JOURNAL LA MULTIAPP

VOL. 05, ISSUE 04 (380-397), 2024
DOI: 10.37899/journallamultiapp.v5i4.1465

Analysis of the Corpus with Naïve Bayes in Determining Sentiment Labeling

M. Arif Aulia[1], Muhammad Siddik Hasibuan[1]

[1]*Computer Science Study Program, State Islamic University of North Sumatra, Indonesia*

**Corresponding Author: M. Arif Aulia*
Email: m.arif0701201009@uinsu.ac.id

Article Info

Article history:
Received 2 July 2024
Received in revised form 18 July 2024
Accepted 7 August 2024

Keywords:
Corpus
Naïve Bayes
Pre-Processing
NLP

Abstract

The raw form of data is also an issue that creates a lot of problems while attempting to extract useful insight, thus requiring the use of NLP algorithms for text mining. This paper discusses sentiment analysis, with emphasis on user comments regarding cars on the microblog X that was formerly known as Twitter, work which employs Naïve Bayes Algorithm in text categorisation. The steps involved are the formation of the corpus and use of InsetLexicon dictionary for sentiment analysis with the help of weighted keywords and then going through pre-processing of the text data that includes cleaning, normalization and tokenization. The Naive Bayes algorithm estimates the probability of text under positive or negative sentiment class. The work shows that the "Comfortable" component of car reviews obtained the highest score in terms of recall, precision, and F1-score, which equals 0.83, 0.85, and 0.563, and the second set consists of 87 instances overall including an overall data set accuracy of 71%. The result validates the use of lexicon-based sentiment analysis in specific domain and at the same time exposes the weakness of the Naive Bayes, especially with complex word dependencies. Further studies should incorporate more advanced models and suitable dictionaries which facilitate sentiment analysis in ever-shifting online media settings.

Introduction

The rapid development of science in the current digital era is the ever- increasing text data that includes such as emails, documents and social media and the way to extract information from this unstructured text data is *text mining*. However, the complexity of the data is the main problem in producing deep understanding so that artificial intelligence assistance is needed, namely *Natural Languange Processing* (NLP) which is the key to the problem. This can allow us to process text more effectively and increase the opportunity to allow computer interaction with human language to become more natural (Prasetyo et al., 2021).

NLP can also be used to classify or label text with one of the text classification methods, namely sentiment analysis, which aims to identify opinions or feelings in text. Using various techniques such as *keyword* analysis, syntactic and semantic understanding of text can help identify the structure and sentiment contained in sentences and understand the meaning of words and phrases in a broader context in order to be able to interpret the sentiment more thoroughly. In addition to techniques, the approach taken in sentiment analysis is *supervised learning* where the model is trained using pre-labeled data (Ismail & Hakim, 2023)

This approach uses data that has been labeled with sentiments such as positive or negative and then modeled as training data to learn existing sentiment patterns The process involves

ISSN: 2716-3865 (Print), 2721-1290 (Online)

feature extraction in the form of key words, for example such as "good", "disappointed" or "beautiful" can be considered as relevant key words. And word frequency to convert text into a numerical representation that classification models can understand by counting how often a word appears in a text in order to be used as learning in the model (Isam & Abd Mutalib, 2019). However, the quality of the dataset used to determine success depends on the text corpus used. A quality text corpus can affect the performance and generalization of the classification model. In particular, the quality of sentiment labels affects the effectiveness of the classification model. However, if the labels are inconsistent, ambiguous and unrepresentative, it can lead to bias in the classification model, inaccurate and untenable results when applied to new data. So the quality of sentiment labels is an important factor in the formation of a good text corpus for sentiment analysis. Labeling sentiment as positive or negative currently uses the Naïve Bayes algorithm using chance and probability (Rachman & Handayani, 2021). This classification algorithm divides text into categories and then calculates the likelihood of the text appearing in a particular category by looking at how often they occur with the words used. However, its limitation in assuming every word in the text is unrelated to each other becomes one of its biggest drawbacks. Nonetheless, in various situations it can still provide satisfactory results so many researchers continue to use it especially if the text data is large enough (Khoirunisa, 2020).

In this study, an analysis will be carried out on user comments on social media X which will be summarized into a corpus so that the comments can be analyzed by labeling sentiments that are positive or negative based on certain aspects that will be given using the Naïve Bayes algorithm which uses opportunities and probabilities and then calculates them. Based on the description above, the researcher raised the title, namely "Analysis of the Corpus with Naïve Bayes in Determining Sentiment Labeling".

Text Mining is a process of interaction between a *user* and a set of documents with *analysis tools* that can provide solutions to problems by processing and analyzing large amounts of unstructured text. Various solutions can be used, one of which is *Natural Language Processing (*NLP). The purpose of using it in *text mining is* to group text (text *clustering)* and categorize text (*text categorization*) (Fitri et al., 2019). In this case, NLP is needed to extract information from text that was previously difficult or time consuming if done manually. Sentiment analysis is a field that encapsulates *text mining*, one of which is in natural language processing (NLP) which aims to analyze the sentiments, judgments and emotions of a person related to such as a particular organization, topic or activity. So sentiment analysis has the main goal of dividing a text in a document or sentence so that they can be categorized and determined whether they are positive or negative (Obiedat et al., 2021).

A corpus refers to a larger collection of texts. They are commonly used in the context of natural language processing (NLP) and computational linguistics. A corpus can contain a variety of text types, including news articles, books, academic documents, and more. A corpus is used to represent the variety of languages and contexts required in the development of NLP models (Sutabri et al., 2018). InsetLexicon is one example of a corpus. This dictionary can determine a sentence into positive and negative groups in sentiment comments or can be used as a dictionary used to find *sentiment scores.* The way the InSetLexicon corpus works is: (a) Polarity Marking, this stage analyzes the sentiment words in the sentence, then determines the polarity of the words; (b) Word frequency, this stage counts the number of occurrences of sentiment words; (c) Word Attitude, this stage determines the word attitude on each sentiment word, the value is +1 if the sentiment has a positive polarity, and -1 if it is negative; (d) Overall attitude, this step calculates the overall attitude with the following equation.

$$Overall\ Attitude = sikap\ kata * frekuensi\ kata$$

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Sentiment score, this stage calculates the sentiment score by summing up all the attitudes that have been calculated previously (A. Yani et al., 2019).

$$Sentiment\ score = \sum_{i=1}^{n} OverallAttitude\ (i)$$

Naïve Bayes Classifier is a classification algorithm that is often used in natural language processing (NLP) and text classification to predict a situation (Mayasari & Indarti, 2022). It works by predicting the probability of a condition occurring in the future based on current data (Nugroho & Religia, 2021). In text classification, probability is used in an example case for example to determine the possibility of spam or not spam based on the words that appear in the email. For example, if it contains the words "free", "offer", "special", "only" and so on, there is a possibility that the email is spam.

The following is the mathematical formula for the Naïve Bayes method:

$P(H\ |X) = P\ (X\ |H)P(H)$

$P(X)$

Ket:

$(H\ |\ X)$ = Probability of hypothesis H based on state X

X = Sample data with unknown label

H = Hypothesis that X is data with label

P(H) = Probability of hypothesis H

P(X) = Probability of X

Table 1. Example of Spam Email Dataset

Email	Spam	Keyword	UpperCase	Long
Email 1	0	“gratis”,”tawaran”	2	15
Email 2	1	“menang”,”hadiah”	5	20
Email 3	0	“berita”,”update”	1	10
Email 4	1	“diskon”,”jual”	3	18
Email 5	0	“halo”,”dunia”	0	12

Next is to calculate the posterior probability of each class against the feature for email 1:

Table 2. Posterior Probability of Each Class on Features for Email 1

Features	Spam	No Spam
Keywords.	1/3	0/2
Uppercase	1/3	0/2
Subject Length	1/3	0/2

The posterior probability of spam over all features for email 1 is : P(Spam | Keywords, Uppercase, Subject Length) = P(Spam) * P(Keywords | Spam) * P(Uppercase | Spam) * P(Subject Length | Spam) = 3/5 * 1/3 * 1/3 * 1/3 = 0.022

The posterior probability of not spamming against all features for email 1 is : P(No Spam | Keywords, Uppercase, Subject Length) = P(No Spam) * P(Keywords | No Spam) * P(Uppercase | No Spam) * P(Subject Length | No Spam) = 2/5 * 0/2 * 0/2 * 0/2 = 0.

A dataset is a collection of data that has been confirmed so that it can be used as a reliable source of data to be used in research. In this case, the *dataset* or what is taken is made itself into a corpus sourced from *crawling data* from the X / Twitter platform which is already in CSV format. Meanwhile, the weight data for machine learning is sourced from Github in

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Excel format (Firdaus et al., 2021). *Confusion Matrix* is a method to calculate accuracy with 4 *outputs*, namely *Recall*, *Precision*, *F1-score* and *Accuracy*. This concept evaluates the classification model based on the calculation of object trials that have correct or incorrect prediction values. This calculation process is implemented into a table with 2 *datasets* that are positive and negative: Jupyter Notebook is a data *science software* that consists of *data preprocessing* techniques, *machine learning* and even *natural language processing* (NLP). Jupyter Notebook provides an easy approach such as a flexible and responsive *user interface* so that users can compile and run code in various programming languages such as Python, R and others.

Methods

This study uses quantitative research methodology which is an approach in research that emphasizes the collection and analysis of numerical data to find causal and statistical relationships between variables. The research stages are in the following figure.

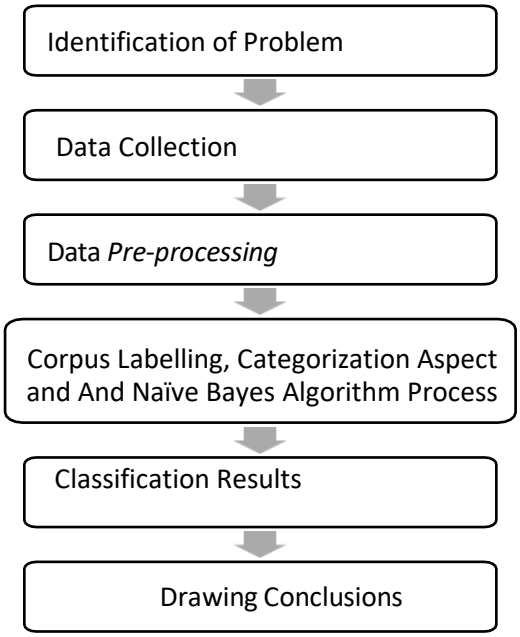

Figure 1. Research Stages

Data volume being one of the obvious drawbacks in text data collection and tagging further complicates the problem. Sentiment analysis models may not only be less accurate, but also prone to unwanted bias. Therefore, in this case, the researcher wants to use NLP as a solution to the problem with Naïve Bayes as the algorithm to analyze and see the influencing factors in the performance of the corpus in sentiment analysis.

The data for this research was obtained from X social media platform and then saved in CSV format to be used for further analysis and research. This technique uses Data Crawling (Goel et al., 2016). The population of this study is a collection of text from a corpus of 3728 datasets in CSV form. A sample is a portion or representative of the total population.. In this study, the samples used by researchers were around 464 datasets used as training data and 117 as test data (Dey et al., 2016).

Overall in this research, the *dataset* obtained in CSV format from the X social media platform is *crawled in the* form of a summary of public comments on car brands in Indonesia. After *processing, the* data is divided into a ratio of 80:20 where 80% is used for training data for model learning where the training data will be directly labeled each word into positive or negative labels. While the remaining 20% is used for unlabeled test data to test the model with model learning that has been done on the previous test data. Next, the sentiment labeling

ISSN: 2716-3865 (Print), 2721-1290 (Online)

process is done with the *InsetLexicon* dictionary to find the polarity value and group the comment data to be categorized based on aspects related to cars. The model will then be classified with the Naïve Bayes algorithm to determine the class. After that, the Multinomial Naïve Bayes classification model is used to calculate the probability of documents in the test data based on the probabilities in the training words and then tested for accuracy to determine the average value on aspects and the entire dataset.

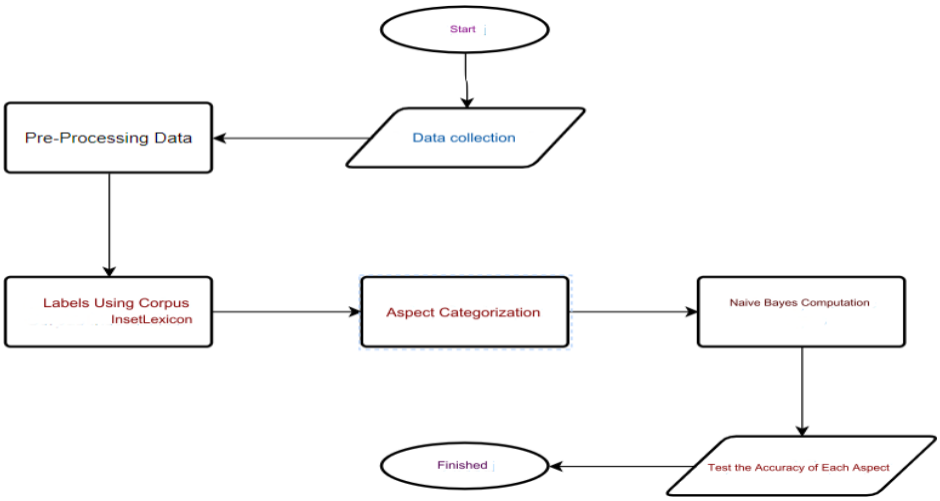

Figure 2. Research Flowchart

Results and Discussion

In this study, the author conducted a data collection process taken from application X (Twitter). Comment data is taken using the scrapping process. The data taken by the scrapping process will be stored in a .csv extension file. The data taken amounted to 3729 comment data, which is based on 4 types of cars, namely Daihatsu, Honda, Toyota and Suzuki. For example in Daihatsu using sigra, terios, and xenia cars. Then, Honda uses brio, civic, HR-V, and CR-V cars. Toyota cars use Avanza, fortuner, innova and raize. Meanwhile, Suzuki cars use ertiga, baleno, APV and ignis cars. The way to collect comment data is to tag all car names based on the type of car. The comments used were obtained from application X (Twitter) in June 2024 and accessed through jupyter notebook. Reviews are taken from X (Twitter) application user comment data with keywords of 4 types of cars with 15 car brands. Examples of reviews used are shown in the table below.

Table 3. All Comment Data

No.	Full_text
1	@akbartruck77688 @SelebtwitMobil I really agree! The choice of car is steady! Suzuki APV and Daihatsu Gran Max are already famous for their imbah and fit a lot. The Toyota Agya Base Model M/T is also economical and agile. Perfect for daily mobility.
2	@SuzukiIndonesia Ya im ba baleno dong
3	@innovacommunity If it's the strength of the body and frame, it's mending ertiga or xenia, right?
4	@aimr0d @ezash Kalo under 200 million KNP org2 on the non-take sigra with the highest spec yes.
5	@SelebtwitMobil Terios/Rush The seat is typical of cheap im bah bgt
3729	@SelebtwitFess Ertiga aja. Kalo doyan jadul ya kijang kapsul

The next step is to separate comments that contain spam and not spam. The way is to remove the aspects that are in the spam message Where the aspect in python is defined by "*keyword*".

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Comment data that is processed or analyzed is comment data that is not spam or not advertising/promotion. Non-spam comments are data submitted directly by X (Twitter) users.

Table 4. Non Spam Data

No.	Full_text
1.	@ravidetan @SelebtwitFess Don't livina. Especially if you are going to climb. Better number 6 (ertiga). Or just ignis. It's economical but you have to get used to the same taste of AGS. That's all
2.	@SelebtwitFess Ertiga is the most suitable to prepare a budget for foot snacks and make sure to replace the legs using SGP original spare parts, especially shockbreakers
3.	@koz_ns @ertigaid Ertiga is the same, it's just a slight difference
4.	@SelebtwitFess Beli R15 Dua viji biar jadi ertiga
5.	@SelebtwitFess The answer is easy. ERTIGA.
2984	Smg thr nya im ba HR-V aamiin

Figure 3. Non Spam Data With Using Aspect

Aims to clean the comment data that has been obtained. The components that are cleaned are components that are meaningless or irrelevant to the data classification process.

Table 5. Cleaning, Case Folding, Lower Case, and Remove HTTP in Text

No.	Full_text	Clean
1	@ravidetan @SelebtwitFess Don't livina. Especially if you are going to climb. Better number 6 (ertiga). Or just ignis. Economical im tapi harus terbiasa the same sense of AGS that is stupid. That's all	Don't livina, especially if you want to climb the number three better or ignis, it's economical, but you have to get used to the same taste of AGS which is so bad
2	@SelebtwitFess Klo butuh RWD Obviously, Terios Klo needs to be comfortable and the rhino is matic, Ertiga/XL7 Klo wants a safe selling price and can be used in the city, Avanza/Xenia	I need rwd obviously Terios klo needs to be comfortable and rhino matik, yes ertigaxl klo want a safe selling price and used in the city of doang, avanzaxenia
3	@efenerr Suzuki Baleno Hatchback Facelift Gagah dan berkelas	im ba baleno hatchback facelift gagah dan berkelas
4	@0tk0il im ba pasang sunroof aja udah next level	im ba Just install a sunroof is the next level
5	@MasMasBiassaa As a civic user, I don't agree with the word imbah. I'm a cool young person who tries to save money even though I fail wkwk	As a Civic user, I don't agree with the word IMBAH I am a cool young person who is trying to save money even though he fails to wkwk
2984	@admiredbysun Naksir banget juga sama HR-V	naksir banget juga sama hrv

Text can be broken down into specific words or tokens, such as “Avanza” (im ba), “sangat” (very), “cool” (cool), etc., and general terms that do not provide much value to the text. These terms can be broken down into specific words or tokens, such as “dan” (and), “or” (or), “di” (in), etc

Table 6. Tokenizing Process on Clean Data

No.	Clean	Tokens

ISSN: 2716-3865 (Print), 2721-1290 (Online)

1	Don't livina, especially if you want to climb the number three better or ignis, it's economical, but you have to get used to the same taste of AGS which is so bad	[‘jangan’, ‘livina’, ‘apalagi’, ‘kalo’, ‘buat’, ‘nanjak’, ‘better’, ‘nomor’, ‘ertiga’, ‘atau’, ‘ignis’, ‘aja’, ‘iritnya’, ‘polll’, ‘tapi’, ‘harus’, ‘terbiasa’, ‘sama’, ‘rasa’, ‘ags’, ‘nya’, ‘yang’, ‘nyebelin’, ‘sekian’]
2	I need rwd obviously Terios klo needs to be comfortable and rhino matik, yes ertigaxl klo want a safe selling price and used in the city of doang, avanzaxenia	[‘klo’, ‘butuh’, ‘rwd’, ‘jelas’, ‘terios’, ‘klo’, ‘butuh’, ‘nyaman’, ‘dan’, ‘badak’, ‘matiknya’, ‘ya’, ‘ertigaxl’, ‘klo’, ‘mau’, ‘harga’, ‘jual’, ‘aman’, ‘dan’, ‘dipake’, ‘di’, ‘kota’, ‘doang’, ‘ya’, ‘avanzaxenia’]
3	im ba baleno hatchback facelift gagah dan berkelas	[‘suzuki’, ‘baleno’, ‘hatchback’, ‘facelift’, ‘gagah’, ‘dan’, ‘berkelas’]
4	im ba pasang sunroof aja udah next level	[‘avanza’, ‘pasang’, ‘sunroof’, ‘aja’, ‘udah’, ‘next’, ‘level’]
5	As a Civic user, I don't agree with the word IMBAH I am a cool young person who is trying to save money even though he fails to wkwk	[‘sebagai’, ‘pengguna’, ‘civic’, ‘saya’, ‘merasa’, ‘tidak’, ‘setuju’, ‘dgn’, ‘kata’, ‘kata’, ‘si’, ‘mbah’, ‘saya’, ‘anak’, ‘muda’, ‘kece’, ‘yang’, ‘batikan’, ‘nyari’, ‘irit’, ‘walau’, ‘gagal’, ‘wkwk’]
2984	I really have a crush on HRV too	[‘naksir’, ‘banget’, ‘juga’, ‘sama’, ‘hrv’]

The text can be converted into good Indonesian according to the KBBI where there are word changes, such as “yg” to yang, “ga” to “tidak” and others. The research uses normalized.csv data accessed from github.

Table 7. Data Normalization

No.	Tokens	Normalisasi
1	[‘jangan’, ‘livina’, ‘apalagi’, ‘kalo’, ‘buat’, ‘nanjak', 'better', 'nomor', 'ertiga', 'atau', 'ignis', 'aja', 'iritnya', 'polll', 'tapi', 'harus', 'terbiasa', 'sama', 'rasa', 'ags', 'nya', 'yang', 'nyebelin', 'sekian']	['jangan', 'livina', 'apalagi', 'kalau', 'buat', 'nanjak', 'better', 'nomor', 'ertiga', 'atau', 'ignis', 'saja', 'iritnya', 'polll', 'tetapi', 'harus', 'terbiasa', 'sama', 'rasa', 'ags', 'nya', 'yang', 'nyebelin', 'sekian']
2	['klo', 'butuh', 'rwd', 'jelas', 'terios', 'klo', 'butuh', 'nyaman', 'dan', 'badak', 'matiknya', 'ya', 'ertigaxl', 'klo', 'mau', 'harga', 'jual', 'aman', 'dan', 'dipake', 'di', 'kota', 'doang', 'ya', 'avanzaxenia']	['kalau', 'butuh', 'rwd', 'jelas', 'terios', 'kalau', 'butuh', 'nyaman', 'dan', 'badak', 'matiknya', 'ya', 'ertigaxl', 'kalau', 'mau', 'harga', 'jual', 'aman', 'dan', 'dipake', 'di', 'kota', 'doang', 'ya', 'avanzaxenia']
3	['suzuki', 'baleno', 'hatchback', 'facelift', 'gagah', 'dan', 'berkelas']	['suzuki', 'baleno', 'hatchback', 'facelift', 'gagah', 'dan', 'berkelas']
4	['avanza', 'pasang', 'sunroof', 'aja', 'udah', 'next', 'level'	['avanza', 'pasang', 'sunroof', 'saja', 'sudah', 'next', 'level']
5	['sebagai', 'pengguna', 'civic', 'saya', 'merasa', 'tidak', 'setuju', 'dgn', 'kata', 'kata', 'si', 'mbah', 'saya', 'anak', 'muda', 'kece', 'yang', 'batikan', 'nyari', 'irit', 'walau', 'gagal', 'wkwk']	['sebagai', 'pengguna', 'civic', 'saya', 'merasa', 'tidak', 'setuju', 'dengan', 'kata', 'kata', 'si', 'mbah', 'saya', 'anak', 'muda', 'kece', 'yang', 'batikan', 'nyari', 'irit', 'walau', 'gagal', 'wkwk']
2984	['naksir', 'banget', 'juga', 'sama', 'hrv']	['naksir', 'banget', 'juga', 'sama', 'hrv']

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Table 8. Tokenizing Process on Clean Data

No.	Clean	Tokens
1	jangan livina apalagi kalo buat nanjak better nomor ertiga atau ignis aja iritnya im tapi harus terbiasa sama rasa ags nya yang nyebelin sekian	[‘jangan’, ‘livina’, ‘apalagi’, ‘kalo’, ‘buat’, ‘nanjak’, ‘better’, ‘nomor’, ‘ertiga’, ‘atau’, ‘ignis’, ‘aja’, ‘iritnya’, ‘polll’, ‘tapi’, ‘harus’, ‘terbiasa’, ‘sama’, ‘rasa’, ‘ags’, ‘nya’, ‘yang’, ‘nyebelin’, ‘sekian’]
2	klo butuh rwd jelas terios klo butuh nyaman dan badak matiknya ya ertigaxl klo mau harga jual aman dan dipake di kota doang ya avanzaxenia	[‘klo’, ‘butuh’, ‘rwd’, ‘jelas’, ‘terios’, ‘klo’, ‘butuh’, ‘nyaman’, ‘dan’, ‘badak’, ‘matiknya’, ‘ya’, ‘ertigaxl’, ‘klo’, ‘mau’, ‘harga’, ‘jual’, ‘aman’, ‘dan’, ‘dipake’, ‘di’, ‘kota’, ‘doang’, ‘ya’, ‘avanzaxenia’]
3	im ba baleno hatchback facelift gagah dan berkelas	[‘suzuki’, ‘baleno’, ‘hatchback’, ‘facelift’, ‘gagah’, ‘dan’, ‘berkelas’]
4	im ba pasang sunroof aja udah next level	[‘avanza’, ‘pasang’, ‘sunroof’, ‘aja’, ‘udah’, ‘next’, ‘level’]
5	sebagai pengguna civic saya merasa tidak setuju dgn kata kata im bah saya anak muda kece yang batikan nyari irit walau gagal wkwk	[‘sebagai’, ‘pengguna’, ‘civic’, ‘saya’, ‘merasa’, ‘tidak’, ‘setuju’, ‘dgn’, ‘kata’, ‘kata’, ‘si’, ‘mbah’, ‘saya’, ‘anak’, ‘muda’, ‘kece’, ‘yang’, ‘batikan’, ‘nyari’, ‘irit’, ‘walau’, ‘gagal’, ‘wkwk’]
2984	naksir banget juga sama hrv	[‘naksir’, ‘banget’, ‘juga’, ‘sama’, ‘hrv’]

The text can be converted into good Indonesian according to the KBBI where there are word changes, such as “yg” to yang, “ga” to “tidak” and others. The research uses normalized.csv data accessed from github.

Table 9. Data Normalization

No.	Tokens	Normalisasi
1	[‘jangan’, ‘livina’, ‘apalagi’, ‘kalo’, ‘buat’, ‘nanjak', 'better', 'nomor', 'ertiga', 'atau', 'ignis', 'aja', 'iritnya', 'polll', 'tapi', 'harus', 'terbiasa', 'sama', 'rasa', 'ags', 'nya', 'yang', 'nyebelin', 'sekian']	['jangan', 'livina', 'apalagi', 'kalau', 'buat', 'nanjak', 'better', 'nomor', 'ertiga', 'atau', 'ignis', 'saja', 'iritnya', 'polll', 'tetapi', 'harus', 'terbiasa', 'sama', 'rasa', 'ags', 'nya', 'yang', 'nyebelin', 'sekian']
2	['klo', 'butuh', 'rwd', 'jelas', 'terios', 'klo', 'butuh', 'nyaman', 'dan', 'badak', 'matiknya', 'ya', 'ertigaxl', 'klo', 'mau', 'harga', 'jual', 'aman', 'dan', 'dipake', 'di', 'kota', 'doang', 'ya', 'avanzaxenia']	['kalau', 'butuh', 'rwd', 'jelas', 'terios', 'kalau', 'butuh', 'nyaman', 'dan', 'badak', 'matiknya', 'ya', 'ertigaxl', 'kalau', 'mau', 'harga', 'jual', 'aman', 'dan', 'dipake', 'di', 'kota', 'doang', 'ya', 'avanzaxenia']
3	['suzuki', 'baleno', 'hatchback', 'facelift', 'gagah', 'dan', 'berkelas']	['suzuki', 'baleno', 'hatchback', 'facelift', 'gagah', 'dan', 'berkelas']
4	['avanza', 'pasang', 'sunroof', 'aja', 'udah', 'next', 'level'	['avanza', 'pasang', 'sunroof', 'saja', 'sudah', 'next', 'level']
5	['sebagai', 'pengguna', 'civic', 'saya', 'merasa', 'tidak', 'setuju', 'dgn', 'kata', 'kata', 'si', 'mbah', 'saya', 'anak', 'muda',	['sebagai', 'pengguna', 'civic', 'saya', 'merasa', 'tidak', 'setuju', 'dengan', 'kata', 'kata', 'si', 'mbah', 'saya', 'anak', 'muda',

ISSN: 2716-3865 (Print), 2721-1290 (Online)

	'kece', 'yang', 'batikan', 'nyari', 'irit', 'walau', 'gagal', 'wkwk']	'kece', 'yang', 'batikan', 'nyari', 'irit', 'walau', 'gagal', 'wkwk']
2984	['naksir', 'banget', 'juga', 'sama', 'hrv']	['naksir', 'banget', 'juga', 'sama', 'hrv']

The detokenized process is a process that makes text that has been stemmed into comments that are suitable for analysis or have passed the data pre-processing period. Detokenized can be seen in the table below

Table 10. Detokenized Text

No.	Stemming	Detokenized
1	['livina', 'nanjak', 'better', 'nomor', 'ertiga', 'ignis', 'irit', 'polll', 'biasa', 'ags', 'nyebelin', 'sekian']	livina nanjak better nomor ertiga ignis irit polll biasa ags nyebelin sekian
2	['butuh', 'rwd', 'terios', 'butuh', 'nyaman', 'badak', 'matiknya', 'ya', 'ertigaxl', 'harga', 'jual', 'aman', 'dipake', 'kota', 'doang', 'ya', 'avanzaxenia']	butuh rwd terios butuh nyaman badak matiknya ya ertigaxl harga jual aman dipake kota doang ya avanzaxenia
3	['suzuki', 'baleno', 'hatchback', 'facelift', 'gagah', 'kelas']	suzuki baleno hatchback facelift gagah kelas
4	['avanza', 'pasang', 'sunroof', 'next', 'level']	avanza pasang sunroof next level
5	['guna', 'civic', 'tuju', 'dgn', 'si', 'mbah', 'anak', 'muda', 'kece', 'bati', 'nyari', 'irit', 'gagal', 'wkwk']	guna civic tuju dgn si mbah anak muda kece bati nyari irit gagal wkwk
2984	['naksir', 'banget', 'hrv']	naksir banget hrv

After the data containing the word spam is deleted, the next step is to delete duplicate or *double comment* data. Data is obtained where initially it amounted to 2984 to 2831 data where 153 data containing duplicate comments were deleted.

Tabel 11. Before and After Removing Duplicated Comments

Before	After
jangan livina apalagi kalo buat nanjak better nomor ertiga atau ignis aja iritnya polll tapi harus terbiasa sama rasa ags nya yang nyebelin sekian	jangan livina apalagi kalo buat nanjak better nomor ertiga atau ignis aja iritnya polll tapi harus terbiasa sama rasa ags nya yang nyebelin sekian
suzuki jimny doors dan suzuki ertiga cruise sukses bius pengunjung iims surabaya	suzuki jimny doors dan suzuki ertiga cruise sukses bius pengunjung iims surabaya
ertiga paling pas siapin budget buat jajan kaki kaki dan pastikan kalau ganti kakikaki pake sparepart ori sgp terutama shockbreaker ertiga sama aja sihh cuma beda tipis aja	ertiga paling pas siapin budget buat jajan kaki kaki dan pastikan kalau ganti kakikaki pake sparepart ori sgp terutama shockbreaker ertiga sama aja sihh cuma beda tipis aja
2984	2831

The next step is the data labeling process which is done automatically by implementing the *InSetLexicon* Corpus dictionary. In the *InSetLexicon* dictionary, words and their weights are used to determine the polarity value. The polarity value is calculated by summing up all the weights of the words in the review text. The result of the calculation is to classify the review for each aspect as positive or negative sentiment. A review is categorized as positive sentiment if its polarity value is greater than zero, and negative sentiment if its polarity value is lower than zero. Table 4.11 shows an example of the data labeling used using the InSet Lexicon dictionary. 1 is categorized as positive and 0 is categorized as negative.

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Table 12. Positive Corpus

No.	Word	Weight
1	Hai	3
2	Tetap	3
3	Detail	2
4	Bagus	2
3610	Orisinal	3

Table 13. Negative Corpus

No.	Word	Weight
1	Isak	-5
2	Sakit	-5
3	Mulu	-1
4	Gamau	-4
6610	Mencoreng	-4

Table 14. Illustration of InsetLexicon Implementation

Review 1 : tetap civic kalau saya mah desain nya								
	tetap	civic	kalau	saya	mah	desain	nya	**Polarity**
Word Weight	3	-	-1	-3	-	4	-	**3**
Label								**Positive**
Review 2 : daihatsu sigra ternyata tidak nyaman								
	daihatsu	sigra	ternyata	tidak	nyaman			**Polarity**
Word Weight	-	-	-	-5	4			**-1**
Label								**Negative**
Review 3 : ertiga emang keren bodi nya kuat								
	ertiga	emang	keren	bodi	nya	kuat		**Polarity**
Word Weight	-	4	-	-	-	2		**6**
Label								**Positive**

Table 15. Labeling Using InsetLexicon Corpus

No.	Text	Label
0	don't livina especially if you want to go uphill better number ertiga or ignis aja economical polll but have to get used to the taste of ags which is annoying at all.	Negative
1	It's still a civic if I like the design.	Positive
2	daihatsu sigra turned out to be uncomfortable	Negative
3	ertiga is just the same, it's just a slight difference.	Positive
4	any ertiga is ok anyway for me everything is ok	Positive
2984	hopefully the thr is bought hrv aamiin	Positive

The aspect categorization stage is carried out to classify comment data based on aspects that affect the quality of the car. In the aspect categorization stage, comment data is manually selected with a total of 580 data, because in the dataset there are many comments that contain spam, for example, car sales, use of people's names and others. This research uses 5 aspects, namely design, comfort, features, *price/money* and service/maintenance. Can be seen in table 16.

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Table 16. Aspect Description of Comment Data

Aspects	Description
Design	Reviews the aesthetic aspects of the vehicle. This includes the style, shape, color and materials used.
Comfortable	Assess comfort in terms of vehicle space. Includes seat comfort, temperature regulation and more.
Features	Reviewing the various features offered, both from safety, entertainment and other advanced technology features.
Price	Assess the value of the vehicle in terms of price compared to the features and quality offered.
Service	Review aspects of vehicle servicing and maintenance, including the ease of obtaining servicing and maintenance costs.

Table 17. Description of Number of Aspects

Aspects	Word Count
Design	13
Comfortable	42
Features	29
Price	59
Maintenance	11

The amount of initial data in comment analysis research using the Naïve Bayes Algorithm is 580, because in the dataset there are many comments that contain spam, for example, car sales, use of people's names and others. The ratio between training data and test data in this study is 0: 2, meaning that the training data totals 2387 data, while the test data totals 597 data. The training process will produce the weight of each word in each class using the TF-IDF weighting method. By using 3 previous sample data as training data, 1 test data is determined as follows.

Table. 18 Sample Test Data

Test Comments
Really Hybrid Ertiga Million Buy WKWK

Sentiment classification is done automatically using the Naïve Bayes algorithm. This process is performed using the MultinomialNB function that compares the weight of each word in the test data with the words in the training data. As a result, each training document will have an equal number of probabilities of positive and negative words. Next, a weight comparison is performed on the test document, where if the positive probability weight is greater, the classified sentiment is positive, while if the negative probability weight is greater, the classified sentiment becomes negative. The class classification process begins by calculating the prior probability, conditional probability, and posterior probability. The following are the stages of the classification process using the Naïve Bayes algorithm on test data. The specified comments use 3 comments obtained from the training data which are then matched with the test data samples.

Table. 19 *Sample* Data Training Data

No.	Commentary	Label
1.	ertiga Really cool strong body	Positive
2.	Suzuki Ertiga Only LMPOVE System Mild Hybrid Gini Breakdown Cost	Negative

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Table. 20 TF Data Train

No.	Vocabulary	Tf(Positive)	Tf(Negative)
1.	Ertiga	1	0
2.	Emang	1	0
3.	Keren	1	0
4.	Bodi	1	0
5.	Kuat	1	0
6.	Suzuki	0	1
7.	Ertiga	0	1
8.	Satusatunya	0	1
9.	Improve	0	1
10.	Sistem	0	1
11.	Mild	0	1
12.	Hybrid	0	1
15.	Gini	0	1
16.	biaya	0	1
17.	rusak	0	1
Number of *Terms*		5	10

Term Total = 15

Calculation of Probability Value

$$P\ (Class \mid Comment) = \frac{Number\ of\ Class\ X}{Number\ of\ Comment}$$

Using the above equation, we will obtain the probability of each class in the sentiment

P (Positive | comment) = $\frac{1}{2}$ = 0.5

P (Negative | comment) = $\frac{1}{2}$ = 0.5

Calculation of Conditional Probability Value :

$$P\ (Term \mid class) = \frac{Total\ weight\ TF\ Term\ on\ class + 1}{TF\ Weight\ Class + Total\ Weight\ TF}$$

Using the above equation, we will obtain the probability of the terms in each sentiment class.

Probability of the word “ertiga”

P(ertiga | Positive) = $\frac{1+1}{5+15}$ = $\frac{2}{20}$ = 0.1

P(ertiga | Negative) = $\frac{1+1}{10+15}$ = $\frac{2}{25}$ = 0.08

Probability of the word “emang”

P(emang | Positive) = $\frac{1+1}{5+15}$ = $\frac{2}{20}$ = 0.1

P(emang | Negative) = $\frac{1+1}{10+15}$ = $\frac{2}{25}$ = 0.08

Probabilitas kata “keren”

P(keren | Positive) = $\frac{1+1}{5+15}$ = $\frac{2}{20}$ = 0.1

P(keren | Negative) = $\frac{1+1}{10+15}$ = $\frac{2}{25}$ = 0.08

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Next is to take the test data, namely by classifying the test data by multiplying all the opportunities. The higher value is the new class of the data. In the test data included in the training data are the words "ertiga" and "hybrid". Calculation of posterior probability values.

$$P\ (Comment \mid Class) = P_{Term\ 1} x \ldots x P_{Term\ n} x\ P\ (Class|Comment)$$

P (Test | Positive) = P(positive) x P(ertiga|positive) x P(hybrid|positive)

= 0.5 x 0.1 x 0.04

= 0.002

P (Test | Negative) = P(Negative) x P(ertiga|Negative) x P(hybrid|Negative)

= 0.5 x 0.08 x 0.08

= 0.0032

Sentiment classification results, with the highest value of sentiment test data is negative with a value of 0.0032. This value was chosen based on the example calculation performed. Next is to look for accuracy based on the aspect category described, namely the aspect category of comment data manually selected with a total of 580 data, this study uses 5 aspects, namely design, comfort, features, price/money and service/maintenance.

Table. 21 Classification Testing Results

Aspect	Comment	Recall	Precision	F1-Score	Accuracy
Design	Positive	0.5	1.0	0.67	0.75
	Negative	1.0	0.67	0.8	
	Average	**0.75**	**0.83**	**0.73**	
Comfortability	Positive	1.0	0.9	0.95	0.91
	Negative	0.67	1.0	0.8	
	Average	**0.83**	**0.85**	**0.87**	
Feature	Positive	0.86	1.0	0.92	0.85
	Negative	0	0	0	
	Average	**0.43**	**0.5**	**0.46**	
Price/Money	Positive	0.87	0.81	0.84	0.72
	Negative	0	0	0	
	Average	**0.43**	**0.41**	**0.42**	
Service/Maintenece	Positive	1.0	0.67	0.8	0.75
	Negative	0.5	1.0	0.67	
	Average	**0.75**	**0.83**	**0.73**	
Average					0.75

The number of comments/sentiment analysis is shown in Figure 4 where 1825 positive comments and 1006 negative comments were obtained.

ISSN: 2716-3865 (Print), 2721-1290 (Online)

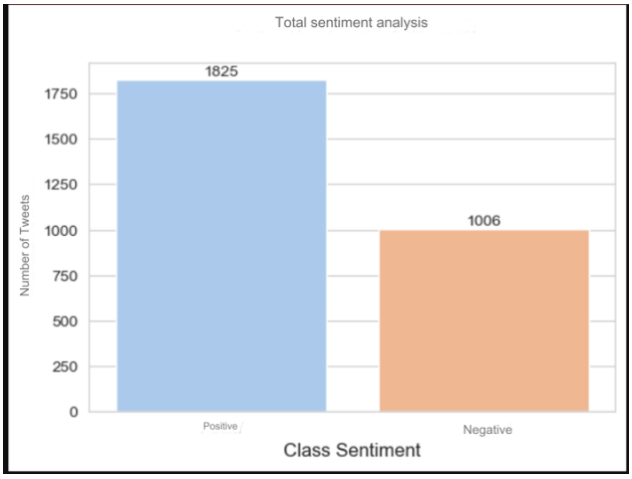

Figure 4. Number of Sentiment Analysis

After sentiment testing using the Naïve Bayes algorithm, the sentiment classification results will be obtained in the form of sentiment labels. This classification result label will be compared with the actual label to determine the *accuration*, *precision*, *recall* and F1 value of the model for the sentiment dataset used.

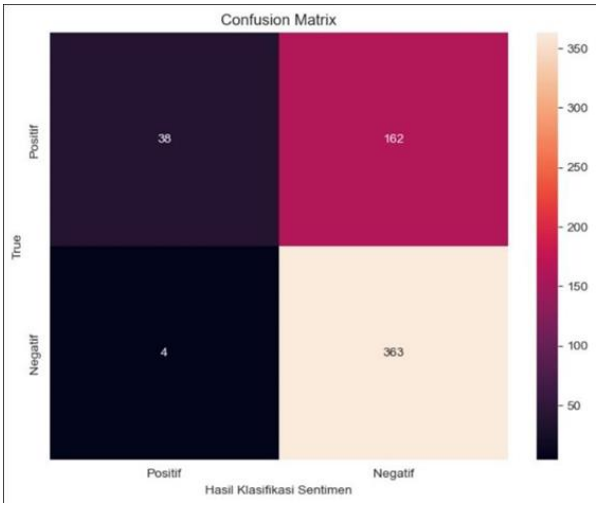

Figure 5. Confusion Matrix

In Figure 5. the classification results for finding *accuration*, *precision*, recall and *F1-score* follow the following formula.

$$accuracy = \frac{363 + 38}{363 + 4 + 38 + 162} x\ 100\% = 70\%$$

$$precision = \frac{38}{38 + 4} x\ 100\% = 90\%$$

$$recall = \frac{38}{38 + 162} x\ 100\% = 19\%$$

$$F1 - Score = \frac{2\ x\ 90 + 19}{90 + 19} x\ 100\% = 31\%$$

ISSN: 2716-3865 (Print), 2721-1290 (Online)

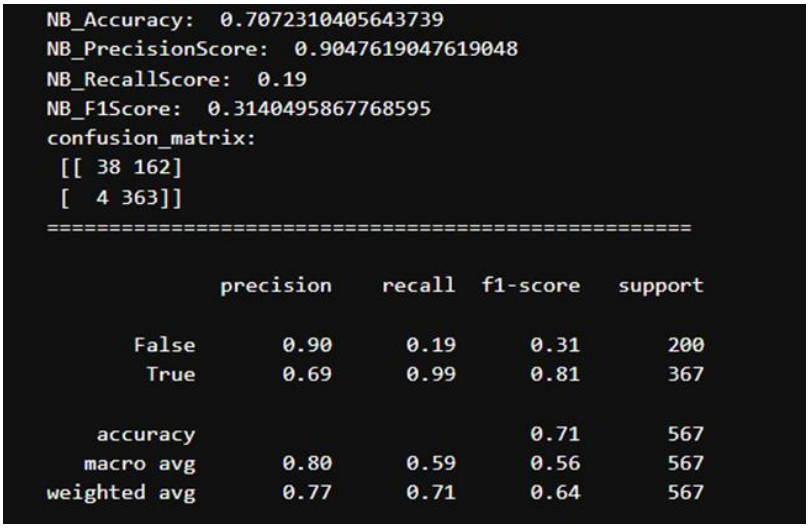

Figure 6. Dataset Accuracy

The use of Naïve Bayes for sentiment analysis in this work shows how far the machine learning models have come and how far they have to go especially in the context of NLP. The combination of InsetLexicon corpus used in the study to provide sentiments from the user comment from social media such as X formerly twitter and Naïve Bayes classification algorithm is an example of using lexicon-based approach with probabilistic model that is increasingly popular in the field.

The present investigation has successfully applied the Naïve Bayes algorithm in collaboration with the InsetLexicon corpus for the purpose of categorizing the sentiments successfully with the highest Recall, Precision and F-measure recorded in the "Comfortability" aspect at 0. 83, 0. 85, and 0. 87 respectively. This finding resonates with current study stressing the effectiveness of the lexicon-based sentiment analysis models in operational contexts (Zhang & Wang, 2023). A high accuracy result observed in the comfortability aspect is therefore in line with the capability of such models in sentiment classification tasks and especially in contexts where users' opinions on comfort and usability are crucial. This was achieved through very strict pre-processing steps such as data cleaning, normalization and tokenization which enhanced the performance of model. As highlighted in the current research studies, the aforementioned pre-processing techniques are extremely important in removing noises and producing high-quality inputs for sentiment analysis models (Chen et al., 2022; Torres et al., 2020; Naseem et al., 2021; Cai et al., 2022). In addition, the pairing of training and testing datasets in equal parts (80:20) used in this study is in line with most of the ML methods and framework that provides a accurate model training and evaluation (Liu et al., 2023).

However, Naïve Bayes have its drawbacks especially the assumption that words are independent which is a problem that is starting to gain much attention in the literature (Romano et al., 2024; Wickramasinghe & Kalutarage, 2021; Rezaeian & Novikova, 2020). Probably the most obvious drawback of the assumption concerns relations between words More often than not, words in natural language depend on one another, which might cause errors in sentiment classification, particularly in subtle contexts (Li et al., 2023). This drawback is most prominent in the classification of comments concerning the "Price/Money" and "Feature" aspects, which have a significantly lower F1-Scores. This indicates that, though Naïve Bayes works well in some environments, it may not perform well with more complicated forms of sentiment such as those that involve the understanding of the relationships between words and context s as observed in other studies by Sun et al. (2024). The fact that the InsetLexicon corpus does not appear to change over time exacerbates the lack of the ability of the model to adjust to current trends in language use on social media sites, where such shifts could be rather frequent (Cahyawijaya et al., 2023). Such static

ISSN: 2716-3865 (Print), 2721-1290 (Online)

approach is in juxtaposition to the dynamic models capable of modifying the lexicon entries based on appearance of new slangs or changes in language use as was pointed out as a key area to future development of sentiment analysis (Wang & Jiang, 2023).

Based on the same, some perspective for further research and development can be evaluated. There is, however, a potential to extend the application of more sophisticated algorithms which have been proven to work more effectively in understanding of natural language since they address interactions between words and context (Gao et al., 2024). These models could in part overcome the constraints that have been identified with Naïve Bayes particularly regarding the complexities of sentiments in expressions (Nhu et al., 2020; Ali et al., 2020; Piryonesi et al., 2020; Tang et al., 2020).

Furthermore, synchronization of dictionaries that are adaptable to the peculiarities of the context is an important step in enhancing sentiment analysis in such dynamic environments as X known today In the contemporary environment, new and more advanced NLP methods that can help to construct dynamic lexicons adaptable to the changes in language usage include such tools as BERT and GPT (Devlin et al., 2023). These models could offer the premise for improved and more dynamic sentiment analysis instruments that are more appropriate to the rapid shifts of language on social media. Finally, adding more articles to the range of sources and more languages would improve the model's applicability and reliability. For instance, multilingual datasets would enable the identification of sentiment patterns per language, a novel field that is becoming critical in today's interlingual society.

Conclusion

Based on the research results, the positive and negative sentiment labeling process carried out using the Indonesian Sentiment Lexicon (InSet Lexicon) dictionary, the results in Table 4.20 show that the value of the total comments on the highest aspect of all aspects of the 4 types of cars and their respective brands, namely the "Comfortable" or comfort aspect, is around 91% with the acquisition of the average Recall, Precision and F1-Score of around 0.83, 0.85 and 0.87 and the overall accuracy of the dataset is around 71%. It is hoped that future research can take review data from more sources with similar topics without being limited to one / two research objects only and can be considered to make many modifications to the lexicon so that the weight given is closer to perfect than before.

References

A. Yani, D. D., Pratiwi, H. S., & Muhardi, H. (2019). Implementasi Web Scraping untuk Pengambilan Data pada Situs Marketplace. *Jurnal Sistem Dan Teknologi Informasi (JUSTIN)*, *7*(4), 257. https://doi.org/10.26418/justin.v7i4.30930

Ali, S. A., Parvin, F., Pham, Q. B., Vojtek, M., Vojteková, J., Costache, R., ... & Ghorbani, M. A. (2020). GIS-based comparative assessment of flood susceptibility mapping using hybrid multi-criteria decision-making approach, naïve Bayes tree, bivariate statistics and logistic regression: a case of Topľa basin, Slovakia. *Ecological Indicators*, *117*, 106620. https://doi.org/10.1016/j.ecolind.2020.106620

Cahyawijaya, S., Lovenia, H., Koto, F., Adhista, D., Dave, E., Oktavianti, S., ... & Fung, P. (2023). Nusawrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages. *arXiv preprint arXiv:2309.10661*. https://doi.org/10.48550/arXiv.2309.10661

Cai, J., Yang, Y., Yang, H., Zhao, X., & Hao, J. (2022). ARIS: a noise insensitive data pre-processing scheme for data reduction using influence space. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *16*(6), 1-39. https://doi.org/10.1145/3522592

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier. *International Journal of Information Engineering and Electronic Business*, *8*(4), 54–62. https://doi.org/10.5815/ijieeb.2016.04.07

Firdaus, R., Asror, I., & Herdiani, A. (2021). Lexicon-Based Sentiment Analysis of Indonesian Language Student Feedback Evaluation. *Indonesia Journal on Computing (Indo-JC)*, *6*(1), 1–12. https://doi.org/10.34818/INDOJC.2021.6.1.408

Fitri, V. A., Andreswari, R., & Hasibuan, M. A. (2019). Sentiment Analysis of Social Media Twitter with Case of Anti-LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm. *Procedia Computer Science*, *161*, 765–772. https://doi.org/10.1016/j.procs.2019.11.181

Goel, A., Gautam, J., & Kumar, S. (2016). Real Time Sentiment Analysis of Tweets Using Naive Bayes. *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, 257–261. https://doi.org/10.1109/NGCT.2016.7877424

Isam, H., & Abd Mutalib, M. (2019). Pemanfaatan Analisis Korpus sebagai Teknik Alternatif Pengajaran dan Pembelajaran Tatabahasa. *International Journal of Language Education and Applied Linguistics*, *9*(1), 13–31. https://doi.org/10.15282/ijleal.v9.594

Ismail, A. R., & Hakim, R. B. F. (2023). Implementasi Lexicon Based untuk Analisis Sentimen dalam Menentukan Rekomendasi Pantai di DI Yogyakarta Berdasarkan Data Twitter. *Emerging Statistics and Data Science Journal*, *1*(1), 37–46. https://doi.org/10.20885/esds.vol1.iss.1.art5

Khoirunisa, R. (2020). Penggunaan Natural Language Processing Pada Chatbot Untuk Media Informasi Pertanian. *Indonesian Journal of Applied Informatics*, *4*(2), 55. https://doi.org/10.20961/ijai.v4i2.38688

Mayasari, L., & Indarti, D. (2022). Klasifikasi Topik Tweet Mengenai Covid Menggunakan Metode Multinomial Naïve Bayes dengan Pembobotan TF-IDF. *Jurnal Ilmiah Informatika Komputer*, *27*(1), 43–53. https://doi.org/10.35760/ik.2022.v27i1.6184

Naseem, U., Razzak, I., & Eklund, P. W. (2021). A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter. *Multimedia Tools and Applications*, *80*, 35239-35266. https://doi.org/10.1007/s11042-020-10082-6

Nhu, V. H., Shirzadi, A., Shahabi, H., Singh, S. K., Al-Ansari, N., Clague, J. J., ... & Ahmad, B. B. (2020). Shallow landslide susceptibility mapping: A comparison between logistic model tree, logistic regression, naïve bayes tree, artificial neural network, and support vector machine algorithms. *International journal of environmental research and public health*, *17*(8), 2749. https://doi.org/10.3390/ijerph17082749

Nugroho, A., & Religia, Y. (2021). Analisis Optimasi Algoritma Klasifikasi Naive Bayes menggunakan Genetic Algorithm dan Bagging. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(3), 504–510. https://doi.org/10.29207/resti.v5i3.3067

Obiedat, R., Al-Darras, D., Alzaghoul, E., & Harfoushi, O. (2021). Arabic Aspect-Based Sentiment Analysis: A Systematic Literature Review. *IEEE Access*, *9*, 152628–152645. https://doi.org/10.1109/ACCESS.2021.3127140

Piryonesi, S. M., & El-Diraby, T. E. (2020). Role of data analytics in infrastructure asset management: Overcoming data size and quality problems. *Journal of Transportation Engineering, Part B: Pavements*, *146*(2), 04020022. https://doi.org/10.1061/JPEODX.0000175

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Prasetyo, V. R., Benarkah, N., & Chrisintha, V. J. (2021). Implementasi Natural Language Processing Dalam Pembuatan Chatbot Pada Program Information Technology Universitas Surabaya. *Teknika*, *10*(2), 114–121. https://doi.org/10.34148/teknika.v10i2.370

Rachman, R., & Handayani, R. N. (2021). Klasifikasi Algoritma Naive Bayes dalam Memprediksi Tingkat Kelancaran Pembayaran Sewa Teras UMKM. *Jurnal Informatika*, *8*(2), 111–122. https://doi.org/10.31294/ji.v8i2.10494

Rezaeian, N., & Novikova, G. (2020). Persian text classification using naive bayes algorithms and support vector machine algorithm. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, *8*(1), 178-188. http://dx.doi.org/10.52549/ijeei.v8i1.1696

Romano, M., Contu, G., Mola, F., & Conversano, C. (2024). Threshold-based naïve bayes classifier. *Advances in Data Analysis and Classification*, *18*(2), 325-361. https://doi.org/10.1007/s11634-023-00536-8

Sutabri, T., Suryatno, A., Setiadi, D., & Negara, E. S. (2018). Improving Naïve Bayes in Sentiment Analysis For Hotel Industry in Indonesia. *2018 Third International Conference on Informatics and Computing (ICIC)*, 1–6. https://doi.org/10.1109/IAC.2018.8780444

Tang, X., Li, J., Liu, M., Liu, W., & Hong, H. (2020). Flood susceptibility assessment based on a novel random Naïve Bayes method: A comparison between different factor discretization methods. *Catena*, *190*, 104536. https://doi.org/10.1016/j.catena.2020.104536

Torres-García, A. A., Mendoza-Montoya, O., Molinas, M., Antelis, J. M., Moctezuma, L. A., & Hernández-Del-Toro, T. (2022). Pre-processing and feature extraction. In *Biosignal processing and classification using computational learning and intelligence* (pp. 59-91). Academic Press. https://doi.org/10.1016/B978-0-12-820125-1.00014-2

Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Computing*, *25*(3), 2277-2293. https://doi.org/10.1007/s00500-020-05297-6

ISSN: 2716-3865 (Print), 2721-1290 (Online)