Jurnal Mantik
Volume 4 Number 1 May 2020, pp. 187-192 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik/index

Classification of Text Mining Review Oil Disfusser Products Using Naive Bayes Classification

Hilda Rachmi[1] , Artika Surniandari[2]

[1]Sistem Informasi, Teknik dan Informatika Kampus Kota Bogor, Universitas Bina Sarana Informatika, Jl. Kramat Raya No. 98, Jakarta Pusat
[2]Sistem Informasi Akuntansi Kampus Kota Bogor, Institution, Universitas Bina Sarana Informatika, Jl. Kramat Raya No. 98, Jakarta Pusat

E-mai : hilda.hlr@bsi.ac.id, artika.ats@bsi.ac.id

ARTICLE INFO

Article history:
Received: 02/04/2020
Revised: 10/04/2020
Accepted: 01/05/2020

Keywords:
Diffuser,
Text mining,
Naivebayes classification

ABSTRACT

Health is the main thing, especially when an outbreak of virus spreads and worries and the possibility of stress increases. Everyone wants to live healthy and avoid stress, that's why the use of natural medicines is the choice of one of them using essential oils or known as aromatherapy. The use of essential oils that are turned into steam and inhaled can produce a calming effect, the function of the pulse is more regular so that it is relaxed and fresher. Is a diffuser which is a tool to convert oil into steam, the use of the diffuser is rife and sales also get negative and positive comments from consumers. This has led researchers to examine consumer opinions about the diffuser product. Using the Naïve Bayes Classifier method to classify reviews based on positive sentiment class and negative sentiment class. From the labeling results, it is seen the association of texts in each sentiment class to find information that is considered important and can be useful for decision making. The classification results using the Naïve Bayes Classifier model obtain an accuracy rate of 76.00% and the value of the split ratio reaches the level of accuracy. The results of this study are expected to be developed and contribute to the development of sentiment analysis research by applying different methods.

1. Introduction

A healthy lifestyle is a choice that many people choose now because awareness of the importance of health is increasing, coupled with the spread of the deadly covid-19 virus. Stress levels can also increase due to uncertain economic conditions coupled with fears of contracting dangerous diseases. To reduce the impact of maintaining health through food intake is very important besides exercise and adequate rest. To get a calm and comfortable effect, many essential oil products are available as ingredients for aromatherapy. To be able to produce aromatherapy requires a tool that can convert oil into steam that can be inhaled, the device is a diffuser. This tool can provide a calming effect and can ward off various diseases depending on the material used therein, enthusiasts of this tool from various circles who really want aromatherapy in his room. In many online stores or marketplaces aromatherapy-producing devices are sold and in this study data will be processed using rapid miners to conclude a review of consumers who have purchased the device.

Based on the results of an Indonesian Poll study in collaboration with the Indonesian Internet Service Providers Association (APJII), in 2018 the number of internet users in Indonesia has reached 171.17 million. This figure is equivalent to 64.8% of Indonesia's total population of 264.16 million people (www.cnbcindonesia.com) [1]. For these results is a survey conducted for internet users in Indonesia alone, while internet users in the world far more than that number. One of the uses of the internet network is the growing interest in online shopping for internet users, one of which is consumers who buy these diffuser products through the amazone.com page. The method used is Naive Bayes classification to classify positive and negative comments from the review given on the diffuser product.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik
Volume 4 Number 1 May 2020, pp. 187-192 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik/index

2. Literature Review

2.1. Text Mining

Text mining is the process of mining data for later analysis with the help of software so that concepts, patterns, topics, keywords and other attributes can be identified in the data. According to [2] data mining is a series of activities used in discovering new patterns that are hidden or unexpected patterns in the data.

Through opinions expressed [3] data mining is a way to find meaningful patterns in large amounts of data. On another occasion [4] suggested that data mining is the application of a specific algorithm to extract patterns from data. Patterns generated from data mining can be used in predicting new data based on these patterns. The pattern is represented in a structural form that can be analyzed, can and is easily understood and can be used in decision making [5].

2.2. Naive bayes Classisfication

Naive Bayes classifier algorithm is an algorithm used to find the highest probability value to classify test data in the most appropriate category [6]. In this study, the test data are buyer comments. There are two stages in the classification of comments. The first stage is testing using training data on data that is already known in its category. Then the second stage is the testing process with data testing.

3. Research Methods

A. Data Collection

The data used in this study is in the form of comments from buyers of oil diffuser products on the amazone.com page. The news data amounted to 200 news data and divided into 2 positive and negative comment categories where each category numbered 100 positive comment data and 100 negative comment data. In accordance with the Paretto Principle, 80% of the data is used as training data and 20% of news data is used as testing data.

B. Pemodelan

The steps taken in text mining modeling are as follows:

a) Text Preprocessing

The action taken at this stage is toLowerCase, which is to change all the letters into lowercase characters, and Tokenizing is the process of decomposing the description that was originally in the form of sentences into words and eliminating delimiter-delimiter such as periods (.), Commas (,) , spaces and numeric characters in the word [7].

b) Feature Selection

At this stage the action taken is to eliminate stopword (stopword removal) and stemming of words that have implications [6] [5] . Stopword is a vocabulary that is not a feature (unique word) of a document [8]. For example "at", "by", "at", "a", "because" and so forth. Stemming is the process of mapping and decomposing various forms (variants) of a word into its basic word form (stem)[10] The purpose of the stemming process is to eliminate affixes in the form of prefixes, suffixes, and confixes that exist in each word..

C. Validasi dan Evaluasi

Stages of testing the model is done by measuring the value to test accuracy by using confusion matrix and split validation as the validation process. The parameters used to evaluate include: accuracy, precision, and recall taken from the confusion matrix table [11].

Tablel 1.
Confusion Matrix

Classification	Predicted Class	
	Prediction = Yes	Prediction = No
Actual = Yes	a (True Positive - TP)	b (False Negative - FN)
Actual = No	c (False Positive - FP)	d (True Negative - TN)

Source: [12]

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik
Volume 4 Number 1 May 2020, pp. 187-192 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik/index

Split validation is an operator provided in Rapidminer to conduct random validation, dividing the dataset into 2 parts, namely training data and test data in evaluating how accurate the model is used. Illustration of split validation can be seen in the image below:

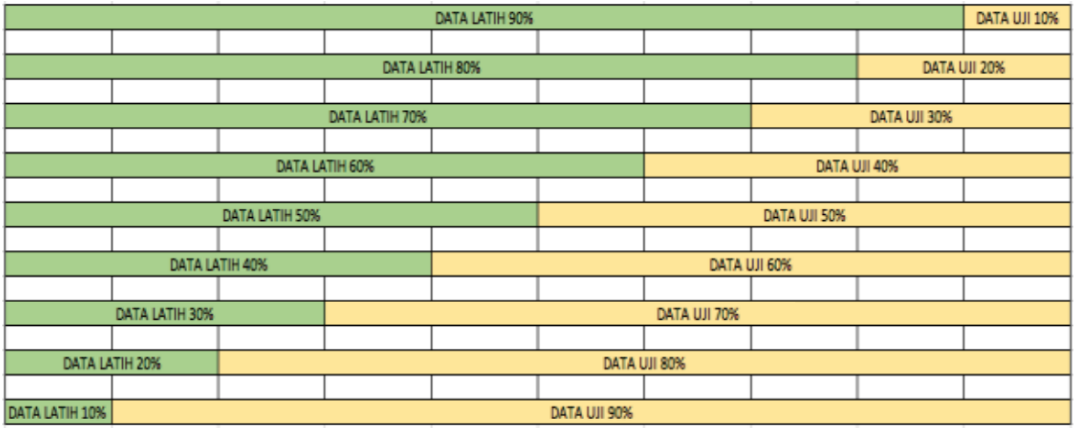

Fig 1. Split Validation Ilustration

source: [12]

Secara keseluruhan,metode penelitian yang diterapkan pada penelitian ini dapat dilihat pada gambar dibawah:

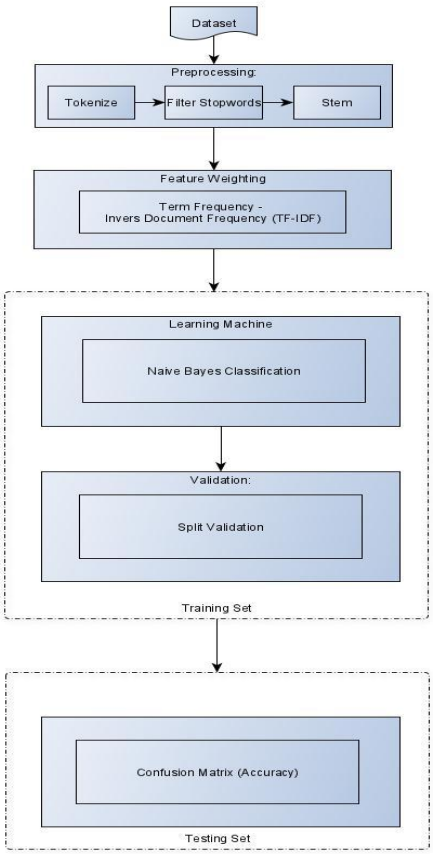

Fig 2. Alur Model Penelitian

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik
Volume 4 Number 1 May 2020, pp. 187-192 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik/index

4. Results and Discussion

This study uses product review data taken from amazon.com, as one marketplace that has more than 1000 consumers. The dataset used is comments on oil disfusser products totaling 200 text files that have been labeled as negative comments and positive comments, each totaling 100 data. 100 data divided into 50 negative comments and 50 positive comments were used as training data and the remainder were used as test data with 100 data and equally distributed 50 comments each for both negative and positive comments. The training data serves as the forming of the classification model and the test data functions for testing in evaluating the quality of the results of the use of the chosen algorithm.

Then the data is processed in preprocessing using the rokenize feature, filter stopwods, and stem. Data is broken down into tokens, the letters in the token are all converted to lowercase letters without punctuation. Next filter out the stopwords. And the last step is to change the data into basic words by removing the prefix and suffix from the data.

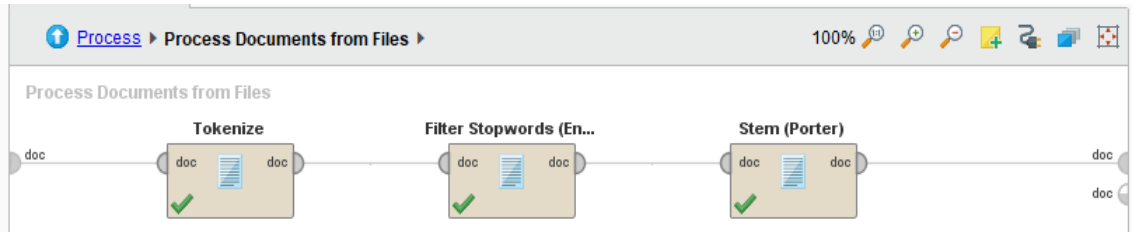

Fig 3. Preprocessing Step

Table 2.
Term Frequency Result

Row No	Label	Metadata_File	Awesome	Bad	Difficult	Good
1	negative	Negative1.txt	0	0	0	0
2	negative	Negative10.txt	0	0	0	0
3	negative	Negative100.txt	0	0	0.178501	0
4	negative	Negative11.txt	0	0	0.091371	0
5	negative	Negative12.txt	0	0	0	0
6	negative	Negative13.txt	0	0	0	0
7	negative	Negative14.txt	0.174137	0	0	0
8	negative	Negative15.txt	0	0	0	0.089073
9	negative	Negative16.txt	0	0	0	0.062161
10	negative	Negative17.txt	0	0	0.082854	0

The results of preprocessing then add the Naïve Bayes Classification model to obtain accuracy, precision and recall of the data processed.

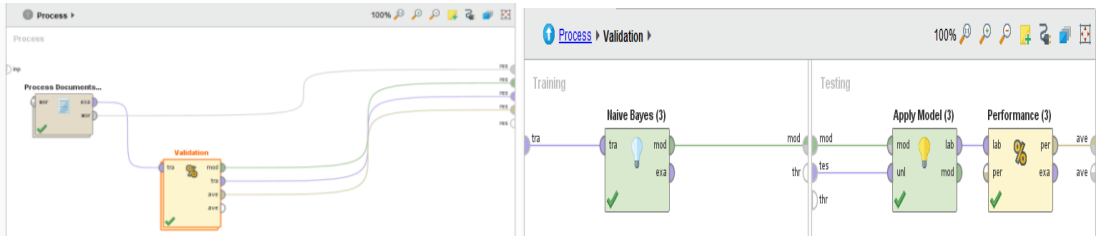

Fig 3. Preprocessing Step

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik
Volume 4 Number 1 May 2020, pp. 187-192 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik/index

Table 3.
Preprocessing from Rapidminer Result

Split Ratio	Accuracy	Precision	Recall
0,1	67,78%	63,33%	84,44%
0,2	68,75%	65,00%	81,25%
0,3	66,43%	61,39%	88,75%
0,4	66,67%	61,63%	88,33%
0,5	76,00%	68,75%	96,00%
0,6	66,25%	60,00%	97,50%
0,7	70,00%	62,50%	100,00%
0,8	67,50%	60,61%	100,00%
0,9	75,00%	66,67%	100,00%

Source : research, 2020

The table above shows the results of accuracy, precision, and recall with different split ratios. The table shows the best results with a split ratio of 0.5 and an accuracy of 76.00%, while the highest precision value with a value of 68.75% is also at a split ratio of 0.5.

Table 4.
Result Table

	True Negative	True Positive	Class Precision
Pred Negative	28	2	93,33%
Pred Positive	22	48	68,57%
Class Recall	56,00%	96,00%	

5. Conclusion

Based on the results of research conducted using Rapiminer 8.2 it can be concluded that the calculation of accuracy using the Naïve Bayes method related to the analysis of oil disfusser product comments obtained the lowest value of 66.25% with a split validation ratio of 0.6 and the highest value of 76% with a split validation ratio of 0 .5.

6. Reference

[1] R. Franedya, "Survei: Pengguna Internet di RI Tembus 171,17 Juta Jiwa," *CNBC Indonesia*, 2019. [Online]. Available: https://www.cnbcindonesia.com/tech/20190516191935-37-73041/survei-pengguna-internet-di-ri-tembus-17117-juta-jiwa 16 mei 2019. [Accessed: 02-Jul-2020].
[2] V. K. Deepa, J. Rexy, and R. Geetha, "Rapid development of applications in data mining," *2013 Int. Conf. Green High Perform. Comput. ICGHPC 2013*, pp. 1–4, 2013, doi: 10.1109/ICGHPC.2013.6533916.
[3] A. E. M. Sadiku, Matthew N. O; Shadare and S. M., "A Brief Introduction to Data Mining," *Eur. Sci. J.*, vol. 11, no. 21, pp. 1–3, 2015.
[4] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge discovery and data mining," *Am. Sci.*, vol. 87, no. 1, pp. 54–61, 1999, doi: 10.1511/1999.16.807.
[5] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. 2016.
[6] R. Feldman and J. Sanger, *The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press, 2007.
[7] F. Weiss, S.M., Indurkhya, N., Zhang, T., Damerau, *Text Mining Predictive Methods for Analyzing Unstructured Information*. New York: Springer, 2005.
[8] E. Dragut, F. Fang, P. Sistla, C. Yu, and W. Meng, "Stop word and related problems in web interface integration," *Proc. VLDB Endow.*, vol. 2, no. 1, pp. 349–360, 2009, doi: 10.14778/1687627.1687667. Diakses tanggal 8 Desember 2011
[9] M. W. Berry and J. Kogan, *Text Mining: Applications and Theory*. 2010.
[10] F. Z. Tala, "A Study of Stemming Efects on Information Retrieval in Bahasa Indonesia. Institute for Logic,

Jurnal Mantik
Volume 4 Number 1 May 2020, pp. 187-192 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik/index

Language and ComputationUniversiteit van Amsterdam The Netherlands," 2003. Diakses tanggal 29 September 2011.

[11] A. P. Wibowo and E. Jumiati, "Sentiment Analysis Masyarakat Pekalongan Terhadap Pembangunan Jalan Tol Pemalang-Batang Di Media Sosial," *IC-Tech*, vol. XIII, no. 0285, pp. 42–48, 2018.

[12] G. Galih, "Data Mining di Bidang Pendidikan untuk Analisa Prediksi Kinerja Mahasiswa dengan Komparasi 2 Model Klasifikasi pada STMIK Jabar," *J. Teknol. Sist. Inf. dan Apl.*, vol. 2, no. 1, p. 23, 2019, doi: 10.32493/jtsi.v2i1.2643.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.