SENTIMENT ANALYSIS ON COSMETIC PRODUCT IN SEPHORA USING NAÏVE BAYES CLASSIFIER

FADLY[1,3], TRI BASUKI KURNIAWAN[2,*],
DESHINTA ARROVA DEWI[3], MOHD ZAKI ZAKARIA[4],
NAZZATUL FARAHIDAYAH BINTI MOHD NAZZIRI[4]

[1]Departement of Pharmacy, Politeknik Kesehatan Kemenkes, Palembang, Indonesia and
INTI International University, Malaysia
[2]Faculty of Computer Science, University of Bina Darma, Palembang, Indonesia
[3]Faculty of Data Science and Information Technology, INTI International University, Malaysia
[4]Faculty of Computer & Mathematics Sciences, University Technology Mara, Malaysia
*Corresponding Author: tribasukikurniawan@binadarma.ac.id

Abstract

At present, digital communication and data have become the higher use, and expressing their message through reviews and many more. The cosmetics industry has developed into a place where every business and sector competes to market and enhance its brand. Sephora, one of the biggest cosmetic industries, has higher sales and promotion, and its website has more reviews than it could ever get. The consumer can access the reviews and view them to give their opinion. The user's opinion can be predicted to know the positive and negative. It brought us to sentiment analysis as the focus of research to see the review. The Nave Bayes classifier, which is automatically pre-processed using natural language processing, is modelled. In building the model, the process goes through data collection, such as the reviews from each product brand. Then the pre-processing is done to get the bag of words trained in the Naïve Bayes model. The data have been trained with different split ratios and the number of iterations to find the highest accuracy. Then the data will be fine-tuning to get higher accuracy results to measure the prediction. As the model goes through, the visualization shows the prediction data. As a result, the Naïve Bayes model showed 94.7% accuracy after measuring using the cross-validation technique.

Keywords: Consumer behaviour, Consumption, Cosmetic product, Naive Bayes, Production, Sentiment analysis, Sephora,

1. Introduction

In the world of beauty, Sephora is the most popular company that sells many products from around the world [1]. It is the most inspiring company that inspires people around the world. The store opened in France in 1970; Sephora is known for its luxury branding as it on the biggest LVMH company. Now Sephora operates more than 2,300 stores worldwide and will expand more around the world of beauty [2]. As the beauty world becomes more trending and a must in many people's lives, Sephora needs to compete with the local stores as this product also has its store and platform, such as a website. Many websites sell products such as Nar's, Tarte, and Urban Decay. They have a store to sell their product. What makes Sephora different is that Sephora offers luxury option and how Sephora treat its customers [3].

In the beauty world, Sephora offers more luxury options, from the fragrances from the most popular brands, such as Channel and Chloe, to the Naked eye shadow pallet from Urban Decay and the most popular Yves Saint Laurent's for its gloss. Sephora also markets its product collection called Sephora Collection. Sephora offers luxury brand options such as Chloe, Urban Decay, Yves Saint Laurent, and many more. Sephora also brings in products from celebrities with their beautiful lines, such as Kat von D, Rihanna Fenty Beauty, Huda Beauty, and many more. Sephora also has its website, launched in 1999, and the customer can view and review it at Sephora.com [1].

Sephora's website provides many products and services to its customers. It is not the only platform for the end user to buy the product; it is also a platform for the customer to learn about each product and more about beauty. It is also a platform where all customers can learn more and learn how to make up and news about fashion and beauty products. From this website, customers can review all the products. The review helps Sephora take information from the user to analyse each product [4].

This review will help the analyser to make a sentiment analysis and allow the end user and Sephora itself to tack down each product and help in the sale of Sephora. As the data analyser does that, it tracks down all the reviews and analyses all the positive and negative reviews [5].

1.1. Cosmetic

Now in the world of beauty, many things can help enhance our appearances, such as Botox, plastic surgery, and many more. Directly the more comprehensive the cosmetic sometimes harms them. They then spend more money on the surgery to get a great look and get good compliments. With the influencer's influence, it becomes more aggressive for the user. With the broader use of technologies, the world has become more fearful, especially in makeup. In Malaysia, we are concerned about this harmful product because this product can cause cancer. We in Malaysia emphasize the good is Halal and shariah compliant [6].

In early 1900, they introduced the use of powder in women's lives as the powder is skin colour powder. They started to establish more product base makeup for women stated in Europe. In this era, they introduced essential product for different skin types, from oil-based to regular-based products. We are also being introduced to the "Eyelash curling instrument." Popular products like Lancôme and L'Oréal were invented at this time, significantly impacting Paris.

1.2. Industry

In the growth of the beauty industry, we can see that the number has grown over the years. Figure 1 shows that the Asia Pacific has the highest use of cosmetic products compared to others. Its growth steadily throughout the years.

With the help of social media, the world of beauty growth rapidly. The non-western county like Asia-pacific has a vast influence now as we can think. The beauty world in Asia, such as Korea, has grown quickly and significantly impacted the society involved. The beauty of Korea has more influence now in beauty from the product through the appearance. The diversity of products that gives such power makes the Korean trend proliferate. The beauty industry has rapidly grown to men's and natural/organic cosmetics. As this world evolves, consumers can choose their favourite product. Now we can see many beauty products from around the world.
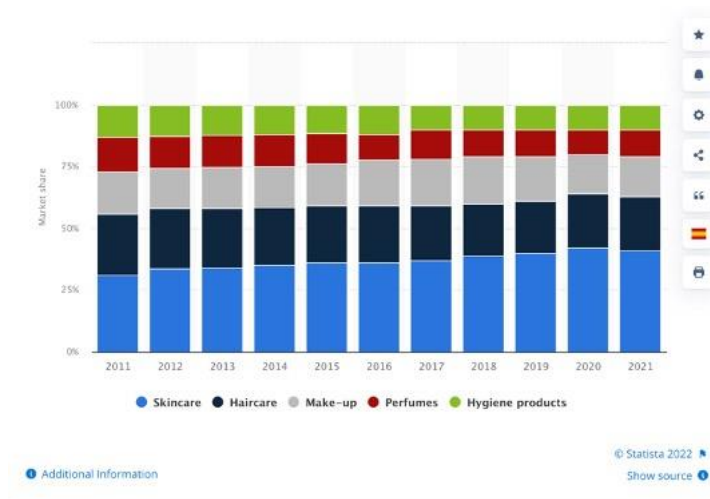

**Fig. 1. Breakdown the cosmetic market
worldwide from 2011 to 2022 by product category [7].**

1.3. Sephora online

Sephora's website (sephora.com) was launched in the United States in years 1999, and the website serves as Sephora's largest retail cosmetics store in North America [7]. Each county has its own Sephora website, such as Malaysia has Sephora MY. When considering the product selection, the customer can also shop online, where they can get the product with the click of a mouse. To satisfy their Sephora website, they have a column where their customer can give a review on any product. They also can provide a rating of the product. Sephora takes online reviews seriously to know their customer satisfaction with the product. The sale shifts online retailers will figure out how to handle online and in-store sales to give the best to their customers [8].

Figure 2 shows the site Sephora shows their customer rating and review of a specific product. 1–5 scale of the star represents the product's terrible and the excellent [9]. This cooperation of the shopper will help the shopper market enhance customer satisfaction and shopping behaviour [10].

Fig. 2. Sephora review.

2. Material and Methods

The research framework is fundamental to the big picture of the project development in this system. It correctly shows each research phase, such as preliminary study, knowledge acquisition, data collection, knowledge representation, system design, system development and documentation. Figure 3 shows this research framework to accomplish all the research objectives. The purpose is to show the research methodology done in describing the process of the data collection and how the data has been gathered.
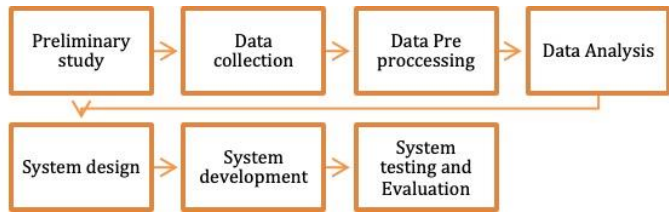

Fig. 3. Research framework.

2.1. Preliminary study

The research domain, approach, and problem in cosmetics are all studied in the preliminary study. At this point, extensive study has been done to establish the best method for the analysis. Online sources for research, including research papers, journal articles, conference papers, books, and other resources, have been used to expand their reach. The researcher is done in knowing why many people use cosmetics, how they choose cosmetics and other questions that need to be answered.

2.2. Data collection

For the data collection phase, data were collected by scraping the data from the Sephora website at www.sephora.com using Data miner. This google extension helps collect data, as shown in Figures 4 and 4. It has been managed is the comment and the recommended data that helped in the process of labelling the data.
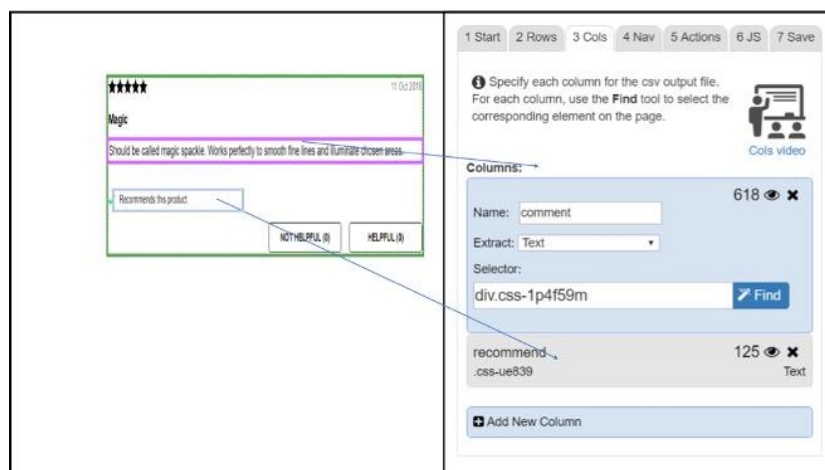
Fig. 4. Review of product from Sephora.

Fig. 5. The column selected for data collection.

Figure 4 shows how the data has been scraped using Data miner by selecting the row and columns of the data needed. From Fig. 5, the green box is the selected row, and the purple and blue are the selected columns. Then the chosen data were stored neatly in an Excel file to be used in the prediction later.

2.3. Data pre-processing

After the data has been retrieved, the process of cleaning happens. It checks the unfamiliar characters that cannot be recognized. The unwanted attributes, stop words, and connective words must be clean. Pre-processing data involves data cleaning, integration, transformation, and reduction.

2.4. Tokenization

Tokenization helps split big data into small and helps in the next step of the process. Tokenization has two (2) modules: word tokenize, and sentence tokenize. In this process, the tokenization of a word is used to split a sentence into words. The review was in long comments, and tokenization was used to convert all the words to the data frame for further process.

2.5. Stemming, lemmatization and stop word

Stemming is the process of changing the words to root words or based words to make the process more understandable for the procedure. The words that contained (-ed, -ing, -s, -ed, mis-) are the words that need to be changed.

2.6. Bag of words

Bag of words helps to extract features in the dataset. The process of transforming the data into matrix form counts the number of times the words appear. With Scikit-learn, it provides a method, Count Vectorizer, to count the number of words. Then create a bag of word that contain the number of words.

3. Result and Discussion

This chapter explains in detail the results that the project has evaluated in this project. It includes all the findings that have been done to the development of the system.

3.1. Data sampling

The data is sampled in the form of a graph chart on the number of datasets from 10 different brand products. The data is represented in Fig. 6.
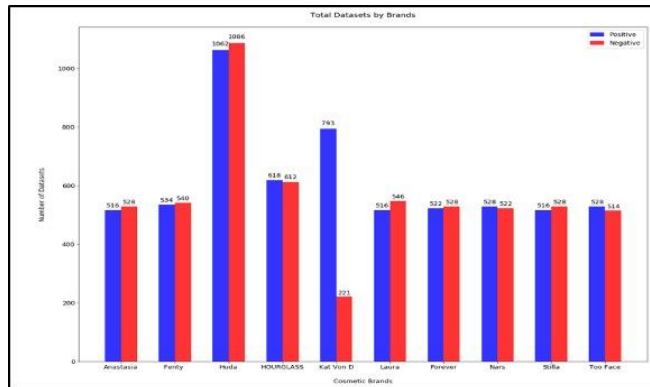

Fig. 6. Data representation.

We can see that the size of the data of each document has a different amount of data. Huda Beauty (on the 4th item in the table) has the highest positive review compared to the other brands. It has the highest number of data sets compared to others. Kat Von D (on the 5th item in the table) has more positive reviews than negative ones. It shows that Huda Beauty is a good product as it shows the highest reviews, as shown in Table 1.

The case of Kat Von D shows the many comparisons between negative and positive reviews. It will affect the performance of the data as positive reviews are highest than negative reviews. It will affect the accuracy of the data and can show overfitting data. As the data have been clean, this shows the overfitting of data in Fig. 7 as it shows the data imbalance. It shows that the data is classified more on

the positive side than the negative area. All the data are sampled to have similar data and utilized for data training and testing, as shown in Table 2.

Table 1. Total data of each brand.

Brands	Positive Reviews	Negative Reviews	Total Reviews
Anastasia Beverly Hills	516	528	1044
Fenty Beauty by Rihanna	534	540	1074
Hourglass	618	612	1230
Huda Beauty	1062	1086	2148
Kat Von D	793	221	1014
Laura Mercier	516	546	1062
Make Up For Ever	522	528	1050
Nars	528	522	1050
Stila	516	528	1044
Too Faced	528	514	1042

```
Documents name : KAT VON D Lock-It Foundation.csv.csv

Accuracy : 0.797
Iteration number : 3

Confusion Matrix :

Predicted  Negative  Positive  All
Actual
Negative          2        61   63
Positive          1       241  242
All               3       302  305

Classification Report :

            precision   recall  f1-score   support

   Negative      0.67     0.03      0.06        63
   Positive      0.80     1.00      0.89       242

avg / total      0.77     0.80      0.72       305
```

Fig. 7. Overfitting data.

Table 2. Data resample.

Brands	Positive Reviews	Negative Reviews	Resample Required	Total Reviews
Anastasia Beverly Hills	528	528	Positive	1044
Fenty Beauty by Rihanna	534	534	Negative	1074
Hourglass	612	612	Positive	1230
Huda Beauty	1086	1086	Negative	2148
Kat Von D	221	221	Positive	1014
Laura Mercier	546	546	Negative	1062
Make Up For Ever	522	522	Negative	1050
Nars	522	522	Positive	1050
Stila	516	516	Negative	1044
Too Faced	514	514	Positive	1042

3.2. Splitting data ratio

In doing classification, the data need to be split into training and testing. The first process is done by testing a different number of iterations. In this process, the data were set with 100 iterations. A different number of iterations will test the highest number of accuracies. Table 3 shows the result of each different splitting accuracy. Splitting of 80% training and 20% testing offers the highest number of accuracies, as shown in Table 3.

Table 3. Result accuracy split ratio.

Training Size (%)	Testing Size (%)	Accuracy (%)
80	20	0.941
70	30	0.933
60	40	0.927
50	50	0.922

3.3. The testing result with different number of iterations

After selecting the ratio, the data was also tested with various iterations to see the model's performance. The model was tested with 100, 200, 500 and 1000 iteration sizes. It shows the accuracy trends of each iteration, and each accuracy was established in the graph Fig. 8.
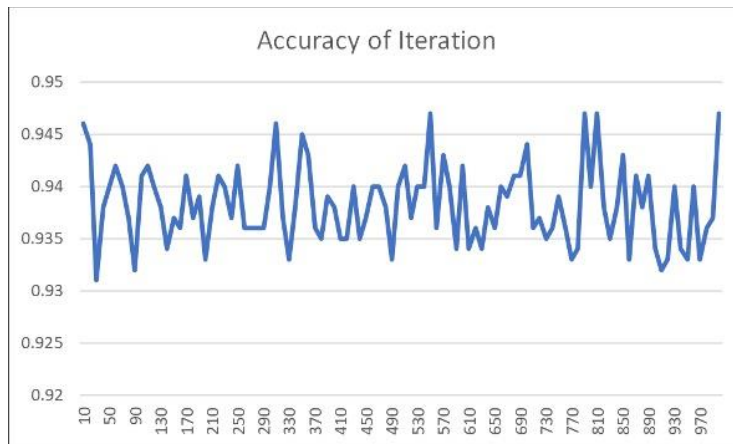

Fig. 8. Graph with accuracy in each iteration.

From the graph in Fig. 8, iteration 1000 show the highest accuracy among other iteration. The data show less gap between each iteration between each process. The process of the data did not go below 90%, and the highest is 94% in this process. The average accuracy is 92.8% between the iteration.

3.4. Testing result using 10-fold cross validation

To make the performance have the highest accuracy, fine-tuning is used to adjust the model. The model of Naïve Bayes in this model uses 10-fold cross-validation to get the highest data. In this process, 1000 iteration is used, and 80% of training and 20% of testing data are split, as shown in Fig. 9.
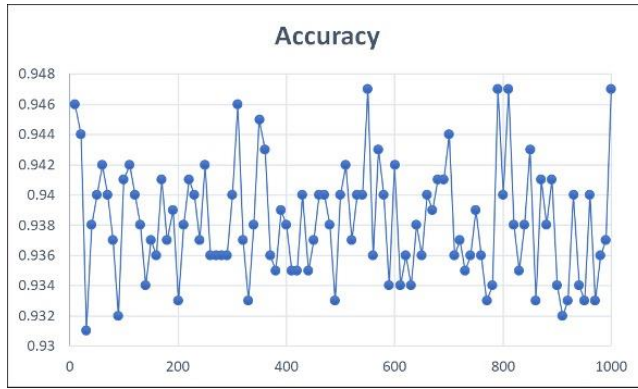
Fig. 9. Graph with accuracy in each iteration.

We can see the difference in each iteration show different results between the iteration. When the model was using cross-validation, the accuracy became higher. Cross-validation predicts randomly with the k equal subsample. The cross-validation process repeats the k time the fold, with each k subsample used in the validation data. At this stage, the maximum number of folds is ten (10), showing the highest number of folds and accuracy.

3.5. Visualization of sentiment analysis result

In sentiment analysis, visualization is the most crucial part of data analysis. The data is visualized by positive and negative reviews that have been predicted and actual data. The percentage of each data is calculated to know the sentiment score of each data. Each brand product is shown in the graph.

3.6. Predicted score

Figure 10 shows the predicted score for each brand. It shows that the actual and the expected difference did not differ substantially as the data have been sampled. The actual data of Kat Von D for the positive is 793, and been sampling and show the sampling data, which are 221.
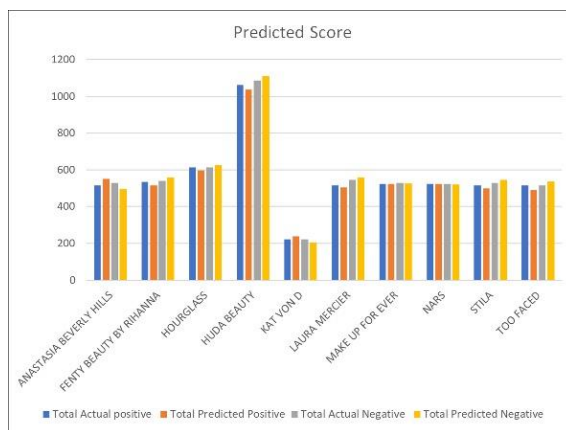

Fig. 10. The predicted score of each brand.

3.7. System result

For the system, the user needs to input the review of the data uploaded to the system. After the user enters the data, the system will show the result of the actual and predicted data of the corpus. Figure 11 shows the system that delivers the data that has the actual and predicted review. Then the system will show the result in a bar graph that shows the prediction of the dataset. In the system shown, blue is the actual and red is the prediction.

Figure 11 shows the bar graph of the system that indicates the predicted and actual data. It gives better visualization of data for better understanding. From this, we can expect the product's performance but know the product's positive and negative scores.
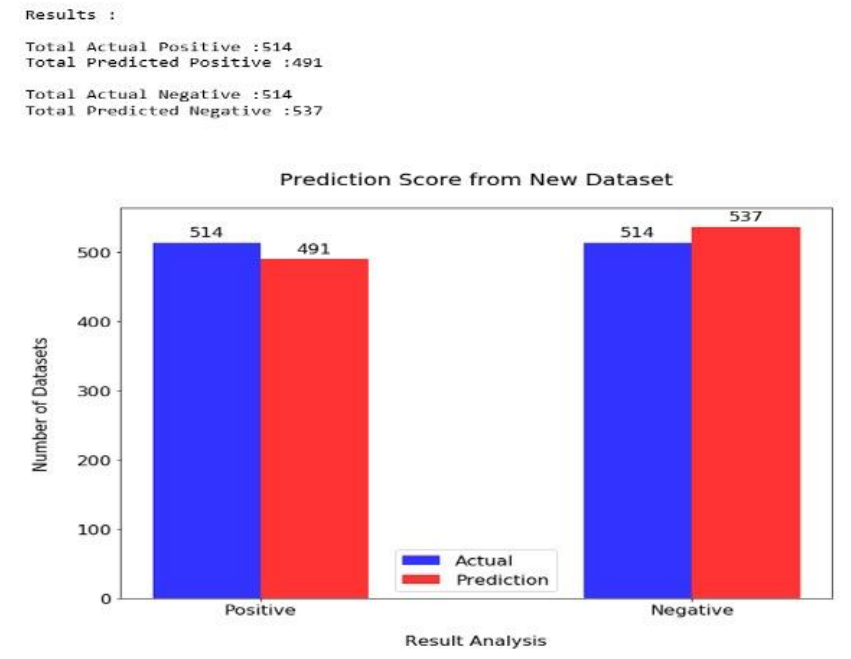
```
Results :

Total Actual Positive :514
Total Predicted Positive :491

Total Actual Negative :514
Total Predicted Negative :537
```


Fig. 11. Graph of data prediction.

4. Conclusion

This paper outlines the efforts made in researching the cosmetic product using a sentiment analysis approach and a Naïve Bayes classifier in order to construct a model that can anticipate new datasets and aid in business monitoring in terms of brand and product sentiment. Datasets from ten different brands of products were used, and the model's performance was assessed using 100, 200, 500, and 1000 iterations. With an average accuracy of 92.8%, Iteration 1000 displays the highest accuracy compared to other iterations. To obtain the best data, the model was then adjusted using Naïve Bayes 10-fold cross-validation. In this procedure, 80% of the training data and 20% of the testing data are split over 1000 iterations. At this point, there can be a maximum of ten (10) folds, demonstrating the largest number of folds and accuracy.

References

1. Sephora. (2021). Sephora. Retrieved August 12, 2021, from https://www.sephora.my/

2. Sephora.(2020). About Sephora. Retrieved August 20, 2020, from https://www.sephora.com.au/about.

3. Y. Laurencia. (2021). Sephora retail strategies As the Leading Beauty Retailer in the World. Retrieved August 20, 2020, from https://bbs.binus.ac.id/gbm/2021/01/21/sephora-retail-strategies-as-the-leading-beauty-retailer-in-the-world/

4. Honigman, B. (2021). How Sephora integrates retail & online marketing. Retrieved August 20, 2020, from https://etailwest.wbresearch.com/blog/how-sephora-integrates-retail-online-marketing.

5. Le, N.B.-V.; and Huh, J.-H. (2021). Applying sentiment product reviews and visualization for BI systems in Vietnamese E-commerce website: Focusing on vietnamese context. *Electronics*, 10(20), 2481.

6. Krishnan, S.; Amira, N.S.; Atilla, U.N.; Syafawani, S.; and Hafiz, M. (2017). The usage of cosmetic in Malaysia: Understanding the major factors that affect the users. *Management*, 7(1), 48-51.

7. Eastburn, J.C. (2015). *Conglomerate branding within the luxury goods sector.* Doctoral dissertation, University of Southern California.

8. Milnes, H. (2018). Why Sephora merged its digital and physical retail teams into one department. Retrieved August 20, 2020, from *https://digiday.com/marketing/sephora-merged-digital-physical-retail-teams-one-department/#:~:text=But%20the%20biggest%20goal%20of,customers%20at%20the%20right%20time.*

9. Valdivia, A.; Luzón, M.V.; and Herrera, F. (2017). Sentiment analysis in Tripadvisor. *IEEE Intelligent Systems*, 32(4), 72-77.

10. Gavilan, D.; Avello, M.; and Abril, C. (2014). Shopper marketing: A new challenge for Spanish community pharmacies. *Research in Social and Administrative Pharmacy*, 10(6), e125-e136.