International Research Journal of Computer Science (IRJCS)
Issue 12, Volume 4 (December 2017)
ISSN: 2393-9842
www.irjcs.com

Naive Bayes Algorithm with Chi Square and NGram Feature for Reviewing Laptop Product on Amazon Site

SUHENDRA
Information System , Computer Science Faculty
Mercubuana University, Indonesia
suhendra.mercu@mercubuana.ac.id

INDRA RANGGADARA
Information System Computer Science Faculty.
Mercubuana University, Indonesia
indra.ranggadara@mercubuana.ac.id

Manuscript History
Number: **IRJCS/RS/Vol.04/Issue12/DCCS10087**
DOI: 10.26562/IRJCS.2017.DCCS10087
Received: 11, November 2017
Final Correction: 28, November 2017
Final Accepted: 07, December 2017
Published: December 2017
Citation: **SUHENDRA & RANGGADARA, I. (2017). Naive Bayes Algorithm with Chi Square and NGram Feature for Reviewing Laptop Product on Amazon Site. International Research Journal of Computer Science, Volume IV, 28-33. doi: 10.26562/IRJCS.2017.DCCS10087**
Editor: Dr.A.Arul L.S, Chief Editor, IRJCS, AM Publications, India

Abstract–Sentiment analysis that will be performed on this research using a laptop product review datasets derived fromthe amazon.com web site. Because the currently the results of the laptop product reviews are provided online, itis easy to obtain such information. The number of datasets taken from that site is 500 reviews which is dividedinto 250 positive reviews and negative reviews. Python programming language and the plug-in NaturalLanguage Toolkit (NLTK) in this study are used in the preprocessing, training, testing and evaluation stages.This research aims to know the extent of the influence of the features selection of Chi Square and the N-Gram inanalyzing sentiment laptop product review with Naïve Bayes classification techniques. Evaluation of the resultsemployed accuracy, precision, recall and also f-measure. The result of the features selection stages are used toimprove the accuracy of the Naïve Bayes Algorithm from 80.15% to 94.44%.

Keywords: Naïve Bayes; N-Gram; Chi Square; Review ; Laptop;

I. INTRODUCTION

The laptop is a computer device that can be carried easily while travelling, so can be used to complete a variety of jobs. In terms of sales of Laptops not only done conventionally but currently such devices can be obtained online shop. Features and specifications are also owned by the Laptop is also very diverse, so in order to make it easier in selecting them, then the consumer can read a review about the product from the Laptop. Review of the consumer products can read online or also through magazines about computers. But now this Laptop product reviews online, so that consumers can get it by way of browsing through the internet. But the more review reads thus time required by consumers more and more time consuming. So needed a way to classify a product and also features. Classification is the process of document classifies a document into any specific category [1].Sentiment classification aims to address this problem by automatically classify user review be opinions positive or negative [2].

Classification techniques used for the analysis include review sentiment, Naïve Bayes, Support Vector Machines (SVM) and K-Nearest Neighbor (KNN) [3]. Sentiment analysis is the study of computation of opinions, sentiment, and the emotions expressed in the text [4]. In addition there is a wide range of research on the sentiment analysis done as sentiment Analysis, opinion review movies using classification Support Vector Machines and Particle Swarm Optimization [5]. Sentiment analysis on movie reviews and some products from Amazon.com using the classification Support Vector Machines and Artificial Neural Network [6]. Naïve Bayes Classifier (NBC) is a simple method but has the accuracy as well as high performance in text classification [7]. According to research [8] selection features of Chi Square are used in selecting appropriate features compared with the method of frequency-based. As well as suggesting merging NBC's machine learning techniques with some N-gram model to examine the accuracy of the classification results obtained. So by doing a data classification and selection also features on product reviews of laptops are expected to be able to improve the accuracy at the time of analysis of sentiment.

A. Research Problems

Based on the background of the aboveproblems, then the author can formulate theproblem as follows:

1) Naïve Bayes algorithm, how big can affectthe accuracy of sentiment analysis productreview laptop on Amazon site?

2) The extent to which the Chi Squarealgorithms and N-Grams are used asselection feature can improve the accuracyof the Naïve Bayes algorithm on sentimentanalysis product review laptop on Amazon site?

B. Limitation of Research

Limitation problem from this research are:

1) Algorithm in this research used Naïve Bayes forclassification of product review sentimentslaptops and the Chi Square and NGram algorithmsto perform feature selection.
2) The authors use laptop review data froman online media http://www.amazon.com/gp/toprated/electronics/565108 with the numberof datasets used was 500 review.
3) On the stages of the process of sentimentanalysis research was not done on Part ofSpeech Tagging.
4) Stopword removal process, the porterstemmingg and lemmatization only applyon English words only.
5) Evaluation of the results of the use ofaccuracy, precision, recall and also fmeasurein the form of a percentage.

C. Purpose And Objectives

The purpose of this research is to know how big influencefrom the selection of features of Chi Square and N-Gram in analyzing sentimentreview of laptop products with Naïve Bayes classification techniques and how Chi Square and N-Grams algorithms are used as selection feature can improve the accuracy of the Naïve Bayes algorithm on sentiment analysis product review laptop on Amazon site.

II. THEORY FUNDAMENTAL

A. Text Mining

Text mining, mining is done by a computer to get something new, something not previously known or rediscover the information implicitly implied, that comes from information that isextracted automatically from data sources of different text[9]. The stages of text mining is generally text preprocessing and feature selection [10].Text mining is a technique used to deal with classification, clustering, information extraction and information retrieval. Basically the work process of text mining many adopted from research Data Mining but the difference is the pattern used by the text mining is taken from the set of natural languages that are not structured in Data Mining patterns taken from structured databases [11].

B. Sentiment Analysis

According Thelwall in Haddi [12], sentiment analysis is treated as a task Classification classify an orientation text into positive or negative. According Moraes [6], steps that are commonly found in text classification analysis of sentiment are on below:

1) Define Domain DatasetThe collection of datasets that include adomain, for example dataset movie reviews, product reviews, and more.
2) Pre-processingThe initial processing stage is generallydone with the process of Tokenization, Stemming, and Stop words removal.
3) TransformationThe process of representations of numbersare counted from the textual data. Binaryrepresentation that is commonly used andonly count the presence or absence of aword in your document. How many timesa word appears in a document is also usedas a weighting scheme of textual data. Theprocess generally used i.e. TF-IDF, Binarytransformation and Frequency transformation.
4) Feature Selection Featured selection could make classification a more efficient or effective by reducing the amount of data to be analyzed by identifying the relevant features whichwill then be processed. Feature selectionmethods that are commonly used are the Expert.

C. Naïve Bayes Algorithm

According to Larose [13], Bayesian approach used to determine chances against the assumption around it. In Bayesian statistics, the parameters taken into consideration against the random and variable data are considered against possibilities results.

The Bayesian Approach first performed by Reverend Thomas Bayes (1702-1761) on "Essay Towards Solving a Problem in the Doctrine of Chances "published in 1763.Naïve Bayes, also called idiot "s Bayes, simple Bayes, and independence, Bayes is a good method because it is easy to make, require no parameter estimation scheme of complex iteration, this means can be applied to large datasets [14]. Naive Bayes classification it is assumed that there are certain characteristics or not of a class has nothing to do with the characteristics of other classes. Bayes theorem is an equation of:

$$P(H \mid X) = \frac{P(X \mid H).P(H)}{P(X)} \text{..... [14]}$$

Description:

X : Data with a class that is not yet known Hypothesis
H : X is a class data specific
P (H | X) : The probability of a hypothesis based on condition (a posteriori probability)
P (H) : the probability of the hypothesis H (the prior probability)
P (X | H) : the probability of X based on the conditions on the hypothesis H
P (X) : the probability of X

D. Chi Square

In this research for test data normality used chi square. According to Afandi and Hartati [15] definition is statistic analysis technique for test differentiation of frequency. Chi Square theorem is an equation of:

$$X^2 = \sum_{i-1}^{k} \frac{(fo-fe)^2}{fe} \text{.....[15]}$$

Description:

X^2 = Chi Kuadrat
fo = Get Frequency
fe = First Frequency

E. NGram

N-grams are n characters in a piece of string or a particular piece of n words in a phrase[16]. For example, in the word "TEKNIK" will be obtained by n-gram as follows [17]:

Table 1. example of cuts Ngram-based characters

Name	n-gram character
Uni-gram	T, E, K, N, I, K
Bi-gram	_T, TE, EK, KN, NI, IK, K_
Tri-gram	_TE, TEK, EKN, KNI, NIK, IK_, K__
Quad-gram	_TEK, TEKN, EKNI, KNIK, NIK_, IK_ _, K_ _ _

Blank character "_" on table 1 is used to represent a space at the front and end of the word and for word based n-gram for example is as follows. The sentence: "N-gram is a piece of string n characters in particular"

III. METHODOLOGY

A. Research Methodology

The research method used was experimental method, which will be conducted on the analysis of the sentiments of product reviews of laptops using the naive Bayes algorithm with feature selection of chi square and the N-Gram consists of several stages. Starting with a review of the data collection web site, preprocessing (tokenization, lower case, stop word removal, porter stemmer, lemmatization), feature selection with chi square and the N-gram algorithm with classification Naïve Bayes, evaluation results with accuracy, precision, recall, F-Measure.

B. Research Sample

Selection in this study sampling data taken from website http://www.amazon.com/gp/toprated/electronics/565108 With a range of time between January 2014 until January 2015 is randomly chosen by the researcher in order that the results of research conducted to approach optimal results. The amount of data that is retrieved is 500 laptop reviews, with 250 positive reviews that has a rating of 5 stars and also the negative review that has 250 rating is 1 star.

C. Testing Model

Testing the model aims to assess whether the model has been made already. In addition, the author does a process for testing the model which can be seen in Figure III-1 using software Python with version 2.7, wxpyht on version 3.0 and Natural Language Toolkit (NLTK) version 3.0 with dataset review product laptop from online siteshttp://www.amazon.com/gp/toprated/electronics/565108 which has been categorized into positive reviews and negative reviews.

D. Validation Testing

Testing method of validation by Focus Group Discussion that is held a group discussion participant is limited and participants were selected 6 students. In the implementation of the researcher as a Focus Group Discussion and the meeting's moderator, researchers presenting and demonstrating sentiment analysis review model to participants.

Researchers also describe any existing functions based on instruments that had already been prepared. FGD participants provide a response to the model analysis of the sentiments of the next review; researchers make conclusions based on the results of the responses from the participants of the FGD results analysis for testing.

E. Quality Testing

Testing the quality of the model of sentiment analysis was conducted to test the level of software quality that is generated based on the four factors model McCall, i.e.: Correctness, reliability, usability, and maintainability. The results of the identification of McCall, of the eleven characteristics of the quality of an application designated only four characteristics are made variable in this study, i.e. correctness, reliability, usability, and maintainability [18]. Testing is only done on the use of model analysis of the sentiments that have been made. Engineering quality testing conducted in this study by using questionnaire. Respondent characteristics as selection criteria research samples for testing the quality of the model is based on the analysis of the sentiments of its users.

IV. RESULT AND DISCUSSION

A. Testing Model

On the test model prototype stages sentiment analysis conducted after the first dataset review text is readily available then the next step is to process such preprocessing (lower case, tokenization, stop word removal, porter stemmer, lemmatization) when the process chosen is the preprocessing, whereas if you select the process accuracy then the next stage is the selection feature (Chi square, Ngram), classification (Naïve Bayes), and evaluation of results in the form of a percentage (accuracy, precision, recall, f-measure).

B. Naïve Bayes Algorithm Result

In figure 1is the result of the use of the naïve Bayes algorithm and is also equipped with a result of precision, recall, and f-measure. It can be seen that the accuracy of the results obtained at the time of naïve Bayes algorithm accuracy results obtained is 80.15%. Below is a picture of the results of the use of the naïve Bayes algorithm

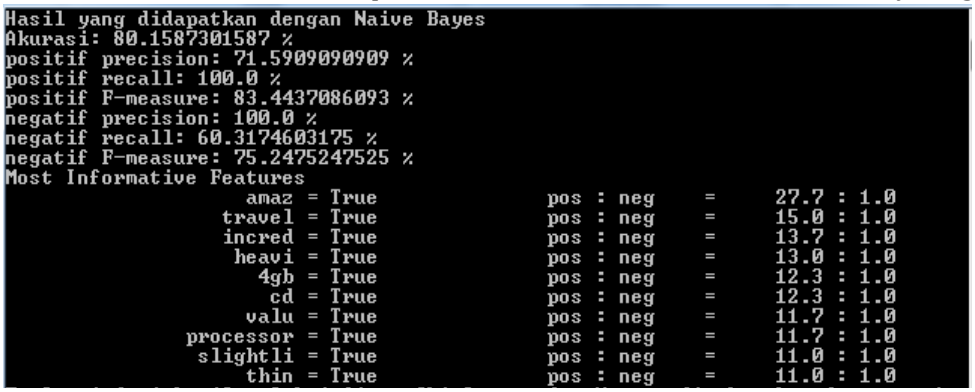

Figure 1. The result of the use of the Naïve Bayes Algorithm.

For ease in reading the result of precision, recall and f-measure in figure 2. Then a single table in order to get the average. Below is a table 2 of discussion of results:

Table 2. Average Results from Evaluation results.

Pos Precision	71.59 %	Pos Recall	100 %	Pos Recall	83.44 %
Neg Precision	100 %	Neg Recall	60.31 %	Neg Recall	75.24 %
Average	85.79 %	Average	80.15 %	Average	79.34 %

So the results of the evaluation of the results that it has taken the average value of such precision that is the level of accuracy of the classification results, and the overall number of introduction is done the system achieve 85,79%, to recall that the evaluation is to find out the success rate the performance of the user in the observation made on testing reached 80,15%, while the F-measure that is a value that is more influenced by the performance of the system compared to the user reached 79,34%. The value of the polarity that exists in figure 1 overall is positive.

C. Results after the merger of Chi Square and the N-gram

In figure 2 it can be seen that the results of the use of the naïve Bayes algorithm combined with the use of the selection features chi square as well as N-gram. Equipped with the result of accuracy, precision, recall, and f-measure. Below is a picture of the results after the use of chi square and the n-grams. In figure 2 showed that after using chi square and the n-gram on naïve Bayes classification accuracy which previously amounted 80,15% rose to 94,44% with the results so obtained are having a very good improvement. As for the result of precision, recall and f-measure can be seen in table 3. In table 3 of evaluations that have taken the average value of such precision that is the level of accuracy of the classification results, and the overall number of introduction is done the system reached 94,72%, to recall that the evaluation is to find out the success rate the performance of the user in the observation made on testing reached 94,44%, While the F-measure that is a value that is more influenced by the performance of the system compared to the user reached 94,43%.

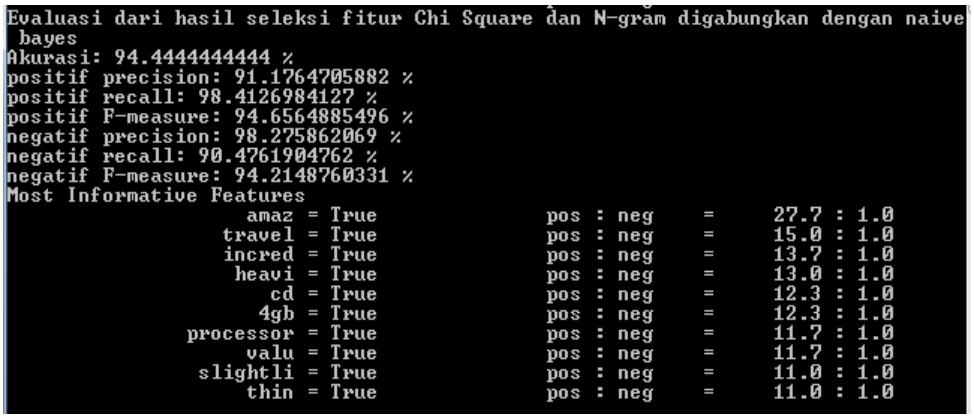

Figure 2. Results after the Use Of Chi Square and The N-Gram

Table 3. Evaluation of The Results by Chi Square and The N-gram

Pos Precision	91.17 %	Pos Recall	98.41%	Pos Recall	94.65 %
Neg Precision	98.27 %	Neg Recall	90.47 %	Neg Recall	94.21 %
Average	94.72 %	Average	94.44 %	Average	94.43 %

Of course with these results that the combination between naïve Bayes algorithm of classification and selection as well as the features of chi square and the n-gram is very good. Then to the results of the most informative feature has the same result as before which has positive polarity results.

D. Testing and Result of Validation

Focus Group Discussion activities are implemented in the Food Court area of the University of Budi Luhur on July 25, 2015 at 10:00 – 11:00 pm. Attended by 6 participants as respondents, the respondent is a student. Start a discussion, researchers conducting a presentation and a demo model analysis of the sentiments of the already designed. Researchers describe every function there is based on instruments that had already been prepared. After listening to the explanation of the respondent's researchers and find out how the use of model analysis of sentiment, the respondents are allowed to give feedback and approval via the form prepared researchers. Testing based on Focus Group Discussion that has been conducted in the Food Court area of the University of Budi Luhur can further is testing as follows:

Table 4. Test Results of FGD

The results of the Responses of the participants in the Focus Group Discussion
1. What is your opinion about the Input and Output generated from the model is already sentiment analysis in accordance with the needs of the users.
Responses of participants:
From all the responses of the respondents against input and output the resulting overall is already quite good, but there are some respondents expect enhanced features for more.
2. How do you think the sentiment analysis models designed, does allow a user in its use (user friendly)?
Responses of participants
All respondents argued that sentiment analysis models are easy to use from either side of the students.
3. How do you feel about the output generated from the model analysis of sentiment, does have benefits for you?
Response of the participants
All the respondents argue that the resulting outputs are very useful.
4. Whether the model analysis of the sentiments that are designed to have a fast response time for displaying the results of preprocessing and accuracy?
Response of the participants
The results of all respondents gave responses that fast enough response time of less than 1 minute to the dataset as much as 500.
5. How your response about the sentiment analysis models designed, whether this sentiment analysis model can be well received for its function?
Responses of participants
Conclusions of all respondents responded that this sentiment analysis model can be well received for its function.

E. Quality Test Result

Based on the analysis of data obtained from the questionnaire, the following recap of quality test results:

Results = Correctness + Reliability + Usability + Maintainability

= (7,067 + 7,147 + 7,093 + 6,833) / 4
= 7,035
= ((7,035 x 10) / 100) x 100%
= 70,35 % (Good)

Based on the result can be concluded that the level of quality of sentiment analysis software product review naïve Bayes algorithm using a laptop with the features of chi square selection overalls in the criteria, with the percentage of 70.35%. The highest quality factor is based on Reliability factor with percentage equal to 71.47%, next usability factor with 70,93%. Highest quality factor is based on the factors of reliability with the percentage of 71.47%, the next factor of usability with 70.93%. The next factor is Correctness with percentage of 70.67%, while the lowest is the quality factor of the Maintainability aspect with the percentage of 68.33%

V. CONCLUSION

Based on the results of the deliberations of the research that has been discussed in the previous chapter, then sentiment analysis research in product review laptop with naïve Bayes algorithm selection and also features the chi square can be drawn the conclusion as follows:

1. Naïve Bayes algorithm of usage that is used as a classification of data get the accuracy 80,15% of sentiment analysis product review laptops.
2. The use of Chi Square and algorithm of Ngrams are used as the selection of proven features can improve the accuracy results naïve Bayes algorithm originally 80.15%to 94.44%.

REFERENCES

1. C.D. Manning, H. Schütze, and P. Raghavan. Introduction to Information Retrieval. Cambridge: Cambridge University Press. 2008.
2. Z. Zhang, Q. Ye, Z. Zhang, Y. Li. Sentiment classification of Internet restaurant reviews written in Cantonese. Expert Systems with Applications, Vol.5, Issue 6, pp. 7674-7682. 2011.
3. R. Dehkharghani, H. Mercan, A. Javeed and Y. Saygn. Sentimental causal rule discovery from Twitter. Expert Systems with Applications, Vol. 41, Issue 10, pp. 4950-4958.2014.
4. B. Liu. Sentiment Analysis andOpinion Mining. San Rafael : Morgan &Claypool Publishers. 2012.
5. A. S. H. Basari, B. Hussin, I. G. P. Ananta.Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. Malaysian Technical Universities Conference on Engineering and Technology (MUCET), 2012.
6. R. Moraes, J. F. Valiati, W. P. G. Neto. Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications. Volume 40, Issue 2, pp. 621-633. 2013.
7. P. Routray, C. K. Swain, and S. P. Mishra. A Survey on SentimentAnalysis. International Journal of ComputerApplications, vol.76, no.10, pp. 1-8. 2013.
8. J. Ling. N. E. I.Kencana, T. B. Oka. Analisis Sentimen DenganMenggunakan Metode Naïve Bayes Classifier Dengan Seleksi Fitur Chi Square. E-jurnalmatematika, vol.3, no.3, pp.92-99. 2014.
9. R. Feldman and J. Sanger. The Text Mining Handbook : AdvancedApproaches in Analyzing Unstructured Data.Cambridge University Press : New York.2007.
10. M. W. Berry and J. Kogan. TextMining Aplication and theory. WILEY:United Kingdom. 2010.
11. Han, J., and Kamber, M., DataMining: Concepts and Techniques SecondEdition. Morgan Kaufmann publisher : SanFrancisco. 2006.
12. E. Haddi,X. Liu, Shi,Yong, The Role of Text Pre-processing inSentiment Analysis. London : BrunelUniversity, 2013.
13. D.T. Larose. DiscoveringKnowledge in Data: An Introduction to Data Mining. John Willey & Son Inc., New Jersey,2005.
14. X. Wu. and V. Kumar. The TopTen Algorithm in Data Mining. Boca Raton. CRC Press. 2009.
15. M. W. Affandi and S. C. Y. Hartati. Pengaruh Permainan Kecil Terhadap Minat Siswa Dalam Pembelajaran Pendidikan Jasmani, Olahraga Dan Kesehatan Pada Siswa Kelas V Minahdlatul Ulama Kecamatan Candi Kabupaten Sidoarjo. Jurnal Pendidikan Olahraga dan Kesehatan., Vol. 5 No.2, pp.253 – 259. 2017.
16. W.B. Cavnar, and J..M. Trenkle. N-Gram-Based Text Categorization. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. 1994.
17. Manalu and Boy Utomo. AnalisisSentimen Pada Twitter Menggunakan TextMining.Thesis. Universitas Sumatra Utara. 2014.
18. Winanti. Sistem PengambilanKeputusan Memilih MasakanBerdasarkan Jenis Penyakit KronisMenggunakan Analytical Hierarchy Process. Thesis. Universitas Budi Luhur. 2015.

