Twitter,” *Comput. Intell.*, vol. 31, no. 1, pp. 133–164, Feb. 2015, DOI: 10.1111/coin.12017.

[9] E. Georgiadou, S. Angelopoulos, and H. Drake, “Big data analytics and international negotiations: Sentiment analysis of Brexit negotiating outcomes,” *Int. J. Inf. Manage.*, vol. 51, no. Oct. 2019, p. 102048, 2020, DOI: 10.1016/j.ijinfomgt.2019.102048.

[10] M. Valera and Y. Patel, “A peculiar sentiment analysis advancement in big data,” in *Journal of Physics: Conference Series*, 2018, vol. 933, no. 1, DOI: 10.1088/1742-6596/933/1/012015.

[11] A. R. Susanti, T. Djatna, and W. A. Kusuma, “Twitter’s sentiment analysis on gsm services using multinomial naïve bayes,” *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 15, no. 3, pp. 1354–1361, Sep. 2017, DOI: 10.12928/TELKOMNIKA.v15i3.4284.

[12] M. A. Burhanuddin *et al.,* “Analysis of mobile service providers performance using naive bayes data mining technique,” *Int. J. Electr. Comput. Eng.*, vol. 8, no. 6, p. 5153, Dec. 2018, DOI: 10.11591/ijece.v8i6.pp5153-5161.

[13] Y. Vernanda, M. B. Kristanda, and S. Hansun, “Indonesian language e-mail spam detection using n-gram and naïve bayes algorithm,” *Bull. Electr. Eng. Informatics*, vol. 9, no. 5, pp. 2012–2019, Oct. 2020, DOI: 10.11591/eei.v9i5.2444.

[14] M. Z. H. Jesmeen *et al.,* “A survey on cleaning dirty data using machine learning paradigm for big data analytics,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 10, no. 3, pp. 1234–1243, Jun. 2018, DOI: 10.11591/ijeecs.v10.i3.pp1234-1243.

[15] M. K. Yusof and M. Man, “Efficiency of json for data retrieval in big data,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 7, no. 1, pp. 250–262, Jul. 2017, DOI: 10.11591/ijeecs.v7.i1.pp250-262.

[16] B. Wilie *et al.,* “IndoNLU: Benchmark and resources for evaluating indonesian natural language understanding,” *arXiv*, Oct. 2020.

[17] D. H. Wahid and A. SN, “Peringkasan sentimen esktraktif di twitter menggunakan hybrid tf-idf dan cosine similarity,” *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 10, no. 2, p. 207, Jul. 2016, DOI: 10.22146/ijccs.16625.

[18] P. K. Novak *et al.,* “Sentiment of emojis,” *PLoS One*, vol. 10, no. 12, pp. 1–22, Dec. 2015, DOI: 10.1371/journal.pone.0144296.

[19] P. Jennifer and A. Muthukumaravel, “A study on stopwords, stemming and text mining,” *Eur. J. Mol. Clin. Med.*, vol. 7, no. 6, pp. 1675–1682, 2020.

[20] T. Kudo, “A Proposal of transaction processing method for mongodb,” *Procedia Comput. Sci.*, vol. 96, pp. 801–810, 2016, DOI: 10.1016/j.procs.2016.08.251.

[21] M. Schonlau, N. Guenther, and I. Sucholutsky, “Text mining with n-gram variables,” *Stata J.*, vol. 17, no. 4, pp. 866–881, Jan. 2018, DOI: 10.1177/1536867X1801700406.

[22] D. Gamal *et al.,* “Implementation of machine learning algorithms in arabic sentiment analysis using n-gram features,” in *Procedia Computer Science*, 2018, vol. 154, pp. 332–340, DOI: 10.1016/j.procs.2019.06.048.

[23] J. Song *et al.,* “A novel classification approach based on naïve bayes for twitter sentiment analysis,” *KSII Trans. Internet Inf. Syst.*, vol. 11, no. 6, pp. 2996–3011, Jun. 2017, DOI: 10.3837/tiis.2017.06.011.

[24] V. D. Chaithra, “Hybrid approach: naive bayes and sentiment vader for analyzing sentiment of mobile unboxing video comments,” *Int. J. Electr. Comput. Eng.*, vol. 9, no. 5, pp. 4452–4459, Oct. 2019, DOI: 10.11591/ijece.v9i5.pp4452-459.

[25] S. Batt *et al.,* “Learning tableau: a data visualization tool,” *J. Econ. Educ.*, vol. 51, no. 3–4, pp. 317–328, Aug. 2020, DOI: 10.1080/00220485.2020.1804503.