JOURNAL OF DATA SCIENCE | Vol.2022:06
eISSN:2805-5160

Twitter Sentiment Analysis on Automotive Companies

Mohd Zaki Zakaria[1], Tri Basuki Kurniawan[2*], Misinem[3], Azizah Binti Soh[1]

[1]Faculty of Computer & Mathematics Sciences, University Technology Mara, Malaysia
[2]Faculty of Computer Science, University of Bina Darma, Palembang, Indonesia
[3]Faculty of Vocational, University of Bina Darma, Palembang, Indonesia

[*]**Email**: tribasukikurniawan@binadarma.ac.id

Abstract

Many users would use social media to express their opinions on their products or services. The expression can be good or bad. This project proposed a sentiment analysis on the automotive company where the users' opinions are analysed through this feedback. The data are collected from the social media of Twitter and followed by data mining techniques which are tokenization, removing stop words, and stemming. A sentiment classifier is implemented after the data have been converted into valuable data. Naïve Bayes classification is employed in this project by using Python language. Based on the dataset that we use, the article may analyse market demand for the automobile industry. According to the findings, Honda and Mazda had the highest positive sentiment, with more than 85 percent. This project is beneficial to the automotive industry, especially to teams' production. This finding supports a better understanding between the industry and their customer, to enhance the business strategies and find out the weaknesses.

Keywords

Sentiment Analysis, Automotive, Naïve Bayes, Python

Introduction

The automotive industry is known as one of the largest economic sectors in the world, with more than 90 million cars and other vehicles(Shukri, Yaghi, Aljarah, & Alsawalqah, 2015). In this modern era, cars have been brought to a new level to adapt to current technology to provide better comfort for their users. The design of a vehicle has several categories such as coupes, convertibles, sedans, SUVs, trucks, vans, sports, and (Shankar, 2013).

In business, the companies might have a problem collecting the feedback from their customers since most of them only give the overall ratings. The information from customers is essential for the company to understand better and find out the best way to improve the offering and get ideas for the new offerings. A traditional method to collect the information from the

Submission: 2 June 2022; **Acceptance:** 10 June 2022

customers by conducting a survey is not as effective as not every customer would respond to it. A new alternative by sentiment analysis can be used by developing a classification system for reviews and analyzing them. The reviews are scraped on social media by using sentiment and studying the overall ratings and the customer's opinions.

Four companies with the highest number of users in Malaysia are chosen for this research. The highest users here are determined by the number of sales in a particular year. The selected companies are the international companies: Honda, Nissan, Mazda, and Kia. Table 1 shows the highest sales by car brands in Malaysia. The reviews on the company product are scraped and classified into positive or negative reviews.

Table 1. Example of the caption for the table

Company	Number of vehicles sold
Honda	227,243
Nissan	28,610
Mazda	16,038
Kia	5,658

Source: Statista Research Department, 2018

Important Features in Car Purchasing

In purchasing a car, a customer would consider many factors to satisfy their desires. Some elements are the price, the brand, the design, the performance that the vehicle can maximize, and the fuel consumption. Among all these, the most influential features are fuel efficiency, safety, suitability for daily use, low price, and the brand's quality. It can be proved by figure 1 below, which shows the result of a survey conducted by Statista Global Consumer about the essential factors in buying a car in the United Kingdom (UK) from 2017 to 2018 (Armstrong, 2018).

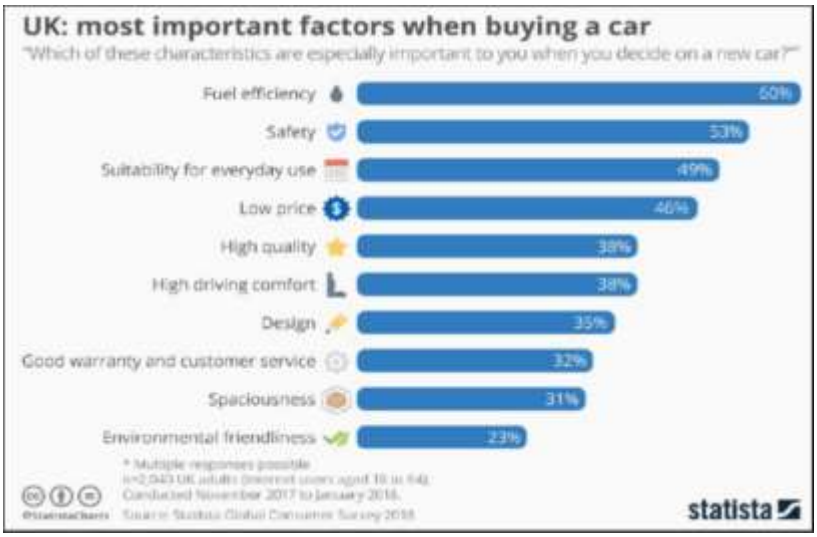

Figure 1. Important factors in purchasing a car in the UK between 2017 to 2018
Source: Armstrong, 2018

Consumer Buying Behaviors

Research by (Ramya & Ali, 2016) said that buying behaviour is widely used in marketing and has grown year by year. It is an attribute to perceive consumer buying behaviour in crating effect on product purchasing. It is natural when customers are being demanded on company production because they want to fulfill their needs, and if it is not in their requirements, the possibility for them to buy the product is low. Thus, the automotive industry must understand customer behaviors to stay active in the market. Using sentiment analysis can help the company to gain insight by performing research on what the customers feel unsatisfied with the products. Thus, the organization can always keep up-to-date on customer requirements by collecting and analyzing the comments from a user about their production. Besides fulfilling the customer demands to keep the number of customer retention, it can also lead the company to gain more potential customers. Measuring customer satisfaction can be done like what is shown in figure 2 and figure 3.

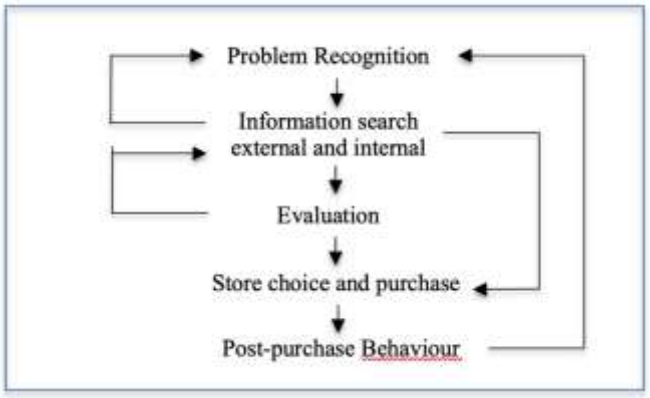

Figure 2. Model for a decision process
Source: Hawkins & Mothersbaugh , 2010

Figure 2 describes the steps in the decision process based on research by (Hawkins & Mothersbaugh, 2010). Based on their research, the decision process had five main phases: problem recognition, external and internal information search, evaluation and selection, store choice and purchases, and post-purchase behavior. In the problem recognition phase, the problem is identified by finding the difference between the existing product used by the user with the desired state. The customer finds out that the product is not what he expects in the information search phase. So, he collects further information about the same domain for comparison. In the evaluation and selection phase, the customer will evaluate the knowledge of the brands and look for problem-solving significant that is more fulfilled his needs. In-store choice and purchase phase, the customer chooses the product that can solve his problems. In the post-purchase behavior phases, the customer will check whether he is interested or not in purchasing the products.

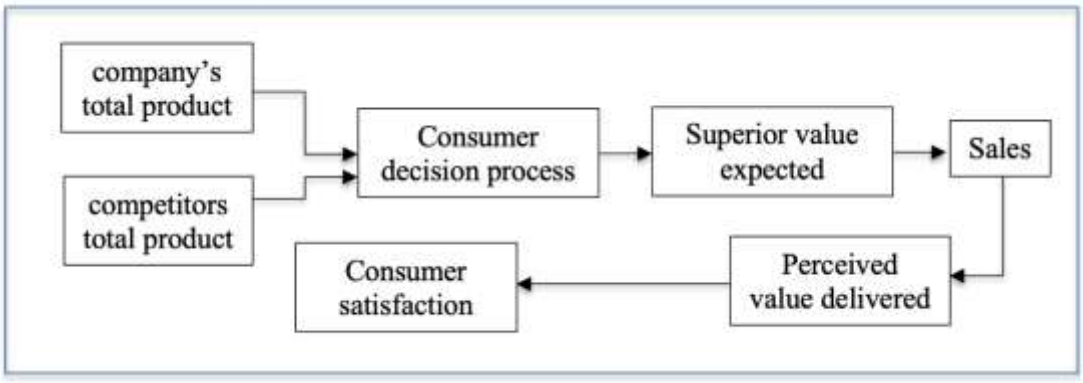

Figure 3. Model for creating satisfied customers
Source: Hawkins & Mothersbaugh , 2010

Figure 3 explains the model for creating a satisfied customer. It includes the company's total products, competitors' total products, a consumer decision process that has been discussed above, the superior value expected, sales, perceived value delivered, and customer satisfaction. All these need to be included to create a satisfied customer.

Methodology

Web Scraping

Web scraping is also known as Web Data Extraction, a technique to extract data from any available websites and save them from owning local files on the computer (Toth, 2020). The data in the websites usually do not offer the user to keep it or download the content inside of it. However, users can collect the data by scraping them programmatically from the websites.

Data Pre-processing

Data pre-processing is a method applied to data mining procedures (García, Ramírez-Gallego, Luengo, Benítez, & Herrera, 2016). Most of the collected data are not perfect and contain inconsistencies and redundancies. All the unnecessary things can be removed from the text by applying data pre-processing. A large amount of data requires more time pre-processing to get clean data for further implementation.

Feature Selection

Feature selection is a process of identifying and eliminating redundant and irrelevant feature dimensions to improve classification accuracy. Feature selection is vital in sentiment analysis as the selected feature subsets need to accurately present the features of the objects commented on by customers. The technique used for Feature Selection (FS) is the Natural Language Processing (NLP) approach which will eventually use a machine learning algorithm. NLP is an investigative approach to gathering rich knowledge resources through abstracting and retrieving thoughts or features from unstructured text. The examples of FS techniques using the NLP approach are parts of speech tagging (POST), opinion words and target relations, topic modeling, negation word, and rules of opinion. The procedures of the search method used in research are essential in time efficiency for feature selection. In sentiment analysis, the data dimension will be too large without feature selection and affect the accuracy of selected feature subsets. It has been discovered that it

is helpful to reduce the dimension of data used in the learning model (Yousefpour, Ibrahim, & Hamed, 2014).

Term Frequency-Inverse Frequency

Term Frequency-Inverse Frequency (TF-IDF)is a statistical method for indexing the term (Ghag & Shah, 2014). It is based on the text and term vector, which act as term frequency and term presence. TF-IDF is used as a weighting factor in searches for information retrieval, text mining, and user modeling. The value of TF-IDF is linearly proportional to the number of times a word appears in a document, where the value will increase when the number of the term increases.

The TF-IDF value is offset by the frequency of the word in the corpus, which can help adjust the fact that some words may appear more frequently in general. Due to the document changes in only one update count, TF-IDF is known to be robust and efficient on dynamic content. The equation used in TF-IDF calculation is as below.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Based on Equation 1, the explanation of the equation above is $tf_{i,j}$ represents the number of occurrences of *i* in *j*, df_i is the number of documents containing *i*, and *N* is the total number of the document.

Bag of Word

A bag of words (BoW) is one of the most used feature representation methods. The BoW is first proposed in the text retrieval problem domain for text analysis (Tsai, 2012). That process is where the text from the sentences is elicited by ignoring the grammar and the order of words. The frequency of words will be recorded while the document's structure is disregarded (Joachims, 2002).

Machine Learning Approach

The machine learning algorithm is a composition of methods to automatically recognize the available pattern in a given set of data (Amaral, Lopes, Jansen, Faria, & Melo, 2012). Machine learning commonly used in text analysis is such as Naïve Bayes (NB), Support Vector Machine (SVM), and Decision Tree. NB, SVM, and Decision Tree have supervised learning algorithms. NB uses the Bayes Theorem, which assumes independence among predictors to generate a hypothesis.

Multinomial Naïve Bayes can be explain by $P(c|X) = P(x_1|c) \times P(x2|c) \times \ldots \times P(x_n|c) \times P(c)$, where $P(c|X)$ *i*s the posterior probability, $P(x|c)$ is the likelihood, and $P(c)$ is the predictor prior probability. SVM is formally based on a hyperplane that carries out the process by building a hyperplane in a multidimensional space that separates cases of different class labels. Figure 4 will explains the concept of SVM.

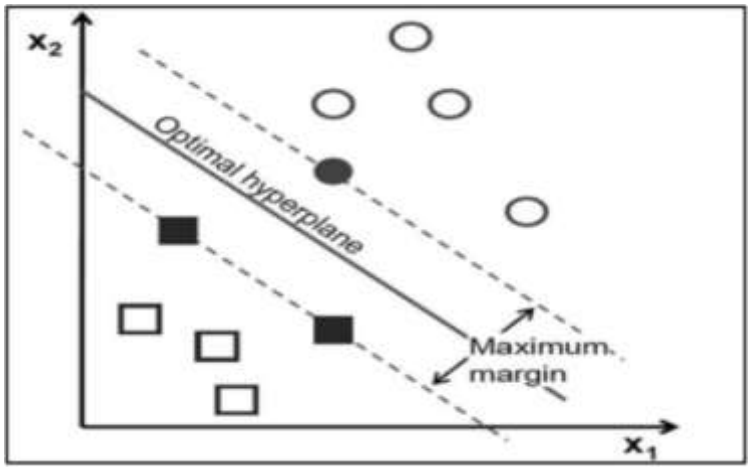

Figure 4. The Support Vector Machine concept

Result And Analysis

Data Sampling

A column chart is represented as a visualization of the number of reviews from four automotive companies, Honda, Kia, Mazda, and Nissan, as shown in figure 5.

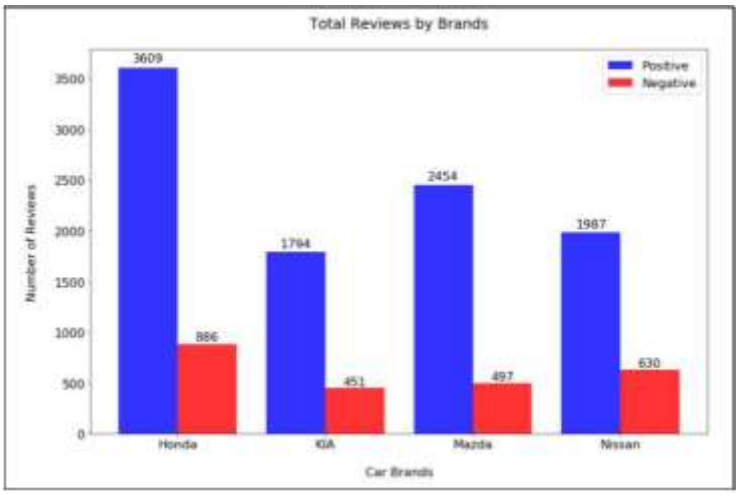

Figure 5. Total reviews by brand

As shown in figure 5, each document has different reviews. The chart shows that Honda and Mazda are quite trending among Twitter users, resulting in a more significant amount of data than Kia and Nissan. The query search is set with 2000 data retrieval for a single scraping process during the data scraping. However, none of these documents can reach that amount of tweets. Table 2 shows the total number of combined datasets of four cars used in this project.

Table 2. Total number of combined datasets

Positive	9844
Negative	2464
Total	12308

Figure 4 shows that the learning model with training data with fewer negative sets fits too well with positive sets. It caused the model to incorrectly classify most of the negative sets and classify all positive sets accurately. Data sampling is decided to be used as an alternative to avoid this overfitting. Hence, after the sampling process, the new number of selected positive data is only 2464, which means the total number of data used in the Naïve Bayes model would be 4928 as recorded in Table 3.

Table 3. Total number of balance datasets

Positive	2464
Negative	2464
Total	4928

Result Analysis Using Balanced Dataset

Testing Result

Testing is needed before the classification task. The first step is to split the dataset where 70% for training and 30% for testing. The training and testing processes are then iterated with ten (10) iterations, and the highest accuracy is updated throughout the iteration. After the last iteration, the highest accuracy is displayed, followed by its confusion matrix, precision, recall, and f1-score. The highest accuracy obtained in the classification is as shown in figure 6.

```
Documents name : Cars Reviews.csv
Testing size : 30.0%
Total positive tweets : 2464
Total negative tweets : 2464

Accuracy : 0.824
Iteration number : 2

Confusion Matrix :

Predicted  Negative  Positive   All
Actual
Negative        609       123   732
Positive        138       609   747
All             747       732  1479

Classification Report :

              precision    recall  f1-score   support

   Negative       0.82      0.83      0.82       732
   Positive       0.83      0.82      0.82       747

avg / total       0.82      0.82      0.82      1479
```

Figure 6. Highest accuracy in early result

Based on figure 5, 1479 data which is equivalent to 30% of the whole dataset, are selected for the testing set. The remaining 3449 data, referred to as 70% of the entire dataset, are chosen for the training set. The highest accuracy obtained is on the 2nd iteration, as shown in the figure above. Significantly, the result obtained is better precision after implementing the data sampling process, at 0.824%.

Testing Results on Different number of Iterations

To analyze the result's performance, another experiment was done to see the performance of the Naïve Bayes model with the different iteration numbers. The iteration sizes used in the model were three sizes which are 10, 20, and 30. The results obtained from each iteration size were then stored and compared to get better insights into accuracy trends. figure 7, figure 8, figure 9, and table 4. depict the results obtained from this experiment.

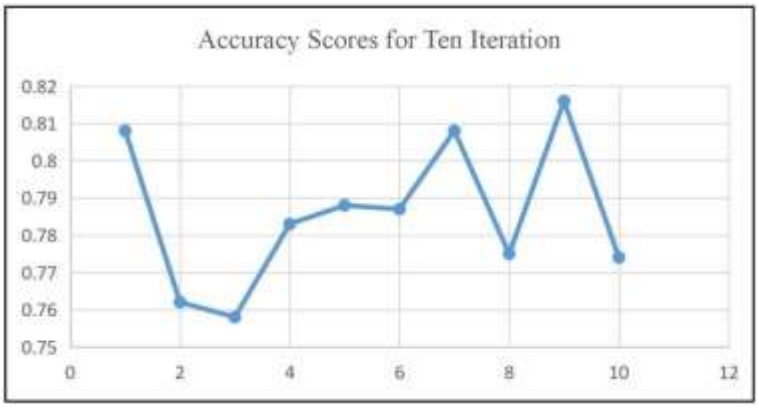

Figure 7. Accuracy trend with ten number of iteration

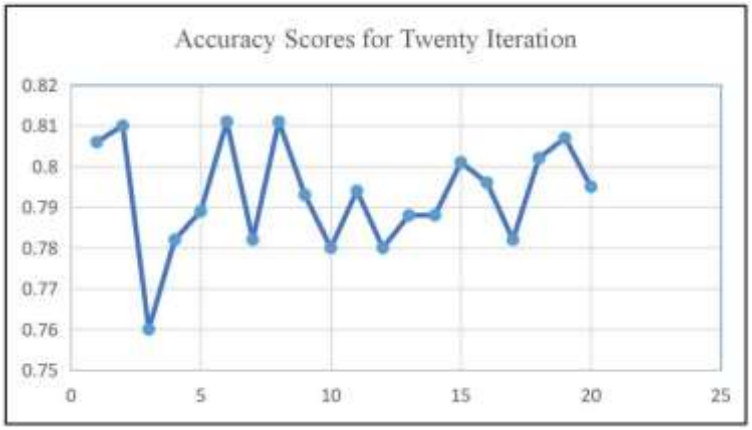

Figure 8. Accuracy trend with twenty iteration

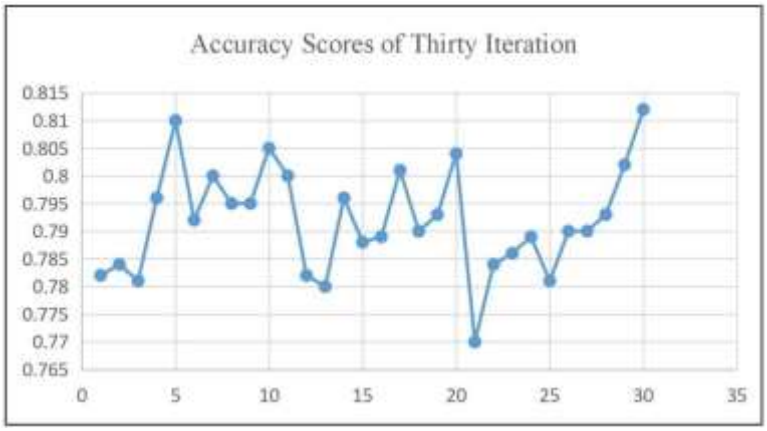

Figure 9. Accuracy trend with thirty number of iteration

Table 4. Comparison of the highest accuracy

Size of Iteration	Highest Accuracy
10	0.824
20	0.834
30	0.825

Fine-Tuning and Results Analysis

A minor modification and adjustment were implemented to boost the performance of the NB model to get the highest possible accuracy. This step is known as fine-tuning, and the results of this analysis are recorded. K=fold Cross Validation was applied in this study as it is a statistical method that has the potential to estimate the skill of the model.

Testing Result from 10-fold Cross-Validation

The result obtained after the 10-fold CV implementation into the model was recorded in figure 10.

```
Documents name : Cars Reviews.csv
Highest accuracy on KFold Validation is on fold 5 in iteration 5 :
Accuracy : 0.850

Confusion Matrix :

Predicted  Negative  Positive  All
Actual
Negative        201        41  242
Positive         33       218  251
All             234       259  493

Classification Report :

             precision    recall  f1-score   support

   Negative       0.86      0.83      0.84       242
   Positive       0.84      0.87      0.85       251

avg / total       0.85      0.85      0.85       493
```

Figure 10. 10-fold validation

The NB model is iterated with a new random split to decrease the variance based on figure 10. The default iteration is set as ten (10), and every iteration process was used with a new and different random split of training and testing sets. The set number 493 above is the outcome of the division of the dataset, which is 4928 respected with ten (10) folds. The result shows that fold five (5) recorded the highest accuracy of 0.850 % in the 5th iteration.

Different K-Fold Values with Different Iteration Sizes

This experiment is furthered by setting the K-fold at 5-fold, 10-fold, and 20-fold. Every K-fold is tested with different iteration sizes of 10, 20, and 30 to produce different results and get the best possible insight. The outcome of this process is analyzed and recorded, as shown in table 5.

Table 5. Comparison of using different K numbers and iteration size

K-fold	Highest Accuracy	Fold Position	Iteration Position	Iteration Size
K = 5	0.856	4th	5th	10
	0.860	10th	4th	20
	0.860	6th	15th	30
K = 10	0.871	8th	8th	10
	0.874	9th	17th	20
	0.863	10th	23th	30
K = 20	0.863	3rd	1st	10
	0.871	4th	3rd	20
	0.967	4th	18th	30

The fold and iteration positions represent the highest accuracy in the three conditions based on the table above. K value is seemed to influence its accuracy a bit. From the table above, it can be seen that applying a more significant number of K values can give better and higher accuracy than smaller K values. It can be concluded that a smaller number of K give minor variance and more bias. Meanwhile, a more significant number of K can potentially provide more variance and less discrimination. A better quality result can be obtained if the variance can be reduced without increasing the bias and repeat the CV process with the same K but different random folds before finalizing the results.

Comparing Accuracy with SVM Model

A simple SVM model is built just to see the comparison of accuracy scores between the Naïve Bayes model and the SVM model. Similar to Naïve Bayes, the dataset is split 30% for the testing process and 70% for the training process. Figure 11 shows the accuracy obtained by the SVM model with ten (10) iterations.

```
Documents name : Car_Review.csv
Testing size : 30.0%
Total positive tweets : 2464
Total negative tweets : 2464

Accuracy : 0.854
Iteration number : 9

Confusion Matrix :

Predicted  Negative  Positive   All
Actual
Negative        647       101   748
Positive        115       616   731
All             762       717  1479

Classification Report :

             precision    recall  f1-score   support

   Negative       0.85      0.86      0.86       748
   Positive       0.86      0.84      0.85       731

avg / total       0.85      0.85      0.85      1479
```

Figure 11. Accuracy of SVM model

Figure 11 shows that the highest accuracy obtained within ten (10) iterations are 0.854 at iteration 9. Overall, the SVM model receives slightly higher accuracy than the Naïve Bayes model. However, the time it takes to train the dataset is much longer compared to Naïve Bayes. Therefore, to avoid huge time consumption, the system thoroughly used the Naïve Bayes model.

The Visualization of Sentiment Analysis Results

After the classification phase, the study proceeded with the most crucial part: visualization of the sentiment analysis results. Reviews collected from four different models of cars are counted specifically to their positive and negative scores. The next step is to calculate their percentage for better reference and put it into a table, as shown in Table 6.

Table 6. Sentiment score for each company

Company	The positive sentiment (%)	The negative sentiment (%)
Honda	80.3	19.7
Kia	79.9	20.1
Mazda	83.2	16.8
Nissan	75.9	24.1

However, the negative opinions are also helpful for the companies where they can find out their weaknesses and get a clear insight into how they can overcome them.

Conclusion

In conclusion, this process is successfully done before the due date, as stated in the course timeline. All the objectives for this study are achieved and deliver the correct output for every single objective. As for the first objective, which is to analyze the customer's reviews on the car's significant features on social media, using a bag of words can be found by looking at the most called features by observing its word counts.

As for the second objective, which is to develop a system for the classification of reviews using sentiment analysis, the Naïve Bayes model has been constructed and performed well in this study. This model can produce high accuracy for the sentiment score at 0.81±% accuracy. Based on the literature reviews, most researchers would recommend the Naïve Bayes classification as it can give higher accuracy than other classifications such as SVM, decision tree, and others. Furthermore, this model can provide value for the organization for them to understand these reviews for future insights. This study has used the suitable methods and techniques suggested by many significant people where text processing, data analysis, and data visualization are used.

References

Amaral, J. L. M., Lopes, A. J., Jansen, J. M., Faria, A. C. D., & Melo, P. L. (2012). Machine learning algorithms and forced oscillation measurements applied to the automatic identification of chronic obstructive pulmonary disease. *Computer Methods and Programs in Biomedicine*, *105*(3), 183–193. https://doi.org/10.1016/j.cmpb.2011.09.009

Armstrong, M. (2018). UK: most important factors when buying a car [online]. Retrieved May 20, 2022, from Statista website: https://www.statista.com/chart/13124/uk_-most-important-factors-when-buying-a-car/

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. *Big Data Analytics*, *1*(1), 9. https://doi.org/10.1186/s41044-016-0014-0

Ghag, K., & Shah, K. (2014). SentiTFIDF – Sentiment Classification using Relative Term Frequency Inverse Document Frequency. *International Journal of Advanced Computer Science and Applications*, *5*. https://doi.org/10.14569/IJACSA.2014.050206

Hawkins, D. I., & Mothersbaugh, D. L. (2010). Consumer Behaviour:Building Marketing Strategies. In *McGraw-Hill*. Retrieved from www.mhhe.com

Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines.* Springer US.

Ramya, N., & Ali, M. (2016). Factors Affecting Buying Behavior. *International Journal of Applied Research*, *2*(I), 76–80. Retrieved from https://www.scirp.org/%28S%28lz5mqp453edsnp55rrgjct55%29%29/reference/referencespapers.aspx?referenceid=2929900

Shankar, S. (2013). Different Car Body Types. Retrieved May 20, 2022, from CarTrade.com website: http://www.cartrade.com/blog/2013/auto-guides/different-car-body-types-494.html

Shukri, S., Yaghi, R., Aljarah, I., & Alsawalqah, H. (2015). *Twitter Sentiment Analysis: A Case Study in the Automotive Industry*. https://doi.org/10.1109/AEECT.2015.7360594

Toth, A. (2020). What is web scraping? Retrieved May 20, 2022, from Scrapinghub website: https://www.scrapinghub.com/what-is-web-scraping

Tsai, C.-F. (2012). Bag-of-Words Representation in Image Annotation: A Review. *ISRN Artificial Intelligence*, *2012*, 376804. https://doi.org/10.5402/2012/376804

Yousefpour, A., Ibrahim, R., & Hamed, H. N. A. (2014). A Novel Feature Reduction Method in Sentiment Analysis. *International Journal of Innovative Computing*, *4*.