Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Sentiment analysis system for movie review in Bahasa Indonesia using naive bayes classifier method

To cite this article: Yanuar Nurdiansyah *et al* 2018 *J. Phys.: Conf. Ser.* **1008** 012011

View the article online for updates and enhancements.

You may also like

- Multi stratagem analysis of sentiments on twitter data using partial phrase harmonizing
  T M Saravanan, T Kavitha, S Hemalatha et al.
- Transmission characteristics of investor sentiment for energy stocks from the perspective of a complex network
  Yajie Qi, Huajiao Li, Nairong Liu et al.
- Sentiment analysis using lexicon-based method with naive bayes classifier algorithm on #newnormal hashtag in twitter
  R L Mustofa and B Prasetiyo

ECS Toyota Young Investigator Fellowship

ECS TOYOTA

For young professionals and scholars pursuing research in batteries, fuel cells and hydrogen, and future sustainable technologies.

At least one $50,000 fellowship is available annually. More than $1.4 million awarded since 2015!

Application deadline: January 31, 2023

Learn more. Apply today!

This content was downloaded from IP address 36.95.236.66 on 19/12/2022 at 04:35

Sentiment analysis system for movie review in Bahasa Indonesia using naive bayes classifier method

Yanuar Nurdiansyah[1], Saiful Bukhori[1], Rahmad Hidayat[1]
[1]Faculty of Computer Science, University of Jember, Indonesia

E-Mail : yanuar_pssi@unej.ac.id, saiful.ilkom@unej.ac.id

Abstract—There are many ways of implementing the use of sentiments often found in documents; one of which is the sentiments found on the product or service reviews. It is so important to be able to process and extract textual data from the documents. Therefore, we propose a system that is able to classify sentiments from review documents into two classes: positive sentiment and negative sentiment. We use Naive Bayes Classifier method in this document classification system that we build. We choose Movienthusiast, a movie reviews in Bahasa Indonesia website as the source of our review documents. From there, we were able to collect 1201 movie reviews: 783 positive reviews and 418 negative reviews that we use as the dataset for this machine learning classifier. The classifying accuracy yields an average of 88.37% from five times of accuracy measuring attempts using aforementioned dataset.

1. Introduction

Information presented in cyberspace is now more diverse and the media used means in the process of information diffusion are growing. One of the main media used in the process diffusing information in cyber media is text or document media. The ability in order to extract information from documents is absolutely necessary. The method of extracting information from data in the form of documents is known as text mining. Over 80% of information is currently stored in the form of text, so that text mining is believed to have high commercial value potential [1]. describe the steps taken in text mining: tokenizing, filtering, stemming, tagging, and analyzing [2].

Sentiment analysis is one of the new branches in the domain of text mining or data extraction in the form of text, consisting of processing and extracting textual data automatically in order to obtain information [3, 4, 5, 6, 7]. Sentiment analysis can be utilized as a tool in seeing the public response of a particular event, either positive or negative response, so that the next strategic steps can be undergone immediately. An example of sentiment data in the form of document is movie review from various sites on the internet. Reviews obtained from movie reviews sites can be used a reference for movie fans to know recommended and also a medium for movie producers to know the public responses towards the movie released. Movie review can be divided into a number of categories based on the sentiments contained in the document.

There are some difficulties in developing sentimental analysis study on movie review using Bahasa Indonesia partly due to the lack of studies on movie review documents using Bahasa Indonesia and the lack of corpus or textual literature for movie review using Bahasa Indonesia as the original language of the document.

Based on the description above, the researchers are interested in creating a system which is capable of categorizing movie review in Bahasa Indonesia into two categories of positive and negative sentiments using the method of Naïve Bayes Classifier [8, 9]. Naïve Bayes Classifier method is one of the supervised learning models based on statistics and probability with high level of accuracy [10].

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

2. Result and Discussion

The steps in this study include the step of literature review, data collection, and system development. The method employed in system development is the method of Naive Bayes Classifier [11]. The flow diagram of the study steps is illustrated in Figure 1.

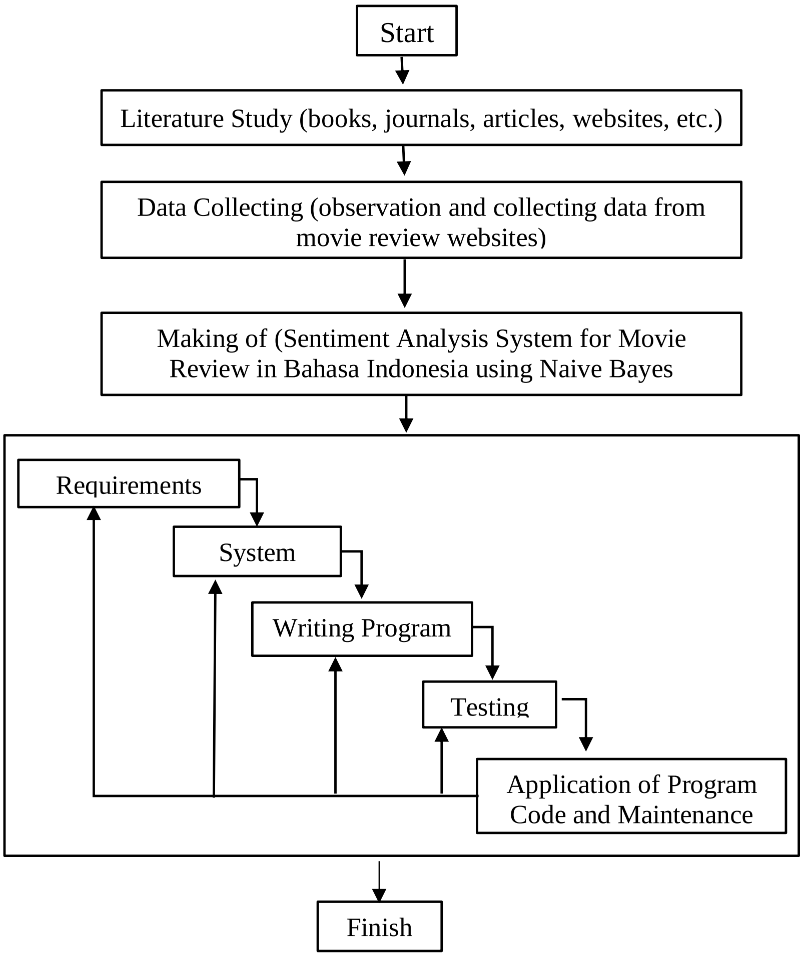
Start
Literature Study (books, journals, articles, websites, etc.)
Data Collecting (observation and collecting data from movie review websites)
Making of (Sentiment Analysis System for Movie Review in Bahasa Indonesia using Naive Bayes
Requirements
System
Writing Program
Testing
Application of Program Code and Maintenance
Finish

Figure 1. Flow Chart of Research Phases

2.1 Data Collection

The data collection is performed in order to obtain the information needed to determine the system requirements. The data used in the study is divided into two categories based on the data source: primary data and secondary data. The primary data required in the development of the Sentiment Analysis System for Movie Review is a dataset of movie review. The movie review dataset is made by the researcher using a movie review obtained from the movie review web page of Movienthusiast. Secondary data required in this study is the root words in the Great Dictionary of the Indonesia Language (KBBI) and stop words dictionary or set of words which are considered less important and is able to be removed from the movie review text [12, 13, 14]. The dictionary of stop words used comes from the study conducted by Tala.

2.2 The Naive Bayes Classifier Method Implementation

The Sentiment Analysis System for Movie Review which is developed implements the Naive Bayes Classifier method in classifying the movie review documents into two classes: positive sentiment class and negative sentiment class. The flowchart regarding the overview of the Sentiment Analysis System for Movie Review is illustrated in Figure 2.

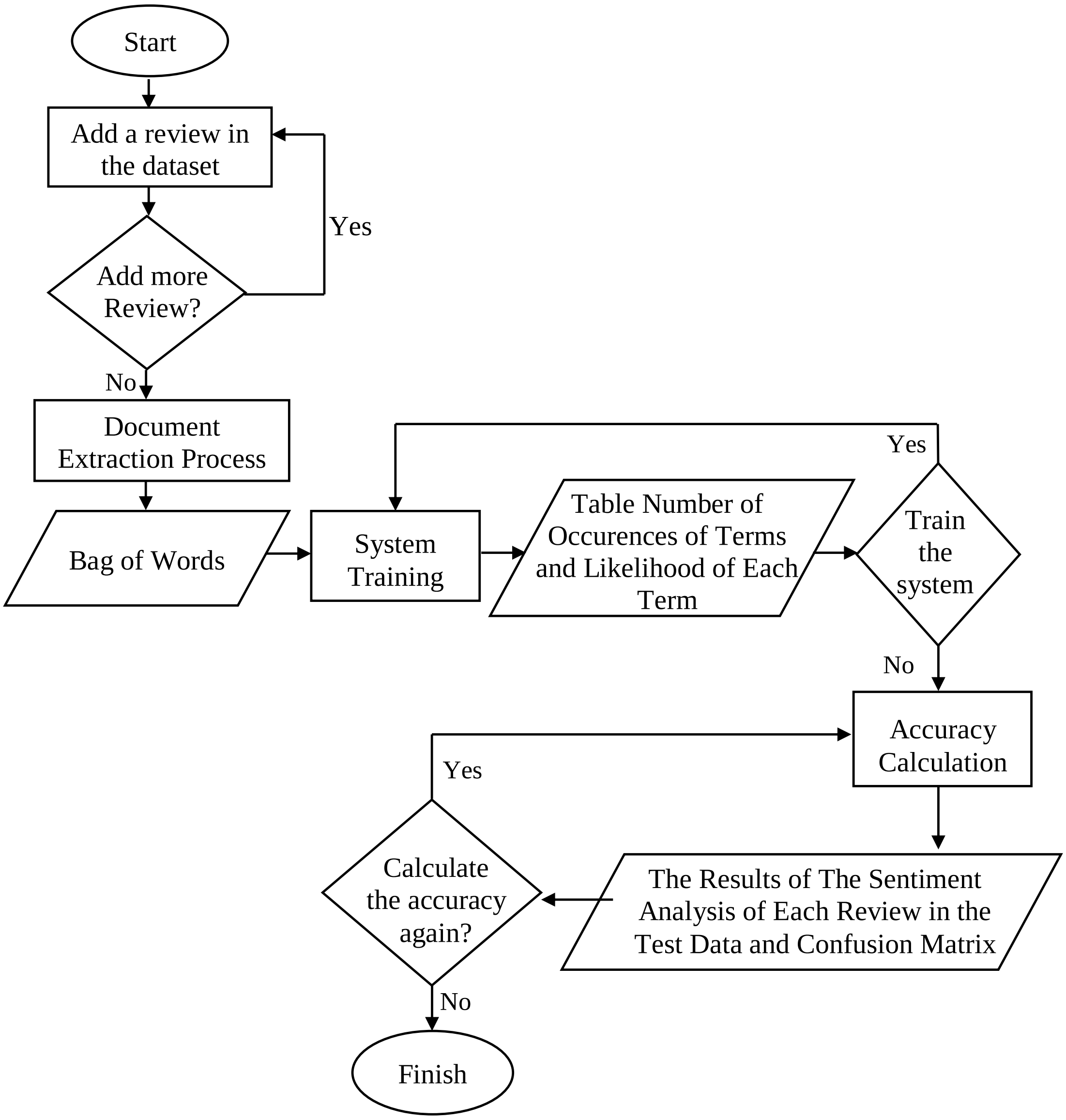
Start
Add a review in the dataset
Add more Review?
Yes
No
Document Extraction Process
Bag of Words
System Training
Table Number of Occurences of Terms and Likelihood of Each Term
Train the system
Yes
No
Accuracy Calculation
The Results of The Sentiment Analysis of Each Review in the Test Data and Confusion Matrix
Calculate the accuracy again?
Yes
No
Finish

Figure 2. Overview of Film Review Sentiment Analysis System

The explanation on Figure 2. above is: firstly, the derived from the movie review is submitted by the user one by one. The submitted review may be categorized into training data or test data based on the user's wishes. The system, then, executes the document extraction process in order to generate a collection of terms from each review text and so-called bag of words. The next step is to train the system. At this step, the system constructs a model of the training data by calculating the occurrences of each term in the training data and likelihood value [15]. The result of the system training process is a table of feature sets containing the number of occurrences and likelihood values of each term in the training data. The likelihood value is, then, incorporated into the next step which is the calculation of the system accuracy [16].

The system performs the classification process using the Naive Bayes Classifier method in order to obtain the best class of sentiment of each review in the category of data train. The results of the accuracy calculation are the sentiments of the classification of each test data review and the confusion matrix table obtained from the original sentiment comparison of the test data review and the sentiment of the analysis result. System accuracy can be concluded from the confusion matrix. The planning steps which are implemented based on the waterfall method [17].

2.3 Requirement Analysis

System requirement is the ability of the system to meet the condition desired by the users. System requirement analysis is performed by grouping the needs into functional requirements and non-functional requirements.

2.3.1 Functional Requirements

The functional requirements of the Sentiment Analysis System for Moview Review are as follows:

1. The system can manage (add, read, update, and delete) movie review datasets,
2. The system can manage (add, read, update, and delete) the dictionary of root words in Bahasa Indonesia,
3. The system can manage (add, read, update, and delete) the dictionary of stop words,
4. The system can perform the classifier training process and display the model in the form of feature sets of the term data from the training data,
5. The system can display the test data result and display confusion matrix generated from the classifier testing,
6. The system can display a set of movie review dataset terms derived from tokenizing, filtering, and stemming processes, and
7. The system can display sentiment analysis result derived from reviews submitted by users.

2.3.2 Non-Functional Requirements

Meanwhile, the non-functional requirements of the Sentiment Analysis System for Movie Review are as follows:

1. The system uses an authentication through login process in order to differentiate user level.
2. The system can run in various web browsers which support the system environment,
3. The system gives a fast response, and
4. The system has a user-friendly interface design.

2.4 System Design

The design of the Sentiment Analysis System for Movie Review uses Unified-Modeling Language (UML) as the modeling language and object-oriented programming concept. The documentation of the created system includes:

2.4.1 Business Process

Business process diagram describes a series of inputs, needs, and outputs processed by the system in order to produce the desired destination by the user. Description of business process system can be seen in Figure 3.

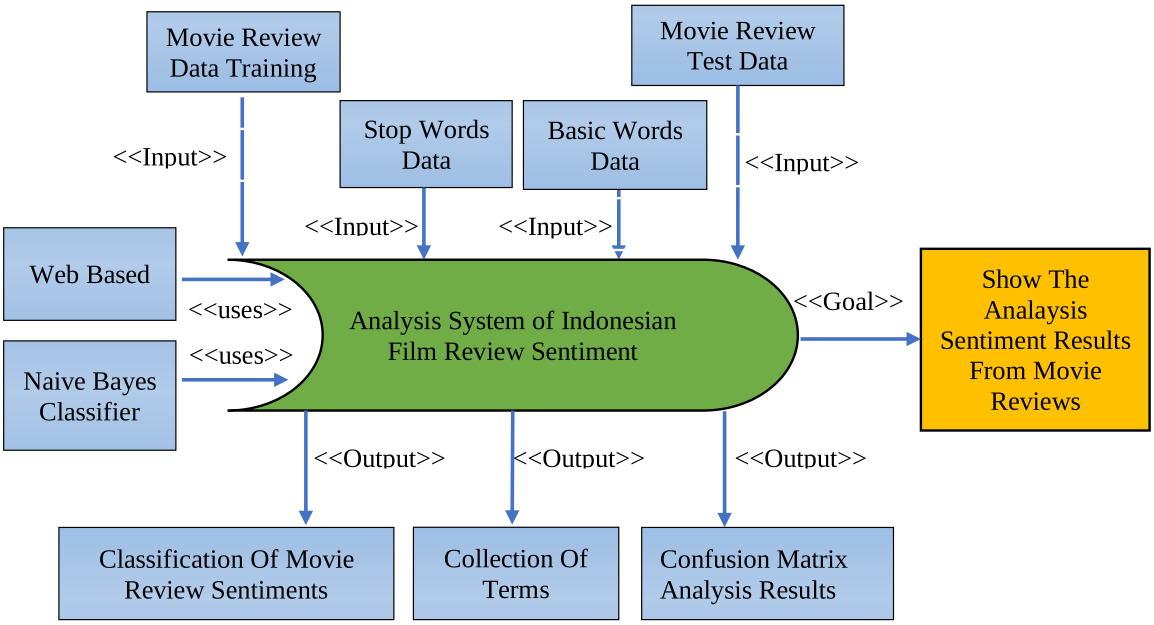
Movie Review Data Training
Movie Review Test Data
Stop Words Data
Basic Words Data
<<Input>>
<<Input>>
<<Input>>
<<Input>>
Web Based
Naive Bayes Classifier
<<uses>>
<<uses>>
Analysis System of Indonesian Film Review Sentiment
<<Goal>>
Show The Analaysis Sentiment Results From Movie Reviews
<<Output>>
<<Output>>
<<Output>>
Classification Of Movie Review Sentiments
Collection Of Terms
Confusion Matrix Analysis Results

Figure 3. Bussiness Process Movie Sentiment Analysis System Review

2.4.2 Use case Diagram

Use case diagram describes the system's functional needs in the form of diagram. Use case is needed to describe typical interactions between users and system. Use case diagram of the Sentiment Analysis System for Movie Review is illustrated in Figure 4.

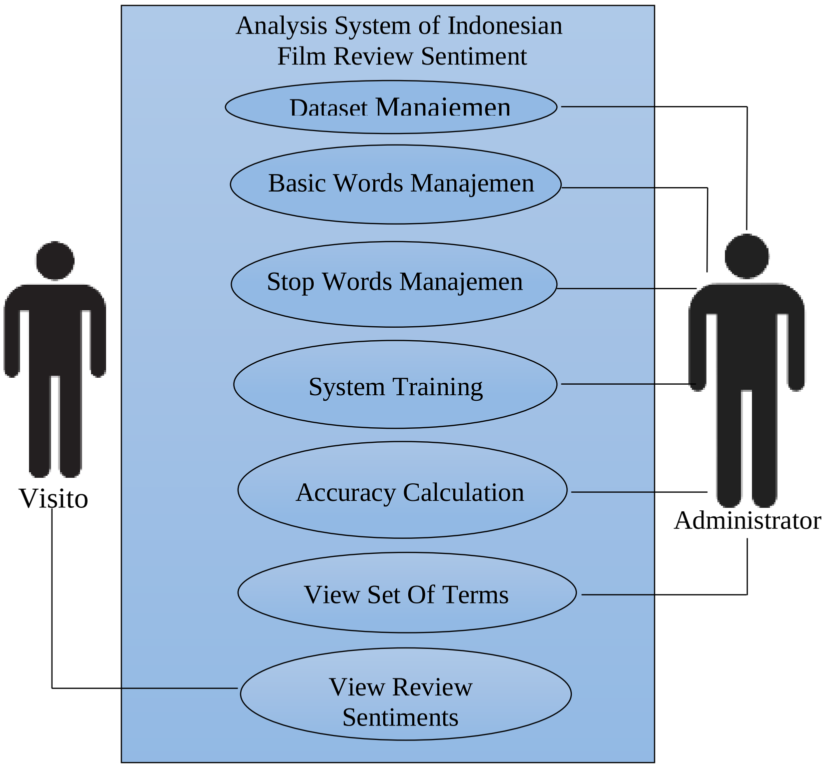

Figure 4. Use Case Diagram Sentiment Analysis Movie Review System

2.4.3 Deployment Diagram

Deployment diagram is a diagram illustrating the physical layout of the system. This diagram is included in order to describe the hardware and software required for the system to run according to the system needs. Deployment diagram of the Sentiment Analysis System for Movie Review can be seen in Figure 5.

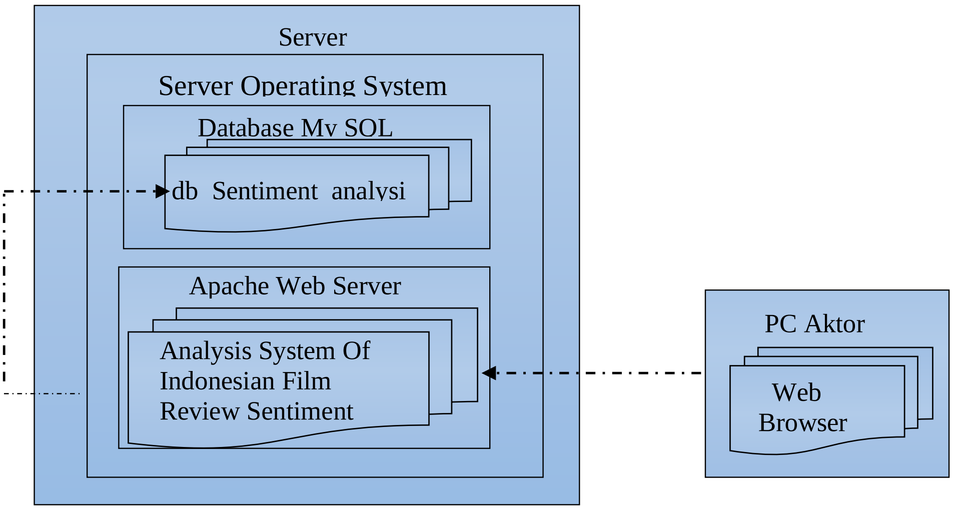

Figure 5. Deployment diagram of the Sentiment Analysis System for Movie

Sentiment Analysis System for Movie Review is developed as a web-based application. The webpage of Sentiment Analysis System for Movie Review is divided into two levels based on the user level: private page for the administrators and public page for the visitors. The private page contains the management dataset menu of the movie review, the stop words management menu, the training system menu, the accuracy calculation menu, and menu containing the set of terms from each movie review on the dataset. The public page contains an interface for visitors wanting to obtain sentiment analysis result from the movie reviews within the system. The main features displayed in the discussion are the dashboard page, system training page, system accuracy calculation page for the visitors.

2.5 System Interface Implementation`

The dashboard page is the first accessed page once the administrator has successfully logged in to the system. The dashboard contains brief information on datasets, such as the number of review for each category, the ratio of training and test data as well as the ratio of positive and negative training data. The information is displayed in the form of numbers and graphs of donut charts. System training page contains a table on the occurrences and likelihood values of each term in all train data. The process of system training in order to form feature sets is performed by the administrators through the interface page by selecting the button of "Latih".

The accuracy calculation page contains a list of reviews which become test data within the dataset, probability of the positive sentiment and negative sentiment for each test data, the result of the class analysis with the highest probability, and the accuracy value of the analysis result towards the original sentiment of the reviews. In addition to the review table of the test data, there are accuracy values of the system's last accuracy test result which have been converted into percentage format and confusion matrix table of the last accuracy test result. Accuracy calculation process is also initiated by the administrators through the interface page by selecting the "Hitung (calculate)" button.

The visitor page is the page which is for the visitors. The visitors can access this page in order to enter the review data they want to know the sentiment of. After selecting the "Lihat Sentiment (view sentiment)" button, the system classifies using the Naive Bayes method in order to obtain positive and negative sentiment probabilities as well as the sentiment result obtained based on the probability ratio of the two sentiment classes.

2.6 Accuracy Calculation Result

The testing process is conducted five times. The first test was conducted by selecting 600 reviews as the training data and 200 reviews as the random test data. Further tests were performed by adding 100 movie reviews as the test data and test data of 20 movie reviews into the dataset utilized for the previous test. Details of the composition of each training and test data for each test process can be seen in Table 1.

Table 1. Composition each training and test data for each test process

Testing no.	Number of Training Data	Number of Testing Data	Number of Positive Training Data	Number of Negative Training Data	Number of Positive Testing Data	Accuracy (%)
1	600	120	390	210	78	87,50%
2	700	140	454	246	91	87,86%
3	800	160	520	280	104	88,13%
4	900	180	585	315	117	88,33%
5	1000	201	650	350	132	90,05%

Based on the five times accuracy calculation processes of the Sentiment Analysis System for Movie Review, the obtained accuracy data for each process is available in the following table 2.

Table 2. Obtained accuracy data for each process

Testing no.	Number of Training Data	Number of Testing Data	Accuracy (%)
1	600	120	87,50%
2	700	140	87,86%

3	800	160	88,13%
4	900	180	88,33%
5	1000	201	90,05%
	Average Value:		88,37%

3. Conclusions

The Sentiment Analysis System for Movie Reviews implements the method of Naive Bayes Classifier in order to obtain the highest posterior probability value of the two review classes of sentiment. The posterior probability value is obtained from the total sum of log prior probability and log likelihood of each term in each review of the training data for each sentiment class. The average accuracy produced by the Sentiment Analysis System for Movie Review is 88.74% out of five time testing processes. The first test was conducted by training 600 movie reviews as the training data and classifying 120 movie reviews as the random test data. For each test, the dataset from the previous conducted test is added with 100 movie reviews as the training data and 20 movie reviews as the test data. The highest accuracy is obtained when the dataset used consists of 1,000 review of training data and 201 test data review and the percentage is 90.05%. The accuracy value is directly proportional to the number of reviews used as the dataset.

References

[1] Grimes S 2008 *Unstructured Data and 80 Percent Rule* Carabridge Bridgepoints

[2] Feldman R and Sanger J 2007 *The Text Mining Handbook* Advanced Approaches in Analyzing Unstructure Data Cambridge: Cambridge University Press.

[3] Pang B and Lee L 2002 Sentiment Classification using Machine Learning Techniques *Proceeding of ACL-02 Conference on Empirical Methods in Natural Language Processing* **10** 79 - 86

[4] Kalaivani P Sentiment Classification of Movie Reviews by supervised machine learning approaches et.al *Indian Journal of Computer Science and Engineering* (*IJCSE*) **4** 4

[5] Rambocas M and Gama J 2013 *Marketing research: The role of sentiment analysis* 489 Universidade do Porto Faculdade de Economia do Porto

[6] Vinodhini, G. and Chandrasekaran, R.M., 2012. Sentiment analysis and opinion mining: a survey. *International Journal*, *2*(6), pp.282-292.

[7] Prabowo R and Thelwall M 2009 Sentiment Analysis: A Combined Approach *Journal of Informatics* **3** 1 143–157

[8] Zulkifli A 2005 *Manajemen Sistem Informasi* Jakarta: Gramedia Pustaka Utama

[9] Blitzer J, Dredze M, Pereira F 2007 Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification *In: Proceedings of ACL* **7** 440 - 447

[10] Kusrini and Lutfi E T 2009 *Algoritma Data Mining* Yogyakarta: Penerbit Andi.

[11] Pakpahan D and Widyastuti H 2014 Aplikasi Opinion Mining dengan Algoritma Naïve Bayes untuk Menilai Berita Online *Jurnal Integrasi Program Studi Teknik Informatika Politeknik Negeri Batam* **6** 1-10

[12] Candra T 2009 *Metode Pembobotan Statistical Concept Based untuk Klastering dan Kategorisasi Dokumen Berbahasa Indonesia* Tugas Akhir. Bandung: Departemen Teknik Informatika Institut Teknologi Telkom Bandung

[13] Tala F Z 2003 *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia* Tesis Amsterdam: Institute for Logic, Language and Computation Universiteit van Amsterdam

[14] Agusta L 2009 Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief & Adriani untuk Stemming Dokumen Teks Bahasa Indonesia *Konferensi Nasional Sistem dan Informatika 2009* 196 – 201

[15] Dhande L L and Patnaik G K 2014 Analyzing sentiment of movie review data using Naive Bayes neural classifier *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)* **3** 4

[16] Kohavi R and R. Provost. 1998. Glossary of Terms. *Machine Learning* **30** 271-274

[17] Rizky S 2011 *Konsep Dasar Rekayasa Perangkat Lunak (Software Engineering)* Jakarta: Prestasi Pustaka