International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-5C, May 2019

Naïve Bayes Classifiers For Tweet Sentiment Analysis Using GPU

Islamiyah, Nataniel Dengen, Eny Maria

ABSTRACT--- The use of computers to solve problems has been done for all areas of work. Along with this, demanded faster computing process. To perform sentiment analysis of data obtained from the internet. Data taken from micro-blogging which at this time became the most popular communication tool and favored by internet users. The method used to construct the classification model of training data in this research is Naive Bayes Method. Training data is collected by utilizing the crontab facility with query emoticons and national media accounts linked to the Twitter API. The collected data will pass certain preprocessing before the training. The weighting feature used is the term frequency with TF-IDF. All data used in this research is a tweet that is delivered in Bahasa Indonesia. From the implementation results obtained 96.61% accuracy for sequential classification conducted using GPU GeForce 930M.

Keywords: GPU; Sentiment Analysis; Microblogging

I. INTRODUCTION

The use of computers to solve problems has been done for all areas of work. This is because computing is considered to be faster in solving problems than manual completion [1]. Along with this, demanded faster computing process. So to increase the speed of computing can be done in two ways, namely the increase in hardware speed and software speed increase. Rare and costly multiprocessor computers make the existing parallel algorithm difficult to implement, apart from the application of parallel computing too difficult to run. The shaping of the application aims to show an increase in speed obtained from the sequential parallelization algorithm. So, to overcome it can be done by designing a pseudo-parallel machine. Parallel machines designed can be done in several ways, with message passing interface, a computer network, or by using GPU graphics card [1]–[3].

The availability of large enough data can be utilized for text mining that refers to the process of retrieving high quality information from the text. High quality information is usually obtained through forecasting patterns and trends through means such as statistical pattern learning. Text mining is an interesting topic to be studied and processed at this time, due to the availability of many text documents and ease in obtaining data, in an effort to provide better information. Organizing textual data for users, researchers have explored the issue of categorizing text automatically [4]. The Naive Bayes method is often called the Naive Bayes Classifiers (NBC). Previous research referred to this study using NBC in his research to perform tweet sentiment analysis. NBC has advantages compared with other method algorithms, because the algorithm used is simple but has a high accuracy [5], [6]. Sentiment analysis is the process of automatically understanding, extracting and processing textual data for information [7]. Sentiment analysis attempts to gather an overall opinion on the comments, for example micro-blog companies trying to learn the reactions of users to get a general sense of their products [8]. Twitter as a micro-blogging builds a model to classify "Tweet" into positive, negative and neutral sentiments [9].

II. EXPERIMENTAL DETAILS

Development of Naïve Bayes Classifiers is done by define attributes, classes and data to be processed first. Each data sample is represented by a set of n-dimensional eigenvector attributes: X = {X_1, X_2, ..., X_n}, with each n data attribute x_1, x_2, ..., x_n. It is assumed that there are some classes of m V_1, V_2, ..., V_m, and given an unknown sample data S. If represented in table form then the value of an attribute, class and sample data can be seen in Table 1.

Table 1: Attributes, classes and naïve Bayes classifiers sample data

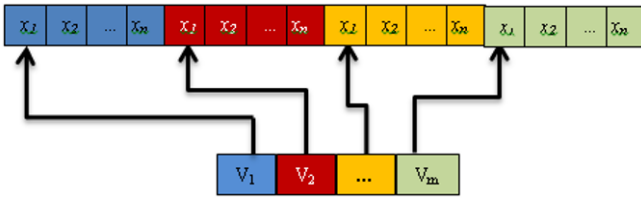
Sample	V_1				V_2				…				V_m			
	x_1	x_2	…	x_n	x_1	x_2	…	x_n	.	.	.	.	x_1	x_2	…	x_n
S_1	1	0	…	0	1	0	…	0	.	.	.	1	0	…	0	1
S_2	0	0	…	0	0	0	…	0	.	.	.	0	0	…	0	0
S_3	1	1	…	0	1	1	…	0	.	.	.	1	1	…	0	1
S_4	1	0	…	0	1	0	…	0	.	.	.	1	0	…	0	1
S_5	1	1	…	0	1	1	…	0	.	.	.	1	1	…	0	1
S_6	1	0	…	0	1	0	…	0	.	.	.	1	0	…	0	1
.	0	1	…	1	0	1	…	1	.	.	.	0	1	…	1	0
.	0	1	…	1	0	1	…	1	.	.	.	0	1	…	1	0
.	0	1	…	1	0	1	…	1	.	.	.	0	1	…	1	0
S_{10}	0	1	…	0	0	1	…	0	.	.	.	0	1	…	0	0

Table 1 is known that there are as many as 10 data samples with a number of m classes, with each class having a number of n attributes. Based on the information obtained then, the next process is to conduct the process of training and testing for sample data in determining the class document.

Manuscript published on 30 May 2019.
* Correspondence Author (s)
Islamiyah, Faculty of Computer Science and Information Technology, Mulawarman University, East Kalimantan, Indonesia(E-mail: islamiyah.unmul@gmail.com)
Nataniel Dengen, Faculty of Computer Science and Information Technology, Mulawarman University, East Kalimantan, Indonesia
Eny Maria, State Polytechnic Agricultural of Samarinda, East Kalimantan, Indonesia

Development of Naïve Bayes Classifiers using GPU is done first for training process. Based on the data in Table 1, the data representation will be used using the vector approach with vector attribute sample as list attribute of

each category. Representation of vector training data can be seen in Figure 1



Summary of algorithms for training process. Inputs are already known samples of their categories.

a. Attributes ← the set of all attributes of all the training samples.

b. For each category V_j do :

- $sample_j$ ← the sample set that is in the category V_j.
- Count $P(V_j)$ by using Eq (1).

$$P(V_j) = \frac{|sample\ j|}{|sample|} \quad (1)$$

c. For each member n_k on attribute do:

- Count $P(x_i|V_j)$ by using Eq (2).

$$P(x_i|V_j) = \frac{n_k+1}{n+|attributes|} \quad (2)$$

Where, n_k is number of times of occurrence; n is a number of occurrence occurrences of each category; $|attributes|$ is the sum of all attributes of all categories; $|sample_j|$ is the amount of data in the category *j*; $|sample|$ is the total number of training data of all categories. At a later stage all probability values are found and stored. This probability value to be transferred to GPU for VMAP value search specifies the sample category.

III. RESULT AND DISCUSSION

The results of the experimental program in this study found several experiments with the results entered on the confusion matrix that will be used to determine the accuracy of the classifier with performance metric accuracy. The accuracy of the classifier in the dataset can be calculated using the following Equation (3).

$$Accuracy = \frac{the\ number\ of\ correct\ classifications}{Data} \quad (3)$$

The first accuracy test is done by labeling the sentiment class on the test data first. Tested test data for positive sentiment 56 tweets, negative sentiment 98 tweets and 20 neutral sentiments. Test data is net data after preprocessing, which is 177 tweets selected and labeled sentiments from the net data owned by a total of 154,503 tweets.

The sentiment analysis was conducted sequentially and found positive sentiments of 55 tweets, negative sentiment of 93 tweets and 28 tweet neutral sentiments. The results of this test can be seen in Figure 2. The process of Naïve Bayes Classifiers for Tweeted Sentiment Analysis using GPU has been done faster than sequentially sequenced classification.

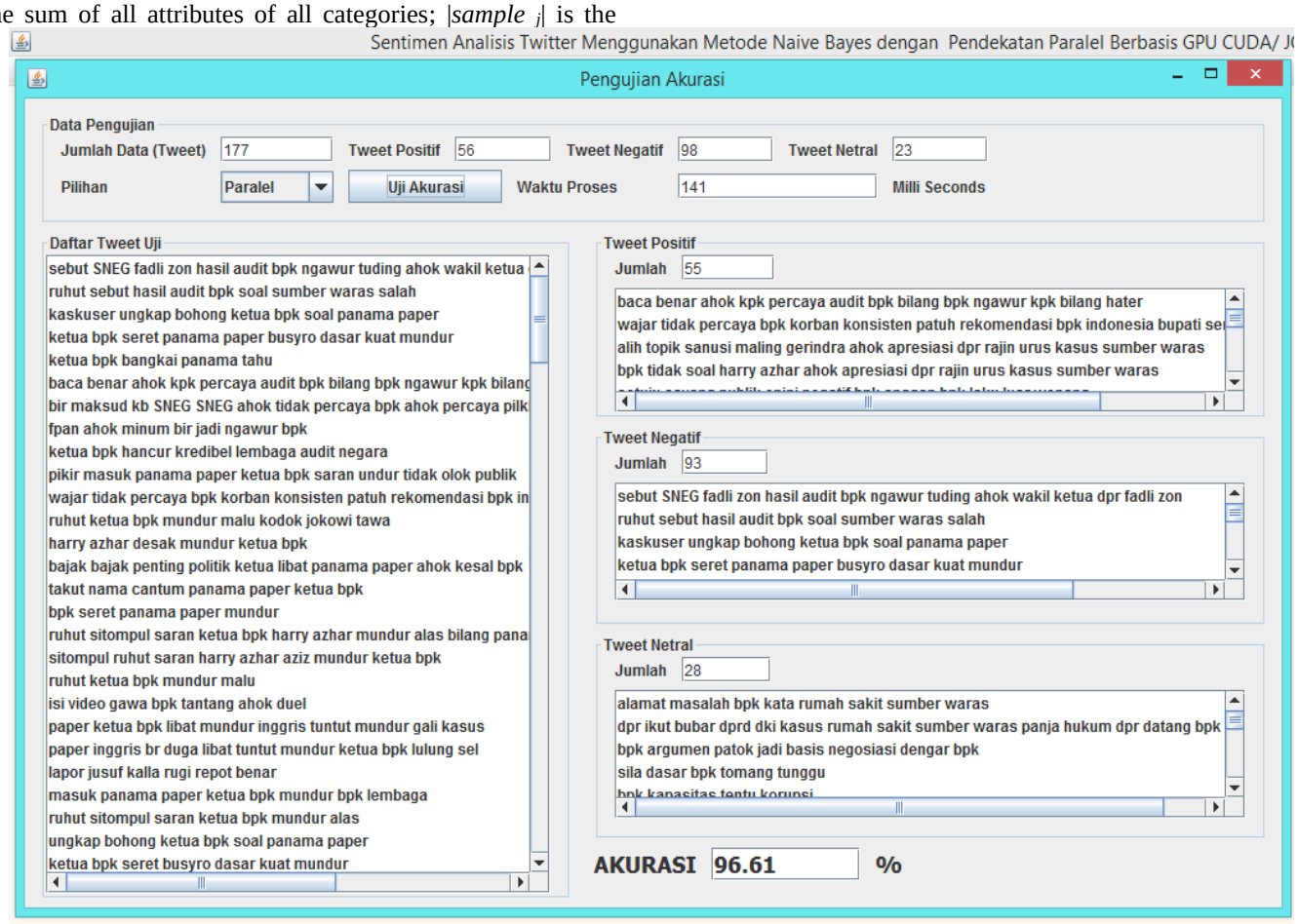


Fig. 2: Sentiment Analysis

IV. CONCLUSION

Based on the process of Naïve Bayes Classifiers for Tweeted Sentiment Analysis using GPU has been done faster than sequentially sequenced classification, then faster processing times compared to sequential processes. Previous research on tweet analysis of sentiments by Naïve Bayes Classifiers method sequentially to perform the test set takes 1102 seconds, when compared with this study for the average time of process conducted in parallel only 108,570.66 ms or 108.57 seconds. So it can be concluded that this study is 10 x faster than previous research about tweet sentiment analysis with Naïve Bayes Classifiers method that is done sequentially.

REFERENCES

1 D. B. Kirk and W. H. Wen-Mei, *Programming massively parallel processors: a hands-on approach*, Third edit. Morgan kaufmann, 2016.
2 N. Matloff, *Programming on parallel machines*. University of California, 2011.
3 C. Böhm, R. Noll, C. Plant, B. Wackersreuther, and A. Zherdin, "Data mining using graphics processing units," in *Transactions on Large-Scale Data-and Knowledge-Centered Systems I*, 2009, pp. 63–90.
4 R. Inam, "A* Algorithm for Multicore Graphics Processors," Chalmers University of Technology, 2010.
5 R. Adawiyah and J. Nugraha, "Sentiment Analysis on Mobile Banking Application Using Naive Bayes Classifer and Association Methods," *Int. J. Eng. Technol.*, vol. 7, no. 4.15, pp. 244–247, 2018.
6 D. S. Reddy, C. N. Harshitha, and C. M. Belinda, "Brain tumor prediction using naïve Bayes' classifier and decision tree algorithms," *Int. J. Eng. Technol.*, vol. 7, no. 1.7, pp. 137–141, 2018.
7 A. I. Wicaksono, E. Nio, and S. H. Myaeng, "Unsupervised Approach for Sentiment Analysis on Indonesian Movie Review," in *the 6th Conference of Indonesian Students Association in Korea (CISAK-2013)*, 2013.
8 E. Boiy, P. Hens, K. Deschach, and M.-F. Moens, "Automatic Sentiment Analysis in On-line Tex," in *Proceedings ELPUB2007 Conference on Electronic Publishing*, 2007, pp. 349–360.
9 A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, 2011, pp. 30–38.