Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 08, No. 03, September 2022.

SENTIMENT ANALYSIS BASED ON PROBABILISTIC CLASSIFIER TECHNIQUES IN VARIOUS INDONESIAN REVIEW DATA

Nur Hayatin[1], Suraya Alias[2], Lai Po Hung[2], Mohd Shamrie Sainin[2]

(Received: 10-Mar.-2022, Revised: 28-Apr.-2022, Accepted: 24-May-2022)

ABSTRACT

Sentiment analysis is the field in data science to achieve a broader holistic view of users' needs and expectations. Indonesian user opinions have the potential to manage to be valuable information using sentiment-analysis tasks. One of the most supervised-learning techniques used in Indonesian sentiment analysis is the Naïve Bayes classifier. The classifier can be optimized and tuned in various models to increase the sentiment analysis model performance. This research aims to examine the performance of various Naïve Bayes models in sentiment analysis, especially when implemented in small datasets to handle overfitting problems. Four different Naïve Bayes models used are Gaussian, Multinomial, Complement and Bernoulli. We also analyze the effect of various pre-processing techniques on the models' performance. Moreover, we build the first fashion dataset from the Indonesian marketplace which has a unique character compared to the datasets from other domains. Finally, we also use various datasets in the experiment to test the Naïve Bayes models' performance. From the experimental results, Complement Naïve Bayes is superior to other models, especially in handling overfitting with an F1-score of approximately 0.82.

KEYWORDS

Naïve Bayes model, Probabilistic classifier, Sentiment analysis, Supervised learning.

1. INTRODUCTION

In Natural Language Processing (NLP), the data is dominated by text that can come from a webpage, social media, online reviews, online news, etc. Sentiment analysis is a field that can achieve a broader holistic view of customers' needs and expectations [1]. Research in this field is not only studied in English, but also in various languages, such as Malay [2], Arabic [3]-[4], as well as Indonesian. It is an important method for social sciences, because of that it is used in various disciplines including analyzing reviews from e-commerce.

Indonesia has a big consumer base of e-commerce consisting of over 8 hundred million visitors in 2019 [5]. This is a potential to identify that Indonesian user opinions from product reviews on the internet become valuable information using sentiment-analysis tasks. Studies on Indonesian sentiment analysis have grown in recent years. We have reviewed more than 100 references related to Indonesian sentiment analysis using Machine learning (ML) techniques. From the review study results, we found some popular ML techniques implemented in Indonesian research; namely, Naïve Bayes, Support Vector Machine, Decision Tree and K-Nearest Neighbour. To the best of our knowledge, there are three various Naïve Bayes models implemented in Indonesian sentiment analysis; namely, Gaussian, Multinomial [6]-[7] and Bernoulli [8]. However, the simple probabilistic classifier is the most popular technique in Indonesian sentiment-analysis research [9]-[10]. Only one research used Bernoulli Naïve Bayes [8], while Complement Naïve Bayes has not been implemented in Indonesian sentiment-analysis research. Naïve Bayes has various probabilistic models; namely, Gaussian, Multinomial, Complement and Bernoulli that can be implemented and can increase the sentiment-analysis model performance.

Some previous research in Indonesian sentiment analysis using a Naïve Bayes classifier was conducted. Priadana & Rizal developed a sentiment-analysis model based on lexicon-based and Naive Bayes Classifiers [6]. The model is used to track trending topics and analyze the sentiment of public opinion on Instagram to figure out government performance in tourism from Instagram during the COVID-19 pandemic. They also implemented some pre-processing techniques, such as lowercase, removing

1. N. Hayatin is with Informatics Department, University of Muhammadiyah Malang, Indonesia. She is also a PhD student at Faculty of Computing and Informatics, Universiti Malaysia Sabah. Email: noorhayatin@umm.ac.id
2. S. Alias (corresponding author), L.P.Hung and M.S.Saini are with Computing and Informatics Faculty, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia. Emails: suealias@ums.edu.my, laipohung@ums.edu.my and shamrie@ums.edu.my

symbol, stemming, tokenizing and bag of words. Sutabri et al. [7] applied multinomial Naïve Bayes to analyze sentiment in Indonesian popular e-travelling sites. Meanwhile, other research applied a similar Naïve Bayes model to the education domain. Akbar et al. proposed a sentiment-analysis model using Bernoulli Naïve Bayes their model can differentiate between pro and contra tweets on the lockdown policy topics using Indonesian tweets [8].

This research focuses on analyzing sentiment in the Indonesian fashion dataset using the probabilistic classifier. The objective statement of the research is as follows:

1) Building a new sentiment dataset in the fashion domain from the Indonesian marketplace.
2) Examining the performance of various Naïve Bayes models in sentiment analysis for small datasets and overfitting handling.
3) Analyzing the effect of various pre-processing techniques on the model performance.

2. RESEARCH METHOD

The research methodology to conduct the research has five steps; namely, data gathering, pre-processing, feature extraction, sentiment classification and evaluation. Figure 1 depicts the methodology architecture of the research.

2.1 Data Gathering

There are various domains of data used in Indonesian sentiment analysis (see Figure 2). To the best of our knowledge, the Indonesian sentiment-analysis dataset in fashion domain has not been created before. We are the pioneers in building the dataset in this domain. From analyzing the dataset, we note some unique keywords of reviews in the Indonesian marketplace, especially in fashion opinions that are different from those in other domains. Those keywords are "*bahan*" (material), "*ukuran*" (size), "*pengiriman*" (delivery service), "*warna"* (colour) and "*harga"* (price). In this chapter, we explain the process of gathering the data.

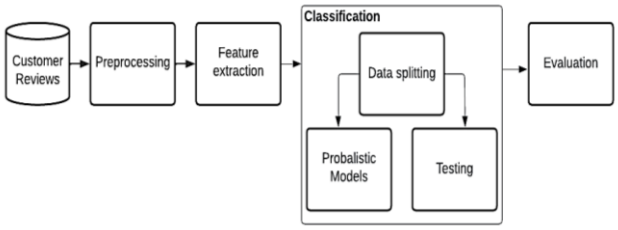

Figure 1. Methodology architecture.

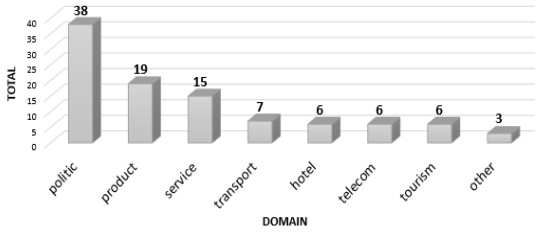

Figure 2. various domains used in Indonesian sentiment analysis.

We collect the data from product reviews scrapped from Shopee Marketplace, one of the big marketplaces in Indonesia [11]. We choose a product-related fashion. There are some criteria for products to be selected; i.e.: Total review is more than 1k comments; there is no far gap of total comments for each rating (1–5).

In gathering the data, the first step is copying the URL of the product which was selected, then scraping the product reviews using Python and Shopee API. The function used to access URL is *requests.get* that returned an object which is the HTML text; furthermore, parsing the HTML with *beautifulsoup* library to extract the HTML elements that are required.

The total data which has been scraped is 3020 reviews. This is the maximum number of data that can be scraped for one store in Shopee using the API. There are three data attributes we need to scrape; i.e.: *id order, comment and star*. We need *id order* to select unique data, because based on manual checking, there is duplicate data that is submitted from a user. This might happen because of accidentally double submitted data by a user or a system error. However, for the experiment, we only used two data; namely, comment and star. The review data that has resulted from scraping is shown in Table 1.

Data Labelling: In supervised learning, the primary step conducted before classifying is labelling the data. We generate label data automatically based on the rating score from the "star" column. Rating represents the acceptance and opinion of the product. There are five categories of rating through the

Table 1. Pairs of review data and star rating.

Comment	Star
Alhamdulillah terimakasih banyak semoga sukses selalu aminnnnn bagusss banget recommended seller deh semoga sukses. (In English: Alhamdulillah, thank you very much, good luck always aminnnnn, really recommended seller, good luck).	5
baguss worthit lahh buat harga segituu, pengirimannya juga lumayan cepat ga nyesel sii beli disinii. (In English: It's good, it's worth for that price, the delivery is also quite fast, I don't regret buying it here).	4
Barang bahanya agak tipis cma lumayan buat dipake sehari2 benang dan jahitanya kurng rapih. (In English: The material is a bit, thin but it's good enough for everyday use, the thread and the stitches are not neat).	3
Oversize nya kecil sekali..kecewa. (In English: The oversize is very small..disappointed).	2
Bahannya rusak gak sesuai fto kaos nya tipis banget gak ska. (In English: The material is ruined, it doesn't match the photo, the shirt is very thin, I don't like it).	1

number of stars given by the author. The range of stars is 1 to 5, where 5 is the highest star score which is interpreted positively *vice versa* 1 is the lowest star score that is interpreted negatively. The total of comments for each rating category from 1 to 5 is 177, 107, 258, 489 and 1787, respectively.

We adopt the Likert scale to convert the rating scores. The Likert scale is a bipolar scale method that measures both positive and negative responses to a statement. In sentiment analysis, data is divided into two classes based on sentiment polarity. In this research, data is labelled as positive (represented by "1") and negative (represented by "0"). A simple rule based on rating scores is used to label data automatically. The 5-star score will be transformed automatically into a positive label ("1") and others will be generated with a negative one ("0"). The data distribution after labelling is as follows: total data in label "1" is 1787, while total data in label "0" is 1031 data.

Data Balancing: Data balancing is a step that aims to achieve similarity to the total data of each label category. There are various techniques to do this process. This research tried to get balancing using minimum data standards. From Figure 3, we know that label "0" has minimum data with a total of 1031. Therefore, label "1" will be pruned, so that the total is like that of label "0". For a simple process, both labels now have similar total data: 1000.

Finally, the clean data is produced with the total data being 2000 selected rows. However, pre-processing techniques are implemented and we clean the data, especially to filter empty data after pre-processing. Therefore, the final data filtered is a total of 520 comments. Table 2 shows the transformation of the data starting from scraping to balancing.

Table 2. Total data transformation.

Original data from scraping	Drop null	Data balancing	
		Before pre-processing	After pre-processing
3020	2818	2000	520

To better understand the data, we visualize the group of data in the form of a word cloud based on a sentiment label. Word cloud contains the terms selected from the dataset and then shown based on the higher frequency of occurrence of the term. For each label, positive and negative, visualization is depicted in Figure 3. From the positive word cloud in Figure 3(a), we can see that there are some dominant words, such as *barang bagus, kiriman cepat, harga sesuai* (in English: good item, fast delivery, good price). Meanwhile, some words represent a negative expression, such as *barang tidak sesuai, size kecil, kecewa, rusak* (in English: the item does not match, the size is small, disappointed, damaged), as presented in Figure 3 (b).

(a) (b)

Figure 3. Word cloud visualization; (a) positive and (b) negative.

2.2 Pre-processing

The second step implemented in this research is pre-processing after gathering the data. There are some stages in pre-processing; to simplify, we group those into three main stages of pre-processing; namely, data cleansing, data transforming and data tagging. Figure 4 depicts the order of the main stages of pre-processing.

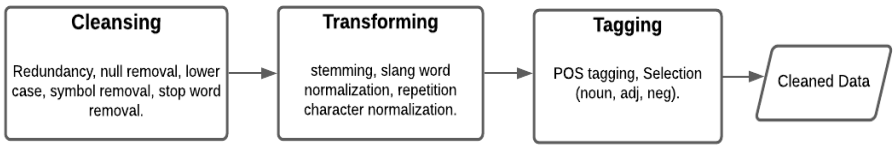

Figure 4. Pre-processing main stages.

Data Cleansing. This is the first stage of pre-processing. There are some stages to clean the data included; i.e.: redundancy, null-value removal, lower case, symbol removal and stop word removal. Redundancy removal is a process aimed to get a unique row; it was selected based on the "*id order*". After this process is carried out to the dataset, the total data is still in 3020 rows, so the data collection is unique or has no redundancy.

The next step is the null-value removal. This process is targeted to select a row that has a null value and then remove it. After it was executed to the dataset, there is a reduction of the total data of the dataset from 3020 rows to 2818 reviews. Lower case is a process to change each abjad of the sentence to be lower. It is used to reduce the data dimension; for example, there is the phrase= {"*Baju bagus", "baju bagus", "Baju Bagus*"}; if we do not use lower case, then after tokenizing these three phrases will be saved as 4-term collection= {"*Baju", "baju", "Bagus", "bagus*"}. However, if we use lower case, then there is only 2-term collection= {"*baju", "bagus*"}. It is produced, because the system will be saved for each unique word based on the character.

This research use symbol removal to clean the data. The symbols that will be removed involve punctuation, ASCII, UNICODE & Newline. This stage automatically includes emoji removal. The last stage of data cleansing is the stop-word removal step. This process is to remove unimportant words. We use a dictionary containing standard words to select stop words; if the word is not in accordance with a word in the dictionary, then that word will be removed.

Data Transforming. At this stage, the words of the sentences will be changed into different word forms. The processes in this stage are stemming, slang word normalization and repetition of character normalization. The last two processes are a part of spelling correction. A previous study has used this technique for pre-processing and the effectiveness of the model has been shown [12].

First, we use stemming to change stem words into root words. The stemming library used in this research is Sastrawi stemmed, an Indonesian stemming algorithm. The stemming result example; i.e.: stem word "*pengiriman*" (in English: delivery), will proceed to be the root word "*kirim*" (in English: send).

Many conversations on online media are done using slang words. This trend also happens in the marketplace, especially in the Indonesian customer reviews with the users' style for expressing their words. Based on the researchers' observation, the type of slang word that is usually used in the marketplace is Collegial. Colloquial is a socio-linguistic term related to a non-formal or informal language which is also referred to as a daily language [13]. The hallmark of this language, among others, is the reduced use of linguistic features, such as letters and syllables in sentences. Slang word normalization is needed to transform slang words, words that are unrecognizable in the dictionary, to be standard words. In this research, we use the slang word dictionary from Okky Ibrahim Github[1].

The next technique used for data transformation is repetition character normalization. This technique is like the previous technique and normalizes unstandardized words to be standard. Unstandardized words are related with that there is the same character mentioned in repetition in the sentence. Table 3 shows an example of a comment before and after being implemented with repetition of character normalization.

[1] https://github.com/okkyibrohim/id-abusive-language-detection/blob/3f511561df6b1ae60f7343f8992d1471209ff10b/kamusalay.csv

Table 3. Repetition character normalization.

Example: "*sumpah **iniii tokooo gercepp bangettt**, mesen kemeja **kemarennn langsungg dikirimmm** juga **hariii ituuuuu**!!! bahannya juga **tebelll pokoknyaa** tidak mengecewakan!!!! **cusss gaisss beliiii disiniiii** di jamin **bagusss**!!"*
Normal: "*sumpah **ini toko gercep banget**, mesen kemeja **kemaren langsung dikirim** juga **hari itu**!!! bahannya juga **tebel pokoknya** tidak mengecewakan!!!! **cus gais beli disini** di jamin **bagus**!!"*

Data Tagging. The final stage of pre-processing is data tagging. This stage will split the data word by word and then give the relevance tag for each word of the sentence; it is generally mentioned as POS (Part of Speech) tagging. In this research, we used *CRFTagger()* library, an Indonesian tagger. After each word is tagged, we can select which words will be used, where this research filters words classified as NN (Noun), JJ (Adjective) and Neg (Negation). This data-tagging result is also needed when we generate word cloud visualization.

2.3 Feature Extraction

We use TFIDF (Term Frequency Inverse Document Frequency) to extract the features [14]. The feature of review sentences is in the form of text. Therefore, TFIDF is needed to generate text to number through term weighting. The concept of TFIDF is that the word T_i is important if it occurs frequently. The values of the vector elements W_i for a document *d* are calculated as a combination of the statistics TF and IDF. The calculation of W_i is as follows:

$$W_i = TF(t_i, d) \cdot IDF(t_i) \tag{1}$$

where W_i is the weight of word t_i in document *d*. The term frequency *TF(t,d)* is the number of times word *t* occurs in document *d*, while the document frequency *DF(t)* is the number of documents in which the word *t* occurs at least once. The inverse document frequency *IDF(t)* can be calculated from the document frequency by:

$$IDF(t) = log\left(\frac{|D|}{DF(t)}\right) \tag{2}$$

where |*D*| is the total number of documents. The inverse document frequency of a word is low if it occurs in many documents and is highest if the word occurs in only one.

2.4 Sentiment Classification

Sentiment classification is a process to classify the data which is grouped based on the relevant sentiment class. In this research, we will classify the data into two classes based on positive sentiment and negative sentiment. We implement a classifier model; namely, Naïve Bayes.

Naive Bayes is a supervised-learning algorithm that is the simplest form of a Bayesian network [15]. This algorithm works based on Bayes' theorem with the "naive" assumption of conditional independence, where all attributes are independent given the value of the class variable. Given class variable *c* and dependent feature vector x_1 to x_n in document *d*; the probability of each sentiment class *c* is calculated as:

$$P(c, x_i) = \frac{P(x_i, c) \cdot P(c)}{P(x_i)} \tag{3}$$

where $P(x_i)$ is the same for all classes; then the class label of x_i can be determined by:

$$label(x_i) = argMax_c\{P(c, x_i)\} \tag{4}$$

There are various models of Naïve Bayes, such as Gaussian, Multinomial, Complement and Bernoulli. The difference between each Naïve Bayes model is determined by the calculation of probability $P(x_i, c)$.

Gaussian Naïve Bayes. In Gaussian NB, feature values of terms for each class *c* are usually generated by a separate Gaussian [16], where σ and μ of the feature values of words are associated with class *c*. The likelihood of feature x_i is given by:

$$P(x_i, c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} exp\left(-\frac{(d_i - \mu_c)^2}{2\sigma_c^2}\right) \tag{5}$$

The parameters σ is the variance vector and μ is the mean, while these parameters are estimated using maximum likelihood from the training document set *d*.

Multinomial Naïve Bayes. The multinomial NB classifier captures the term frequency of the documents [17]. This model is implemented for multinomially distributed data, so that it is suited for discrete feature classification.

For each class *c*, where *n* is the size of the vocabulary in all classes of the training dataset, the probability $P(x_i, c)$ of feature *i* appearing in a sample belonging to class *c* is estimated by a smoothed version of maximum likelihood as follows:

$$P(x_i, c) = \frac{N_{ci}+\alpha}{N_c+\alpha n} \quad (6)$$

where $N_{ci} = \sum_{x \in T} x_i$ is the number of times feature *i* appears in a sample of class *c* in the training set *T* and $N_c = \sum_{i=1}^{n} N_{ci}$ is the total count of all features for class *c*.

The smoothing priors $\alpha \geq 0$ account for features not present in the learning samples and prevent zero probabilities in further computations. Setting $\alpha=1$ is called Laplace smoothing, while $\alpha<1$ is called Lidstone smoothing.

Complement Naïve Bayes. This model is an adaptation of the standard multinomial Naive Bayes (MNB) algorithm that is particularly suited for imbalanced datasets. Complement NB uses statistics from the complement of each class to compute the model's weights [18]. It will lessen the bias in the weight estimates and will improve the classification accuracy. The procedure for calculating the weights is as follows:

$$P(x_i, c) = \frac{N_{\hat{c}i}+\alpha_i}{N_{\hat{c}}+\alpha} \quad (7)$$

where $N_{\hat{c}i} = \sum_{j:y_j \neq c} d_{ij}$ is the number of times word *i* occurred in documents in classes other than *c* and $N_{\hat{c}} = \sum_{j:y_j \neq c} \sum_k d_{kj}$ is the total number of word occurrences in classes other than *c* and α_i and α are smoothing hyperparameters. The classification rule is:

$$\hat{c} = \arg min_c \sum_i t_i w_{ci} \quad (8)$$

where a document is assigned to the class that is the poorest complement match.

Bernoulli Naïve Bayes. This type of classifier assumes that the features are binary and require only 2 values, where each value shows whether a word occurs or does not occur at least once in the document; with a value ranging between 0 and 1 [19]. The decision rule for Bernoulli Naive Bayes is based on:

$$P(x_i, c) = P(x_i, c) b_i + (1 - p(x_i|c))(1 - b_i) \quad (9)$$

where *x* is a word in the document; If the word x_i is present in the document, then $b_i = 1$ and the likelihood was $P(x_i, c)$. If the word x_i is absent, then $b_i = 0$ and the probability is $(1 - p(x_i|c))$.

3. RESULT AND DISCUSSION

In this research, we measure the performance of the probabilistic classifiers in the sentiment-analysis model. For the evaluation model, we start with preparing the datasets needed to build the model and to test the model performance. After that, data is cleaned using various pre-processing techniques. Finally, the data is classified with employing various Naïve Bayes models; namely, Gaussian, Multinomial, Complement and Bernoulli.

3.1 Dataset

The dataset used in the research is from both primary and secondary data. The dataset needs to be split, because we use supervised learning that needs trained data to build the model. Dataset is separated into two groups: train data and test data. In this research, data smoothing for each sentiment class is noticed to produce balanced data.

For the primary dataset, we use a fashion dataset that contained 520 reviews scraped from the Indonesian marketplace. The detailed process for scraping data is explained in Section 2. Total number of data for training is 416, while that of test data is 104. Table 4 shows the proportions of the total primary data split for each sentiment label.

For the secondary data, we gathered various Indonesian reviews as a benchmark from the open public

dataset used in sentiment-analysis research[2]. Four public datasets were utilized especially for conducting the third experiment scenario; namely, cellular [20], cyberbullying [21], movie [22] and politic [23]. The details of proportions for each dataset can be seen in Table 5.

Table 4. Primary data.

Total	Train		Test	
	Positive	Negative	Positive	Negative
520	216	200	53	51

Table 5. Various Indonesian public datasets for sentiment analysis.

Dataset	#Data	Positive	Negative
cellular	300	169	139
cyberbullying	400	200	200
movie	200	100	100
politic	900	450	450

3.2 Experimental Setup

We use Scikit Learn library to implement the algorithms of Naïve Bayes models; namely, Gaussian, Multinomial, Complement and Bernoulli [24]. We have three scenarios of the experiment. In the first scenario, we test some pre-processing techniques to analyze which pre-processing technique affects the model performance. In the second scenario, we test various Naïve Bayes models in sentiment analysis using a fashion dataset. And in the last scenario, we use some public datasets in Indonesian sentiment analysis to measure the models' performance as well as to examine which model is appropriate for handling overfitting.

Considering the amount of data which is under 1000 rows, we implement the K-fold cross-validation method to handle overfitting. We use standard K=5 in the experiment and run standard statistical tools, such as F1-score, precision, recall and accuracy to assess both training and validation performance.

Experiment #1. In the first experiment, we examine the effect of pre-processing techniques on sentiment analysis. Pre-processing is the first step in sentiment analysis or other tasks related to text analyzing. This step is important to understand the data and it was proven that it can improve the model accuracy [25]. However, all of them are not appropriate to be implemented for a small dataset, so there is a need to understand which technique is more influential in increasing the sentiment-model performance.

There are eight pre-processing techniques implemented in this experiment; namely, lower case, punctuation, number and unicode removal, stop-word removal, slang-word normalization, character-repetition normalization, stemming and POS tagging. The detailed explanation for each technique is explained in section 2. In the first step, we design some scenarios by combining some pre-processing techniques into six cases. The six combinations of pre-processing techniques are presented in Table 6.

Case 1 represents a scenario without considering pre-processing techniques; the data proceed in this scenario is from original reviews. Case 2 only uses standard pre-processing techniques, such as lower case, punctuation, symbol removal and stop-word removal. Case 3 and Case 4 implement slang-word and character-repetition handling, respectively. Meanwhile, stemming is added in Case 5; in this experiment we used a standard stemming algorithm from the Sastrawi library. Finally, a complete technique version which uses POS-tagging filtering is employed in Case 6.

Table 6. Combination of pre-processing techniques.

Case	Pre-processing Technique							
	Lower	Punct.	Symbol	Stop	Slang	Repeat	Stem	POS
1	-	-	-	-	-	-	-	-
2	v	v	v	v	-	-	-	-
3	v	v	v	v	v	-	-	-
4	v	v	v	v	v	v	-	-
5	v	v	v	v	v	v	v	-
6	v	v	v	v	v	v	v	v

We implement various Naïve Bayes models in this experiment. The results shows that Complement Naïve Bayes achieves a good performance compared to other models. Table 7 presents F1-score as well as accuracy of Complement Naïve Bayes using variation cases in the fashion dataset. The highest F1-

[2] https://github.com/rizalespe/Dataset-Sentimen-Analisis-Bahasa-Indonesia

score, as well as accuracy, are shown in Case 4, amounting to around 0.87 and 88.08%, respectively. Meanwhile, the lowest scores are shown in Case 2 (F1=0.83, accuracy=84.2%) and Case 4 (F1=0.83, accuracy=84.8%).

Table 7. Experiment results using variation cases for fashion dataset.

Pre-processing	**Case**	**Accuracy (%)**	**F1-score**
No	1	85.58	0.843
Yes	2	84.23	0.832
	3	86.54	0.855
	4	88.08	0.870
	5	87.69	0.866
	6	84.81	0.833

We also analyzed the results of all variation cases implemented for all datasets. Table 8 depicts the average F1-score for each case and Figure 5 presents the trend of the results. From the results, we can see a stable score appearing in Cases 3-5 of approximately 0.82. Meanwhile, the trend shows a significant decrease of F1-score in Case 6 of around 0.7, where this score is the lowest F1-score of all cases. Based on our analysis, the decreased performance in Case 6 is caused by the selection process of some class words based on POS tagging. The class of words that are selected in this pre-processing phase are noun (NN), adjective (JJ) and negation (NEG). This process reduced the dimensions of data and affected data for small datasets significantly.

Table 8. The results of all datasets for each case.

Dataset	**Case**					
	1	2	3	4	5	6
cellular	0,785	0,785	0,800	0,800	0,800	0,697
cyber	0,820	0,822	0,855	0,855	0,855	0,770
movie	0,838	0,838	0,852	0,852	0,852	0,707
politic	0,730	0,728	0,743	0,743	0,742	0,627
fashion	0,843	0,832	0,855	0,870	0,866	0,833
average	0,803	0,801	0,821	0,824	0,823	0,727

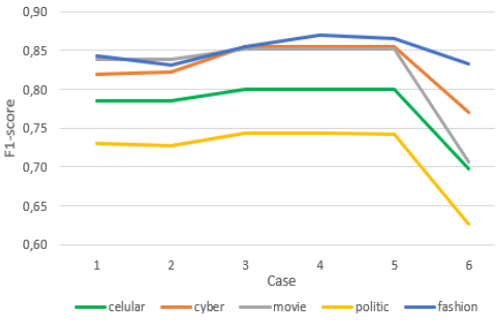

Figure 5. Average F1-score from all domain datasets for each case.

The data scraped from the marketplace causes the format to be unstructured, so pre-processing is needed to clean and prepare the data before analyzing. However, various pre-processing techniques are not appropriate to be implemented, especially for a small dataset. Slang-word and repetition handling and stemming are powerful to be employed. On the other hand, based on the experiment, the selection of words using POS tagging is not recommended for supervised learning with a small dataset, because it reduces the dimensions of data.

Experiment #2. We compare the sentiment-classification results from various types of Naïve Bayes models for the fashion dataset. The fashion dataset is a primary dataset used in the experiment (see Table 4 for the details of the primary data). Four different types of Naïve Bayes models are implemented in this experiment; namely, Gaussian, Multinomial, Complement and Bernoulli. The experiment is conducted to measure the performance of each Naive Bayes model in classifying sentiment sentences. We use the K-fold cross-validation method with K=5. K-fold cross-validation is a measurement method for both training and validation performance that is appropriate for small data. Table 9 presents the validation result for each model of Naïve Bayes in terms of F1-score, accuracy, precision and recall for the fashion dataset.

Table 9. Validation performance results of the second experiment.

Model	**F1**	**Prec**	**Rec**	**Acc(%)**
Gaussian	0.668	0.881	0.538	73.27
Multinomial	**0.876**	**0.919**	0.840	**88.18**
Complement	**0.876**	**0.919**	0.840	**88.18**
Bernoulli	0.847	0.839	**0.855**	84.55

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 08, No. 03, September 2022.

From Table 9, we can see that both Multinomial and Complement present the highest F1-score, precision and accuracy, while the highest recall is produced by Bernoulli. The highest F1-score, precision, accuracy and recall for the fashion dataset are 0.876, 0.919, 0.855 and 88.18%, respectively. Meanwhile, Gaussian has the lowest measurement results with an F1-score of around 0.668 and an accuracy of around 73.27%. The experiment results show that both Complement and Multinomial models have similar performances and are superior to other Naïve Bayes models for the fashion dataset.

In Section 2, we have explained how to gather fashion dataset from Indonesian marketplace; and in Figure 3, we present two groups of words based on sentiment labels which are visualized using cloud word. We note some unique keywords that relate to the experiment results. Some keywords, such as *material* and *size*, are usually followed by context-dependent opinions, such as *thin*, *thick*, *big* and *small*. The context-dependent opinion is an opinion which appears in several aspects with uncertain polarity [26]. The sentiment polarity of context-dependent opinions is caused by the domain of the dataset. For example, the word "*thin*" is positive when mentioned in the context of cellular or electronic products. However, this word will be negative if it appears in fashion. Context-dependent opinions can affect the sentiment-analysis task performance. Therefore, there is a potential of the existence of a concern on this issue for further study.

Experiment #3. In the final experiment, we examine the performance of various Naïve Bayes models using primary data as well as secondary data. The total dataset utilized in the third experiment is comprised of five datasets from various domains; namely, cellular, cyberbullying, movie, politic and fashion. Table 10 shows the average of each statistical measurement result from various sentiment-analysis datasets for both training and validation in the third experiment.

Table 10. Comparing the results of the models in training and validation.

Model	**Training**				**Validation**			
	F1	Prec	Rec	Acc	F1	Prec	Rec	Acc
Gaussian	**0,971**	0,991	0,960	97,53	**0,703**	0,748	0,684	72,09
Multinomial	**0,974**	0,983	0,966	97,51	**0,816**	0,870	0,782	82,79
Complement	**0,978**	0,983	0,973	97,83	**0,820**	0,851	0,798	82,59
Bernoulli	**0,964**	0,963	0,966	96,41	**0,799**	0,827	0,788	80,74

In general, the Complement model is dominant over the other models for both training and validation scoring results. In training, the highest scores for F1-score, recall and accuracy produced by the Complement model are 0.978, 0.973 and 97,83%, respectively. For precision, Gaussian gave the highest precision of around 0.991. In validation evaluation, Complement and Multinomial show excellent results compared to the other two models. The highest F1-score and recall are 0.820 and 0.798, respectively produced by the Complement model. Meanwhile, Multinomial presents higher scores for precision and accuracy of around 0.870 and 82.79% respectively.

From Table 10, we can assess which model has a good performance to handle overfitting. The consistency scores in training and validation can be an indicator of a model for overfitting issues. The Complement model shows the smallest gap in F1-score between training and validation from 0.978 to 0.820. Multinomial has a small distance from training of 0.974 to validation of 0.816. Bernoulli shows a higher F1-score in training of around 0.964, while in validation, the score is under 0.80. On the other hand, the Gaussian model presents a good performance in training above 0.90 for all measurement scores, but it produces the lowest scores under 0.75 in validation.

We compare the experiment results with the baseline. An increased F1-score of the model proposed is shown in cyberbullying dataset at around 0.856, while the baseline using Support Vector Machine produces an F1-score of 0.697 [21]. Meanwhile, the Complement Naïve Bayes for political dataset presents an increased accuracy of around 72,4% compared with the baseline of 70,2% using Multinomial Naïve Bayes [23]. For the cellular dataset, the baseline using Support Vector Machine presents a similar result to that of our model using Complement Naïve Bayes (F1-score=0.800) [20]. On the other hand, the baseline of the movie dataset that used Multinomial Naïve Bayes shows an F1 score=0.917, which is higher than that of the model proposed [22]. This inconsistent result is possibly caused because there is no consideration of cross-validation in the evaluation method of the baseline. The baseline of the movie dataset did not handle the overfitting issue in the experiment.

Referring to Table 11, the Complement model has the highest (average) F1-score at 0.820, followed by the Multinomial model at 0.816. Meanwhile, the Gaussian score is the lowest in performance beingaround 0.703. This result shows that both Complement and Multinomial have good performance in sentiment analyzing, especially to handle small datasets. On the other hand, Gaussian is not good enough to handle overfitting. In terms of *politic* dataset, this dataset is bigger than the others, but this has not increased the performance. Therefore, we can conclude that a lot of data is not enough in supervised learning, but it is important to know the variance as well as the characteristics of the data.

Table 11. Average F1-score of validation results for each NB model.

Dataset	Gaussian	Multinomial	Complement	Bernoulli
cellular	0.734	0.781	0.800	0.760
cyberbullying	0.769	0.856	0.856	0.860
movie	0.704	0.831	0.831	0.786
politic	0.643	0.737	0.737	0.740
fashion	0.668	0.876	0.876	0.847
Average	**0.703**	**0.816**	**0.820**	**0.799**

4. CONCLUSIONS

This research focuses on analyzing sentiment in the fashion domain from Indonesian review data using various Naïve Bayes models. Four different Naïve Bayes models are used in this research; namely, Gaussian, Multinomial, Complement and Bernoulli. From the experiment results, we have three findings: 1) Selection of words using POS tagging is not recommended for supervised learning with a small dataset, because it can reduce the dimensions of data. 2) Complement model is superior to other models, especially to handle overfitting. 3) There are opinion words which appear in several aspects with uncertain polarity called context-dependent opinions, which can affect the sentiment-analysis task performance. For future work, choosing a powerful stemming algorithm in pre-processing can be considered as possible to increase the model performance. Other than that, knowing data characteristics and domain is crucial. Further, it will be of importance to concern the study of context-dependent opinion issues in the next experiments.

ACKNOWLEDGEMENTS

This work is supported by Kementerian Pengajian Tinggi Malaysia, Fundamental Research Grant Scheme (FRGS) by code number FRGS/1/2020/ICT02/UMS/02/2.

REFERENCES

[1] O. Alqaryouti, N. Siyam, A. A. Monem and K. Shaalan, "Aspect-based Sentiment Analysis Using Smart Government Review Data," Applied Computing and Informatics, DOI: 10.1016/j.aci.2019.11.003, 2019.

[2] S. Ainin, A. Feizollah, N. B. Anuar and N. A. Abdullah, "Sentiment Analyzes of Multilingual Tweets on Halal Tourism," Tourism Management Perspect., vol. 34, no. Feb., p. 100658, 2020.

[3] K. M. O. Nahar, A. Jaradat, M. S. Atoum and F. Ibrahim, "Sentiment Analysis and Classification of Arab Jordanian Facebook Comments For Jordanian Telecom Companies Using Lexicon-based Approach and Machine Learning," Jordanian J. of Computers and Inf. Technol. (JJCIT), vol. 6, no. 3, pp. 247–262, 2020.

[4] H. Elfaik and E. H. Nfaoui, "Deep Bidirectional LSTM Network Learning-based Sentiment Analysis for Arabic Text," J. Intellgent Systems, vol. 30, no. 1, pp. 395–412, DOI: 10.1515/jisys-2020-0021, 2021.

[5] M. Gusti, "Ini Dia Nilai Transaksi Marketplace Indonesia 2020," Kompas TV, [Online], Available: kompas.tv/article/107064/ini-dia-nilai-transaksi-marketplace-indonesia-2020. (Accessed Jun. 02, 2021).

[6] A. Priadana and A. A. Rizal, "Sentiment Analysis on Government Performance in Tourism during the COVID-19 Pandemic Period with Lexicon Based," CAUCHY, vol. 7, no. 1, pp. 28–39, Nov. 2021.

[7] T. Sutabri, S. J. Putra, M. R. Effendi, M. N. Gunawan and D. Napitupulu, "Sentiment Analysis for Popular e-traveling Sites in Indonesia Using Naive Bayes," Proc. of the 6th Int. Conf. on Cyber and IT Service Management (CITSM), pp. 1–4, DOI: 10.1109/CITSM.2018.8674262, 2018.

[8] A. F. Akbar, A. B. Santoso, P. K. Putra and I. Budi, "A Classification Model to Identify Public Opinion on the Lockdown Policy Using Indonesian Tweets," J. Theor. Appl. Inf. Technol., vol. 99, no. 14, 2021.

[9] C. C. P. Hapsari, W. Astuti and M. D. Purbolaksono, "Naive Bayes Classifier and Word2Vec for Sentiment Analysis on Bahasa Indonesia Cosmetic Product Reviews," Proc. of the Int. Conf. on Data Science and Its Applications (ICoDSA), pp. 22–27, DOI: 10.1109/ICoDSA53588.2021.9617544, Oct. 2021.

[10] A. Rahmatulloh, R. N. Shofa, I. Darmawan and Ardiansah, "Sentiment Analysis of Ojek Online User Satisfaction Based on the Naïve Bayes and Net Brand Reputation Method," Proc. of the 9th Int. Conf. on Information and Communication Technology (ICoICT), pp. 337–341, 2021.

[11] Webretailer, "Online Marketplaces in Southeast Asia: A Unique Region for Ecommerce," 2020, [Online], Available: https://www.webretailer.com/b/online-marketplaces-southeast-asia/. (Accessed Jun. 26, 2021).

[12] U. Rhohmawati, I. Slamet and H. Pratiwi, "Sentiment Analysis Using Maximum Entropy on Application Reviews (Study Case: Shopee on Google Play)," JITEKI Journal, vol. 5, no. 1, pp. 44–49, 2019.

[13] E. Swandy, "Bahasa Gaul Remaja Dalam Media Sosial Facebook," J. Bastra, vol. 1, no. 4, pp. 1–19, 2017.

[14] L. Jing, H. Huang and H. Shi, "Improved Feature Selection Approach TFIDF in Text Mining," Proc. of the 1st IEEE Int. Conf. on Machine Learning and Cybernetics, pp. 4–5, Beijing, China, 2002.

[15] J. Chen, H. Huang, S. Tian and Y. Qu, "Feature Selection for Text Classification with Naïve Bayes," Expert Syst. Appl., vol. 36, no. 3 PART 1, pp. 5432–5435, DOI: 10.1016/j.eswa.2008.06.054, 2009.

[16] Shuo Xu, "Bayesian Naïve Bayes Classifiers to Text Classification," J. Information Science, no. 15, pp. 1–12, DOI: 10.1177/0165551510000000, 2016.

[17] D. H. Abd, A. T. Sadiq and A. R. Abbas, "Political Articles Categorization Based on Different Naïve Bayes Models," Proc. of the Int. Conf. on Applied Computing to Support Industry: Innovation and Technology (ACRIT 2019), vol. 1174, pp. 286-301, 2020.

[18] J. D. M. Rennie, L. Shih, J. Teevan and D. Karger, "Tackling the Poor Assumptions of Naive Bayes Text Classifiers," Proc. of 21st Int. Conf. on Machine Learning (ICML '04), vol. 2, no. 1973, pp. 616–623, 2003.

[19] V. Metsis, I. Androutsopoulos and G. Paliouras, "Spam Filtering with Naive Bayes - Which Naive Bayes?," Proc. of the 3rd Conf. on Email and Anti-Spam (CEAS 2006), [Online], Available: https://www2.aueb.gr/users/ion/docs/ceas2006_paper.pdf, 2006.

[20] U. Rofiqoh et al. "Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter Dengan Metode Support Vector Machine dan Lexion Based Feature," J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya, vol. 1, no. 12, pp. 1725–1732, 2017.

[21] W. Athira, I. Gholissodin and R. S. Perdana, "Analisis Sentimen Cyberbullying Pada Komentar Instagram Dengan Metode Klasifikasi Support Vector Machine," J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya, vol. 2, no. 11, pp. 4704–4713, 2018.

[22] P. Antinasari, R. S. Perdana and M. A. Fauzi, "Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 1, no. 12, pp. 1718–1724, 2017.

[23] A. R. T. Lestari, R. S. Perdana and M. A. Fauzi, "Analisis Sentimen Tentang Opini Pilkada DKI 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Näive Bayes dan Pembobotan Emoji," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 1, no. 12, pp. 1718–1724, 2017.

[24] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," J. Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[25] A. W. Pradana and M. Hayaty, "The Effect of Stemming and Removal of Stopwords on the Accuracy of Sentiment Analysis on Indonesian-language Texts," Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control, vol. 4, pp. 375–380, DOI: 10.22219/kinetik.v4i4.912, 2019.

[26] H. Kansal and D. Toshniwal, "Aspect Based Summarization of Context Dependent Opinion Words," Procedia Computer Science, vol. 35, pp. 166–175, DOI: 10.1016/j.procs.2014.08.096, 2014.

ملخص البحث:

يهدف هذا البحث الى فحص أداء نماذج مختلفة من مصنِّف (NB) في تحليل المشاعر، خصوصاً عند تطبيقها على مجموعات بيانات صغيرة؛ من أجل معالجة مشكلة فرْط المواءمة. النماذج الأربعة المستخدمة هي: النّموذج الغاوْسيّ، والنّموذج متعدّد الدّوالّ، والنّموذج المتمّم، ونموذج برنولي. كذلك تمّ تحليل أثر التّقنيات المختلفة للمعالجة القَبْليّة على أداء كلٍّ منْ تلك النّماذج. من ناحيةٍ أخرى، قُمنا ببناء مجموعة بياناتٍ للأزياء من الأسواق الإندونيسية، وهي الأولى من نوْعها في حقل الموضة. وهي متميّزة في خصائصها على شبيهاتها في الحقول الأخرى. كذلك استخدمنا عدداً من مجموعات البيانات في تجربة لفحص مصنِّفات (NB) ومقارنة أداء نماذجها المختلفة. واتّضح من النّتائج تفوُّق نموذج مصنِّف (NB) المتمّم على النّماذج الأخرى ، وبخاصّةٍ فيما يتعلّق بمعالجة مشكلة فرْط المواءمة محقّقاً درجة (F1) تصل الى (0.82) .

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).