VOL 1 (2017) NO 2

e-ISSN : 2549-9904
ISSN : 2549-9610

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

Improvement of Email and Twitter Classification Accuracy Based on Pre-processing Bayes Naive Classifier Optimization in Integrated Digital Assistant

Aldo Erianda #, Indri Rahmayuni #

# Information Technology Department, Politeknik Negeri Padang, Padang, Indonesia
E-mail: nidal276@hotmail,com

Abstract— This research focuses on improving the accuracy of email and twitter classification. Spelling mistakes and lack of matches with bag of word causes the low accuracy in classifying. This research used naïve Bayes as a text classification algorithm. Text is divided into three categories: personal, work and family. To achieve maximum likelihood value for the category, a better preprocessing techniques is needed. It is necessary for the process to normalize the preprocessing and search for words that correspond to classes in the bag of word. So that the text can be classified by category or has a higher precision accuracy.

Keywords— **Bayes Naïve Classifier, Pre-processing, Machine Learning**

I. Introduction

Information is very important at this time. It is well balanced with user mobility information is getting higher. Accuracy of information obtained as a very necessary quality of information is essential to decide on decisions.

Integrated Digital Assistant is an application built to make it easier for users to receive information from email and twitter. With Integrated Digital Assistant, the information obtained from various sources are sorted in terms of conformity with the interests and needs of the user and displayed at the right time.

However, a problem arises in the classification of this text. Classification accuracy results with machines that are critical to determining the decision for the user. It is necessary for research to improve the accuracy.

Machine Learning is a scientific field that contains about learning computer / machine to be smart, how to make the machine can be "taught" through various techniques. One algorithm that can be used is Naive Bayes is often called the Naïve Bayes Classifier (NBC). NBC is simple but has a high accuracy. Based on the experimental results, NBC proved it can be used effectively to classify sentiment automatically with accuracy reaches 90.23% [1]. But assuming independent NBC (naïve) refers to the parameter estimation using the maximum likelihood method or the similarity of a document with other documents classes [2].

Machine learning is closely related to (and often overlaps with) computational statistics, which also focuses in prediction-making through the use of computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. [3]

Data mining algorithm is majorly depending upon the type of data and quality of data [4]. To increasing the maximum likelihood optimization needs to be done in the pre-processing techniques [5]. This technique will normalize the word stem and word errors in the text so that the word according to the class of other training documents to get a more accurate classification results.

II. Related Work

Noisy text discusses research conducted by Alexander Clark. In his paper, Alexander Clark noisy text are categorized into Orthographic Model, Model Error, and White Space Type. Orthographic model is used to improve the text, which has the shape of the letter mistake. Error models is used to fix a misspelled word or a typo. While the White Space Type is used to minimize the use of the space that follows a word or paragraph that has no meaning in text processing [6]

III. Analysis And System Design

NBC proved can be used effectively to automatically classify sentiment [7]. But assuming independent NBC (naïve) refers to the parameter estimation using the maximum likelihood method or the similarity of a document with other documents classes.

Problems will occur mistyped a word and the other noise word and the lack of similarity classes of test data and training data. Test data was minimal similarity with training data can be caused by imperfect stemming process. Stemming Indonesian in separating the basic words with affixes and suffixes are often the cause of the lack of similarity. As our sample text "*Softwarenya dapat dibeli di outlet Microsoft di kota anda*", software can be categorized as a text work because software has high enough frequency in work category. However, due to imperfect stemming process in removing the suffix "-nya", the text is not defined as a text work.

Beside that spelling errors can reduce the level of accuracy. Here's a common typing error on email and twitter,

- Replacement of numbers to letters

  N0! I thought Rapunzel was rea1! I wonder h0w much shampoo Rapunzel uses to wash her hair before she cut it.
- Repetition Letter

  Foxconnnn akan menandatangani kerja sama dengan mitra lokal untuk mendirikan pabrik yang akan beroperasi untuk melakukan perakitan dan produksi. Dengan hadirnya Foxconn di Indonesia, diharapkan produksi komponen di dalam negeri dan jangan sampai hal ini hanya menjadi WACANAAA!!!
- Typing Errors

  Generally, consists of metathesis (transposition of two adjacent letters), and substitution errors in which the adjacent key replaced correct, such as 'f' substituted for 'd'

IV. Methodology

A. Noisy Text Normalization

Pre-Processing design for noisy text conducted to prove that by using this process can improve the prediction accuracy of the classification by the Naïve Bayes Classifier algorithm. Normalization is to be performed:

- Replacement of numbers to letters

  Original text:

  N0! I thought Rapunzel was rea1! I wonder h0w much shampoo Rapunzel uses to wash her hair before she cut it.

  Text after normalized:

  No! I thought Rapunzel was real! I wonder how much shampoo Rapunzel uses to wash her hair before she cut it
- Repetition Letter

  Original text:

  Foxconnnn akan menandatangani kerja sama dengan mitra lokal untuk mendirikan pabrik yang akan beroperasi untuk melakukan perakitan dan produksi. Dengan hadirnya Foxconn di Indonesia, diharapkan produksi komponen di dalam negeri dan jangan sampai hal ini hanya menjadi WACANAAA!!!

  Text after normalized:

  Foxconn akan menandatangani kerja sama dengan mitra lokal untuk mendirikan pabrik yang akan beroperasi untuk melakukan perakitan dan produksi. Dengan hadirnya Foxconn di Indonesia, diharapkan produksi komponen di dalam negeri dan jangan sampai hal ini hanya menjadi wacana
- Substitution and metathesis

  Original text:

  Pernah bayangin nggak kalo aplikasi Mailboz yg di iPhone itu, trus dibuat di BB10, seberapa sering salah geser? (kanan-kiri)

  Text after normalized:

  Pernah bayangin tidak kalau aplikasi Mailbox yang di iPhone itu, terus dibuat di BlackBerry 10, seberapa sering salah geser? (kanan-kiri)

B. Stemming Optimization

Porter Stemmer is used in the process of pre-processing. Porter Stemmer is a conflation Stemmer (for English) [8], in particular stemming has five steps and apply certain rules in every step. every word in the process one by one in each rule and so on until all the rules have been performed or there are no more rules that can be in process.

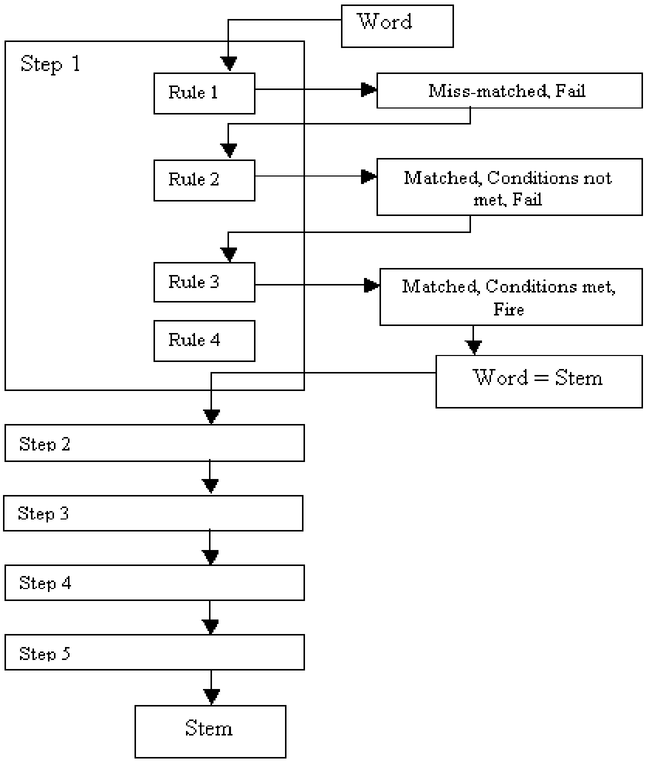

Fig. 1 Stemming Process

The result of stemming process,
Original Text:
Softwarenya dapat dibeli di outlet Microsoft di kota anda. Untuk potongan harga bawalah kartu pelajar dan anda mendapatkan potongan sebesar Rp 100.000 rupiah

Text after stemming:
Software dapat beli di outlet Microsoft di kota anda. Untuk potongan harga bawa kartu pelajar dan anda dapat potongan sebesar Rp 100.000 rupiah

C. Naïve Bayes Classifier

To classify text into a specific category needs an algorithm that serves as a machine learning. Naïve Bayes Classifier is used as algorithms in this study will be conducted two times of testing. The first test classification using pre-processing and subsequent optimization classification without pre-processing.

Here is the method that conducted in the process of classification using Naïve Bayes Classifier,
- The process of training.
Input is a sample documents that have been known category.
Data Training Example,

- Text Work,
Metode yang digunakan untuk mereview penelitian dan merangkumkannya dalam "review paper" sudah mulai distandardkan oleh para dosen Informatika Politeknik Telkom.
- Text Personal,
Nanti mau nonton Chelsea di politeknik telkom ah. berarti sekarang semua harus udah beres atau bisa disambi baca
- Text Family,
Keluarga yg harmonis, karir yg menanjak dan hubungan sosial yg harmonis sangat diinginkan oleh setiap orang

TABLE I
TRAINING DATA

	Text	Sentences	Class
Train	1	*Metode digunakan review penelitian merangkumkan review paper mulai distandardkan dosen informatika politeknik telkom*	Work
	2	*Nanti mau nonton chelsea politeknik telkom ah berarti sekarang semua harus udah beres bisa disambi baca*	Personal
	3	*Keluarga harmonis karir menanjak hubungan sosial harmonis sangat diinginkan setiap orang*	Family
test	4	*Biaya penelitian Departemen Keuangan dapat digunakan dosen Informatika Politeknik Telkom*	?

Calculate *P*(Vj) by the equation

$$P(V_j) = \frac{|\, docs\ J\, |}{|\, example\, |}$$

$Docs\ J$ is the set documents in the V_j category
The Priorities are,
P(work) = 1/3
P(personal) = 1/3
P(family) = 1/3
After that save all the vocabulary words as well as the number of frequency.

- Classification Process

At this stage the system is designed to produce *vMAP* according to the equation

$$V_{MAP} = \arg_{v_j \epsilon V} \max P(v_j) \prod {}_i P(a_i |\, v_j)$$

using P (V_j) and P ($W_K|\, V_j$) which has been obtained from the training.

Calculate P ($W_K|\, V_j$) with the equation

$$P(W_k\, |V_j) = \frac{n_k + 1}{n + |\, vocabulary}$$

The next stage of the whole word in the text four or test text will be saved as a bag of word for class work.
- The final stage is to find the level of accuracy by comparing the number of correct classification of test documents

$$\text{Accuracy} = \frac{Correct\ Classification}{Amount\ of\ Training\ Data} \text{ x } 100\%$$

This stage conducted manually checking the classification results. Any text that has been classified will be checked one by one to know the truth of the classification results.
example,

Test document number as many as 20 test data. Correct classification number as many as 15 documents. Then the accuracy is

$$\text{Accuracy} = \frac{15}{20} * 100\% = 75\%$$

V. Experiment And Result

The result of the test is shown in table 2

TABLE II
EMAIL TEST RESULT

Training Data	Without Optimization	Optimization
50 data	82.2	90.49
100 data	84.58	92.54
160 data	85.13	95.23

Accuracy on each additional test data has increased. Increased rate of approximately 2-3% accuracy on each additional test data. By optimizing the accuracy increased by 6-10% compared with the classification accuracy without performing the optimization

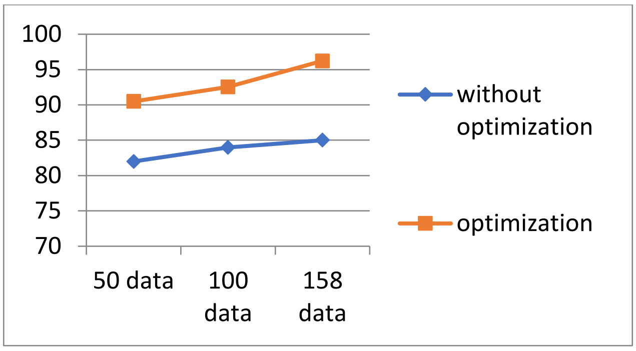

Fig. 2 Email Test Result

Twitter test data test results using optimization and without preprocessing optimization shown by Table 3

TABLE III
Twitter Test Result

Training Data	Without Optimization	Optimization
50 data	75.2	90.49
100 data	75.58	92.54
158 data	76.13	96.23

Increased accuracy on each additional test data. An increase of 1% accuracy on each additional test data. By optimizing an increase of 15-21% accuracy of classification accuracy without performing the optimization.

It is very different to the classification of email as the number of words on twitter is much less than the number of words in the email. Normalization of the word is wrong (noisy word) has more influence on the classification accuracy because fewer classes and each class has a greater tendency.

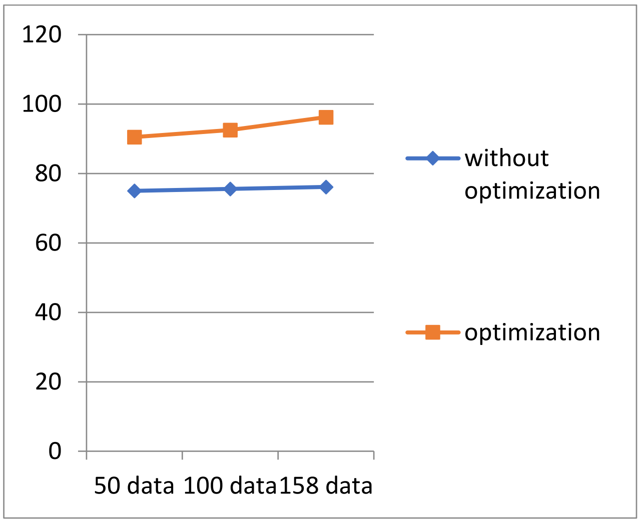

Fig. 3 Email Test Result

Overall classification using preprocessing optimization can indicate a higher accuracy value than without optimization. The accuracy rate obtained was more than 90%

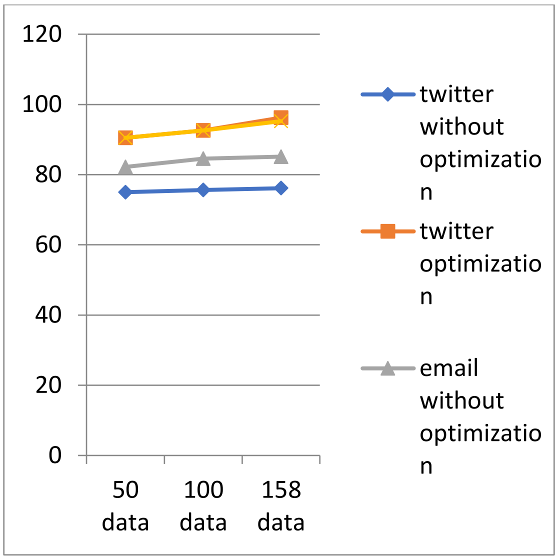

Fig. 4 Email and Twitter Test Result

VI. Conclusion

From the research conducted it can be concluded that the word Noisy may reduce the accuracy of the classification and should be normalized first. Then the method using Naïve Bayes classifier can be used as a pre-processing optimization methods to improve classification accuracy on Email and Twitter.

Naïve Bayes classifier relies on training data, so it requires good training data and the right amount.

References

[1] Lotte Scholten and Daan van Knippenberg and Bernard A. Nijstad and Carsten K.W. De Dreu, Motivated information processing and group decision-making: Effects of process accountability on information processing and decision quality, Journal of Experimental Social Psychology Volume 43 No.4, Elsevier, 2007

[2] Wibisono, Y. Klasifikasi Berita Berbahasa Indonesia. Bandung, 2005

[3] Amnur, Hidra. "Customer Relationship Management and Machine Learning technology for Identifying the Customer," International Journal On Informatics Visualization, 2017

[4] Haykin, Simon. Neural Networks: A Comprehensive Foundation. Prentice Hall New Jerey. 1990

[5] Dixit, S and Gwal, N. A Implementation of Data Preprocessing for Small Data-Set. International Journal of Computer Applications (0975 – 8887) Volume 103 – No.6, October 2014. Genèva.2014

[6] Razavi A.R., Gill H., Åhlfeldt H., Shahsavar N. A Data Pre-processing Method to Increase Efficiency and Accuracy in Data Mining. In: Miksch S., Hunter J., Keravnou E.T. (eds) Artificial Intelligence in Medicine. AIME 2005. Lecture Notes in Computer Science, vol 3581. Springer, Berlin, Heidelberg. 2005

[7] Clark, Alexander. Pre-processing Very Noisy Text, Proceeding of Workshop on Shallow Processing of Large Corpora, 2003.

[8] B, Pang. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics. 2004

[9] Issac, B. Pre-processing Very Noisy Text, Proceeding of Workshop on Shallow Processing of Large Corpora, 2003.