Analysis of Cosmetic Product Opinions on E-Commerce Based on Naïve Bayes Classifier

Claretty Therence[1*], Jimmy Tjen[2]

[1] Department of Digital Business, Faculty of Information Technology, Universitas Widya Dharma Pontianak, Pontianak, Indonesia.
[2]Department of Informatics, Faculty of Information Technology, Universitas Widya Dharma Pontianak, Pontianak, Indonesia.
*Email:
[1] 22430149@widyadharma.ac.id
[2]jimmy.tjen@mathmods.eu

Abstract. This research employs Naïve Bayes classifier to analyze consumer sentiment in concealer reviews, utilizing a dataset of 400 reviews as training data, and testing it to 100 reviews of five different brands on their concealer product. The model's performance achieved an accuracy of 89.37% on training data and 73.75% on the testing data. A dictionary was created highlighting twelve frequently used words in the reviews, along with their frequencies and positive probabilities across five brands, offering insights for improvement. Additionally, the study presents a confusion matrix detailing precision and recall for each brand. The results indicate that Brand D performed the best, followed by Brands B, C, E, and A, providing recommendations for enhancing customer satisfaction based on sentiment analysis.

Keywords: cosmetic product; customer reviews; Naïve Bayes; Sentiment Analysis.

Introduction

The transformation of interaction with products and services has shifted significantly with the emerge of e-commerce during the digital era (Paredes-Corvalan, Pezoa-Fuentes, Silva-Rojas, Rojas, & Castillo-Vergara, 2023). This phenomenon also extends to cosmetic products, with platforms like Shopee playing a prominent role in Indonesia. The openness in sharing reviews of cosmetic products, particularly concealers on Shopee, highlights the importance of sentiment analysis. This analysis is essential for gaining insight into customer opinions, significantly influencing the purchasing decisions of prospective buyers and insights to develop products and precise marketing strategies (Liu, Chen, & Liu, 2022). This approach aligns with the goals of SDG 12, which supports producers in developing sustainable products and encourages responsible consumption. It also aligns with SDG 9, which emphasizes inclusive infrastructure, industrialization, and fostering innovation. The analysis utilizes machine learning and big data to access customer reviews of concealer products available on Shopee in Indonesian.

Sentiment analysis is also known as text mining (Wankhade, Rao, & Kulkarni, 2022), is designed to extract subjective insight that help businesses to understand public sentiment regarding their product or services while tracking online opinions (Daza, Rueda, S{\'a}nchez, Esp{\'\i}ritu, & Qui{\~n}ones, 2024). Recent advancement in machine learning technique have significantly improved their effectiveness as valuable tools for conducting more detailed investigations (Park & Woo, 2019). Extensive research has been conducted on this topic. This research provided a comprehensive overview of approaches, trends, and challenges in sentiment analysis, aiming to deliver an in-depth understanding of the field and its associated areas (Birjali, Kasri, & Beni-Hssane, 2021). Research highlighting an overview of sentiment analysis techniques, focusing on their application in the e-commerce sector to analyze consumer opinions and improve business operations, emphasizing methods such as lexicon-based and supervised machine learning approaches (Marong, Batcha, & Mafas, 2020). Study conducted on sentiment analysis on Twitter data regarding the 2019 indonesian presidental candidates to positive and negative sentiment, with result showing Naïve Bayes outperformed SVM and K-NN in RapidMiner tests (Wongkar & Angdresey, 2019).

Several studies have been conducted in recent years on sentiment analysis across various topics, exploring its applications and methodologies. This study focuses on sentiment analysis of product reviews on Indonesian marketplace, utilizing Natural Langauge Prosessing (NLP) for text preprocessing and comparing machine learning algorithms between Naïve Bayes for unigram datasets and K-Nearest Neighbors (KNN) for bigram datasets, offering beneficial insights for product enchancement and buyer guidance (Rohman,

Musyarofah, Utami, & Raharjo, 2020). Study employs sentiment analysis with Naïve Bayes classifer to analyze Facebook comments about Thai agritech startups providing valuable insight into public and investor opinions, consumer behavior and marketing trends (Kewsuwun & Kajornkasirat, 2022). Research develops sentiment dictionary and applies Naïve Bayes to analyze danmaku video comments, enabling sentimen classification and visualization to monitor emotional trends, predict video popularity, and achieve effective sentiment polarity detection (Li, Li, & Jin, 2020). This study provided researched aimed to classify user review on Fund Application using sentimen analysis, and achieving a high accuracy (84.76%) giving insight to enhance the app by addressing negative feedback (Surohman, Aji, Rousyati, & Wati, 2020).
The structure of this paper is as follows: in Method section, we describe the data collection process, the development of the sentiment dictionary, and the application of the sentiment analysis model, along with the use of evaluation metrics. In the Result and Discussions section, presents model performance, dictionary, matrix and metrics across different brands, and practical insights derived from the analysis, followed by a discussion comparing the results with existing methods and highlighting potential applications. In Conclusion, we finalized the main result finding from this research.

Methods

This section outlines a systematic approach to sentiment analysis of consumer reviews of concealer product, detailing the steps of data collection and labeling, data preprocessing, model development, performance assessment, and insights and application. The final model will be used to compare brands and its model's performance accuracy. We provide the framework in the following structured format:

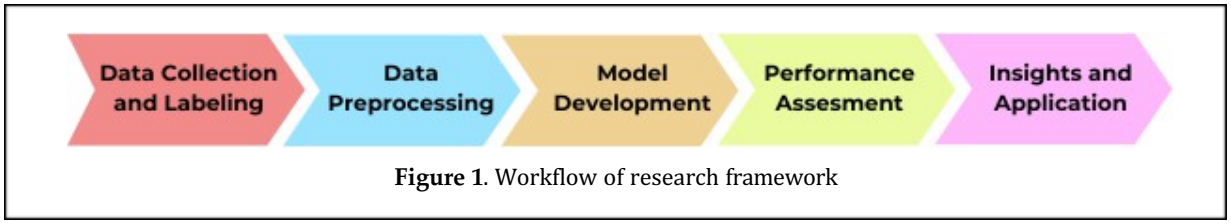

Figure 1. Workflow of research framework

1. **Data Collection and Labeling**
   A total of 600 user-generated reviews from various concealer brands were collected from Shopee. The reviews were obtained using a combination of automated web scraping and manual validation to ensure data quality. The finalized dataset was classified into three categories: "Positive", "Negative", and "Neutral", based on sentiment conveyed in the reviews.
2. **Data Preprocessing**
   The collected dataset was processed to prepare the textual data for analysis, following a specific sequence of steps. Initially, text cleaning was performed, which involved converting all text to lowercase and removing punctuation, special characters, and unnecessary symbols to emphasize meaningful content (He, Zhou, & Zhao, 2022). Following this, reviews were tokenized into individual words, enabling deeper analysis; this was accomplished through a function that split each string into a list of tokens, stored in a new array called "token". An example of the tokenization results includes tokens such as {'i', 'like', 'this', 'product'} and {'too', 'cracky'}. To further refine the analysis, a dictionary of unique words was created to identify the dominant terms within the reviews, and frequency analysis was conducted to highlight key terms associated with each sentiment category.
3. **Model Development**
   With the preprocessed dataset prepared, the next step involved developing and evaluating a sentiment analysis model. The dataset was split into training and testing sets, allocating 80% of the data for training and the remaining 20% for testing. The model will then be used to test five different concealer Brands. A classification model was trained on the annotated training dataset to predict sentiment labels of "Positive," "Negative," and "Neutral."

4. **Performance Assessment**
   The performance of the trained model was evaluated by testing it on the test set, allowing for the determination of its accuracy. The assessment of the model's performance centered on its ability to

accurately classify sentiments, utilizing a comprehensive set of evaluation metrics. Accuracy was defined as the proportion of accurately predicted labels relative to the total number of predictions, serving as an indicator of the model's overall reliability. In addition to accuracy, a confusion matrix was used to visualize classification errors, offering insights into specific areas where the model may have faced challenges. This multifaceted approach to evaluation was essential for understanding the model's strengths and weaknesses, facilitating iterative improvements and greater precision in the future analyses.

5. **Insights and Application**

   Alongside model evaluation, a dictionary of frequency occurring words was compiled to identify the most common themes in user feedback, with dominant words extracted for each sentiment category. For positive sentiments, words reflecting satisfaction, quality, or a favorable experience were emphasized, revealing the attributes consumers value in concealers. Conversely, negative sentiments were marked by expressions of dissatisfaction, complains, or adverse experiences, clarifying potential product shortcomings. Additionally, neutral sentiments included words that conveyed ambiguity or a lack of strong opinions, indicating areas where consumer preferences may be unclear. This through analysis enhance the interpretation of the user sentiments as well as offering valuable insights into the characteristics associated with each sentiment category, ultimately enhancing the overall sentiment analysis process and informing product development strategies.

This robust framework not only enhances the understanding of customer towards concealer products, as well as serving valuable tool for guiding future product development and marketing strategies.

Result and Discussion

This section provides an overview of the result from the sentiment analysis using Naïve Bayes classifier, presenting four main discussion: text processing, model's performance, dictionary words, and the confusion matrix, precision, and recall for each brand. The finalized combination off all will provide information about each brands strength as well as weakness and area for improvement based on each words.

1. Text Preprocessing

This section gives an example result of the text cleaning and tokenizing.

Table 1: Example result of text preprocessing

	Before	**After**
Text Cleaning	'Suka teksturnya ringan gitu. Ga begitu kental dan gampang diblend, warnanya juga cakep 🤗'	'suka teksturnya ringan gitu ga begitu kental dan gampang diblend warnanya juga cakep'
Tokenizing	'suka teksturnya ringan gitu ga begitu kental dan gampang diblend warnanya juga cakep'	'suka' 'teksturnya' 'ringan' 'gitu' 'ga' 'begitu' 'kental' 'dan' 'gampang' 'diblend' 'warnanya' 'juga' 'cakep'

Table 1 outlines the example of text cleaning and tokenizing process. The "Before" column shows the original sentence before being cleaned and being tokenized, while the "After" column breaks it down to individual cleaned and added tokens, enhancing clarity and usability for analysis.

2. Model's Performance

The model's performance is evaluated using its test confusion matrix and accuracy metrics, as illustrated below:

Table 2: Confusion Matrix

Confusion Matrix		Predicted			Recall
		Positive (63)	Negative (13)	Neutral (4)	
True	Positive (57)	50	4	3	0.877
	Negative (11)	3	8	0	0.727
	Neutral (12)	10	1	1	0
Precision		0.793	0.615	0	

Table 3: Model's Accuracy

Data	Accuracy
Test	73.75%
Train	89.37%

Based on table 2, a total of 400 dataset, 320 (80%) are allocated for training data and 80 (20%) for testing, divided into three categories. Among 57 positive reviews, 50 were accurately predicted as positive, out of 11 negative reviews, 8 are predicted correctly negative, and out of 12 neutral reviews, 1 is predicted correctly positive. Overall, positive precision up to 79.3%, indicating the model ability to accurately make positive prediction 79.3% out of 100%. The model has negative precision up to 61.5%, meaning the model has ability to correctly make negative prediction 61.5% out of 100%. This model has positive recall of 87.7%, meaning the model has ability to correctly identify 87.7% of all actual positive instances. The model has negative recall of 72.7%, indicating that it correctly identifies 72.7% of all actual negative instances. According to Table 3, the accuracy for the test data, which reflects the model's performance on unfamiliar data, is 73.75%. In contract, the accuracy for the training data, which consists of data the model has already encountered, is 89.37%.

3. Dictionary Words

In this section, we represent the findings of curated list of domain-specific words that emerged from customer reviews, reflecting key aspect of user sentiment.

Table 4.: Frequent Words and Positive Probability (%)

No	Word	Brand A		Brand B		Brand C		Brand D		Brand E	
		Freq.	Positive	Freq.	Positive	Freq.	Positive	Freq.	Positive	Freq.	Positive
1	'Bagus'	124	79%	114	80%	135	85%	145	83%	132	79%
2	'Cocok'	96	78%	105	82%	91	78%	108	85%	101	74%
3	'Concealer'	93	76%	106	85%	104	82%	110	86%	90	74%
4	'Coverage'	59	73%	57	80%	64	79%	70	83%	64	70%
5	'Kulit'	77	81%	90	86%	82	81%	89	86%	84	77%
6	'Packaging'	49	70%	52	78%	55	76%	61	81%	51	75%
7	'Pas'	112	67%	107	76%	103	74%	110	77%	110	69%
8	'Ringan'	39	90%	48	94%	50	92%	53	94%	39	89%
9	'Sesuai'	47	77%	50	85%	45	83%	46	83%	48	79%
10	'Shade'	88	77%	93	83%	88	82%	107	87%	91	76%
11	'Suka'	51	77%	55	87%	55	88%	60	85%	50	84%
12	'Warna'	51	69%	57	78%	49	74%	52	77%	53	68%

Based on Table 4, the frequent words indicate varying consumer sentiments as well as preferences among the five concealer brands and probability of positive polarity of each word. The word 'Bagus' (meaning "good") is the most frequently mentioned across all brands, with Brand D receiving the highest count (145) and positive association of 83%. Following closely is the word 'Cocok' (meaning "suitable"), with Brand B at 105 mentions and positive association of 82%. The word 'Concealer' ranks third in frequency, with Brand C at 104 mentions with positive association of 82%. Important term 'Coverage' appears frequently, particularly for Brand D with 70 mentions and has a positive association of 83%. The word 'Shade' in Brand D has 107 mentions and a positive association of 87%. The word 'Kulit' (meaning "skin") has the most mentions in Brand B (90) with positive association of 86%.

Table 4 reveals not only the frequency of words associated with each brand but also consumer sentiments and potential areas for improvement. For instance, the word 'Packaging' in Brand A shows the lowest mention (49) and positive association (70%) compared to the other brands, suggesting consumers appreciate its current design but may desire enhancements that further elevate the product's appeal. The word 'Shade' in Brand B is frequently mentioned (93), indicating a strong consumer preference for more variety and options in undertones, which could help attract a broader customer base. Meanwhile, the word 'Coverage' in Brand E is quite frequently mentioned (64) but has the lowest positive association (70%) compared to other brands, indicating a promising opportunity for Brand E to refine its formulation to meet consumer expectation for quality. This information provide source for product improvement in many ways and precise marketing strategy.

4. Confusion Matrix, Precision, and Recall

To evaluate classification model's performance, we utilize the confusion matrix to summarize the model's predictions alongside key metrics: precision dan recall. Providing information on the actual polarity and the predicted polarity from each of the probability for accessing the accuracy of the model's performance. This

metrics provides comparison of each brands of concealer product. The equation used for this metrics is presented like below:

$$Precision = \frac{True\ Positive\ (TP)}{TP + False\ Positive\ (FP)} \quad (1)$$

$$Recall = \frac{True\ Positive\ (TP)}{TP + False\ Negative\ (FN)} \quad (2)$$

Polarity evaluated in precision and recall are "Positive" and "Negative". Neutral polarity is excluded from precision and recall calculation for not significantly contribute to the sentiment classifier (Aoumeur, Li, & Alshari, 2023). Precision is determined by dividing correct predictions by total number of predictions made as seen in Equation (1), while recall is calculated by dividing the number of correct predictions by total number of actual examples of each polarity as seen in Equation (2) (Shaikh & Deshpande, 2016).

Table 5: Confusion Matrix Concealer Brand A

Confusion Matrix		Predicted			Recall
		Positive (66)	Negative (27)	Neutral (7)	
True	Positive (28)	27	0	1	0.964
	Negative (55)	24	25	6	0.445
	Neutral (17)	15	2	0	0
Precision		0.409	0.926	0	

Based on Table 5, Brand A with 100 reviews divided into three categories. 28 positive, 55 negative, and 17 neutral. Of 28 positive reviews, 27 were precisely predicted as positive, 1 as neutral and 0 as negative. Among 55 negative critique, 25 were exactly identified as negative, 24 as positive, and 6 as neutral. From 17 neutral feedback, none were rightly classified as neutral, while 15 were predicted as positive and 2 as negative. Brand A has an accuracy of 52%.

Table 6: Confusion Matrix Concealer Brand B

Confusion Matrix		Predicted			Recall
		Positive (89)	Negative (7)	Neutral (4)	
True	Positive (78)	74	3	1	0.949
	Negative (1)	1	0	0	0
	Neutral (21)	14	4	3	0
Precision		0.831	0	0	

Based on Table 6, Brand B with 100 reviews categorized as 78 positive, 1 negative, and 21 neutral. Of 78 positive reviews, 74 were correctly predicted as positive, 3 as negative, and 1 as neutral. The single negative review was incorrectly predicted as positive. Among 21 neutral reviews, 4 were correctly predicted as neutral, while 14 were predicted as positive and 4 as negative. Brand B has an accuracy of 77%.

Table 7: Confusion Matrix Concealer Brand C

Confusion Matrix		Predicted			Recall
		Positive (92)	Negative (4)	Neutral (4)	
True	Positive (75)	73	1	1	0.973
	Negative (6)	4	2	0	0.333
	Neutral (19)	15	1	3	0
Precision		0.793	0.5	0	

Based on Table 7, Brand C with 100 reviews categorized into 75 positive, 6 negative, and 19 neutral. Out of 75 positive reviews, 73 were correctly predicted as positive, 1 as negative, and 1 as neutral. Among the 6 negative reviews, 2 were correctly identified, while 4 were incorrectly predicted as positive. In existing 19 neutral reviews, 3 were accurately identified as neutral, 15 were predicted as positive, and 1 as negative. Brand C has an accuracy of 78%.

Table 8: Confusion Matrix Concealer Brand D

Confusion Matrix		Predicted			Recall
		Positive (89)	Negative (3)	Neutral (8)	
True	Positive (86)	82	1	3	0.953
	Negative (1)	0	1	0	1
	Neutral (13)	7	1	5	0

Precision	0.921	0.333	0	

Based on Table 8, Brand D has 100 reviews categorized into 86 positive, 1 negative, and 13 neutral. Out of 86 positive reviews, 82 were correctly predicted as positive, 1 as negative, and 3 as neutral. The single negative review was correctly identified, with no incorrect predictions. Among the 13 neutral reviews, 5 were correctly identified as neutral, while 7 were predicted as positive and 1 as negative. Brand D has an accuracy up to 88%.

Table 9: Confusion Matrix Concealer Brand E

Confusion Matrix		Predicted			Recall
		Positive (75)	Negative (18)	Neutral (7)	
True	Positive (42)	37	2	3	0.881
	Negative (33)	18	14	1	0.424
	Neutral (25)	20	2	3	0
Precision		0.493	0.777	0	

Based on Table 9, Brand E has 100 reviews categorized into 42 positive, 33 negative, and 25 neutral. Counted among 42 positive reviews, 37 were correctly forecasted as positive, 2 as negative, and 3 as neutral. Among 33 negative feedback, 14 were rightly predicted as negative, 18 were incorrectly predicted as positive, and 1 as neutral. For 25 neutral feedback, 3 were accurately identified as neutral, 20 were predicted as positive and 2 as negative. Brand E has an accuracy of 54%.

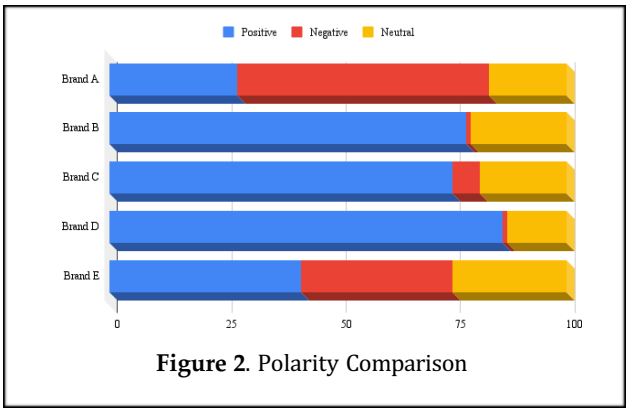

Figure 2. Polarity Comparison

Every Brand has 100 total reviews, categorized into positive, negative, and neutral. Based on Figure 2 and Table 3.1, Brand D stands out with the highest proportion of positive feedback, receiving 86 positive reviews, particularly highlighting the word 'Coverage' mentioned 70 times and 83% positive probability, indicating strong customer satisfaction. Brand B performs well with 78 positive reviews, emphasizing 'Shade', which has 93 mentions with 83% positive probability; these frequent mentions likely resonate with variety of the shades available. Brand C has total 75 positive review, with the word 'Cocok' mentioned 91 times and 78% positive probability, indicating customer feels the concealer is suitable for their needs. Meanwhile, Brand E has 42 positive review, with the word 'Kulit' at 84 mentions and 77% positive probability, the lowest compared to other brands, suggesting improvements are needed in related to the appearance of the skin and the formulation of the concealer. Lastly, Brand A has the lowest positive review, at 28, with word 'Packaging' mentioned 49 times and 70% positive probability, the lowest among the brands; likely due to dissatisfaction with its design and usability that effects product experience.

Conclusion

This analysis unitized Naïve Bayes classifier on a single training data ser to evaluate sentiment across five different brands of concealer product. Sentiment analysis model use training dataset of 400 reviews, with a testing set comprising 100 reviews per brand across five different brands. The model achieved accuracy of 89.37% on the training data and 73.75% on the testing data. We identified twelve frequently used words in concealer reviews, along with their positive probabilities, which offered insights for each brand on areas for improvement. The analysis of the confusion matrix revealed the precision and recall for each brand, concluding that Brand D performed the best, followed by Brand B, C, E, and A in terms on sentiment result, thus providing valuable recommendation for enhancing customer satisfaction.

References

Aoumeur, N. E., Li, Z., & Alshari, E. M. (2023). Improving the polarity of text through word2vec embedding for primary classical arabic sentiment analysis. *Neural processing letters, 55*(3), 2249--2264.

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems, 226*, 107134.

Daza, A., Rueda, N. D., S{\'a}nchez, M. S., Esp{\'\i}ritu, W. F., & Qui{\~n}ones, M. E. (2024). Sentiment Analysis on E-Commerce Product Reviews Using Machine Learning and Deep Learning Algorithms: A Bibliometric Analysisand Systematic Literature Review, Challenges and Future Works. *International Journal of Information Management Data Insights, 4*(2), 100267.

He, H., Zhou, G., & Zhao, S. (2022). Exploring e-commerce product experience based on fusion sentiment analysis method. *IEEE Access, 10*, 110248--110260.

Kewsuwun, N., & Kajornkasirat, S. (2022). A sentiment analysis model of agritech startup on Facebook comments using naive Bayes classifier. *International Journal of Electrical \& Computer Engineering (2088-8708), 12*(3).

Li, Z., Li, R., & Jin, G. (2020). Sentiment analysis of danmaku videos based on na{\"\i}ve bayes and sentiment dictionary. *Ieee Access, 8*, 75073--75084.

Liu, H., Chen, X., & Liu, X. (2022). A study of the application of weight distributing method combining sentiment dictionary and TF-IDF for text sentiment analysis. *IEEE Access, 10*, 32280--32289.

Marong, M., Batcha, N. K., & Mafas, R. (2020). Sentiment analysis in e-commerce: A review on the techniques and algorithms. *4*(1), 6.

Paredes-Corvalan, D., Pezoa-Fuentes, C., Silva-Rojas, G., Rojas, I. V., & Castillo-Vergara, M. (2023). Engagement of the e-commerce industry in the US, according to Twitter in the period of the COVID-19 pandemic. *Heliyon*.

Park, S., & Woo, J. (2019). Gender classification using sentiment analysis and deep learning in a health web forum. *Applied Sciences, 9*(6), 1249.

Rohman, A. N., Musyarofah, R. L., Utami, E., & Raharjo, S. (2020). Natural Language Processing on Marketplace Product Review Sentiment Analysis. *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)* (pp. 1--5). IEEE.

Shaikh, T., & Deshpande, D. (2016). Feature Selection Methods in Sentiment Analysis and Sentiment Classification of Amazon Product Reviews. *International Journal of Computer Trends and Technology, 36*, 225-230.

Surohman, S., Aji, S., Rousyati, R., & Wati, F. F. (2020). Analisa Sentimen Terhadap Review Fintech Dengan Metode Naive Bayes Classifier Dan K-Nearest Neighbor. *EVOLUSI: Jurnal Sains dan Manajemen, 8*(1).

Taboada, M. (2016). Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics, 2*(1), 325--347.

Wankhade, M., Rao, A. C., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review, 55*(7), 5731--5780.

Wongkar, M., & Angdresey, A. (2019). Sentiment analysis using Naive Bayes Algorithm of the data crawler: Twitter. *2019 Fourth International Conference on Informatics and Computing (ICIC)* (pp. 1--5). IEEE.