COMPARING THE CLASSIFICATION METHODS OF SENTIMENT ANALYSIS ON A PUBLIC FIGURE ON INDONESIAN-LANGUAGE SOCIAL MEDIA

[1]DERISMA, [2]DODON YENDRI, [3]MEZA SILVANA

[1,2]Department of Computer System, Universitas Andalas, Padang, Indonesia

[3]Department of Information System, Universitas Andalas, Padang, Indonesia

E-mail: [1]derisma@fti.unand.ac.id, [2]dodon@fti.unand.ac.id, [3]mezasilvana@gmail.com

ABSTRACT

Social media are used by people as the platform to express their opinions and the conditions that happen around them. During the ministerial election for 2019-2024 Cabinet of Indonesia Maju, the choice of ministers of President Indonesia Jokowi always present in the discussed topic on a social media. The most discussed topic is the appointment of Nadiem Makarim as the Minister of Education and Culture. In Indonesia, the ministers have to show their performance and capability. If they failed to do so, they would be reshuffled by the President. One of the sources of information required by President to be able of evaluating the performance of his ministers is through the feedback from citizens. Sentiment analysis is a computation study of opinions, attitudes, and emotions of people toward entities, individuals, issues, events, or topics. Sentiment Analysis The targets of sentiment analysis are to discover arguments, identify the expressed sentiments, and then classify their polarity into positive, negative, or neutral categories. In principle, the classification methods on sentiment analysis can be performed through machine learning approach. Some classification methods of machine learning such as Decision Tree, K-NN, Naïve Bayes, and Random Forest are often used to acquire the best result. Next, several stages were done, including the data collection by crawling Twitter's data via API with "Nadiem Makarim" as the keyword, pre-processing, classification and evaluating the classification performance. Naïve Bayes acquired the best result with 99% accuracy, 94% precision, and 99% Recall. It can be concluded that Naïve Bayes is the best classifier to be used on the dataset of Indonesian-Language social media because it can provide the most accurate and appropriate prediction.

Keywords: *Sentiment Analysis, Machine Learning, Text Classification, Twitter*

1. INTRODUCTION

Sentiment analysis is a computation study of opinions, attitudes, and emotions of people toward entities, individuals, issues, events or topics [1]. The targets of sentiment analysis are to discover arguments, identify the sentiments expressed by them, and then classify their polarities in positive, negative, or neutral categories [2]. Sentiment analysis is a popular field research because it can provide advantages to various aspects, starting from e-commerce [3][4][5][6], politics [7][8], Hotel Reviews [9][10], e-learning [11], Personality Detection [1], and the decision making of investors [6][13][14]. One of the important issues is that opinions can be manifested in various languages (English, Urdu, Arabic, etc.). Overcoming every language based on their orientation is a challenging task. Most of the research works in sentiment mining were done in English. At present, limited studies are being performed on the classification of other languages, such as Arabic [1][15][1], Urdu and Hindi [16]. In this writing, three classification models were used to classify texts by using Waikato Environment for Knowledge Analysis (WEKA).

After the announcement of the members of Indonesia Maju Cabinet, *Nadiem Makarim*; the Minister of Education and Culture became the most lively discussed minister. He is far ahead of *Prabowo Subianto* or the past ministers like *Sri Mulyani Indrawati* or *Basuki Hadimuljono*. The opinions or neutral sentiments toward the founder of Gojek reached 17,148 tweets. There were 9,410 tweets of negative sentiments and 3,294 tweets of positive sentiments. Besides lively discussed, *Nadiem* is the most anticipated minister. Netizens cant wait to see the breakthrough of Nadiem in national education sector (jawapos.com). Next Policy—a think-tank institution has held a sophisticated survey, i.e. *Analisis Media Social Nusantara Berbasis Artifical Intelligence (AMENA*) which stated that *Nadiem* received high attention from the netizens of Twitter. Based on that survey, the positive opinions toward Nadiem reached 3,294 and the negative reached 9,410. However, the score of neutral sentiment reached 17,148. Twitter was selected as the research object because it's considered as the most influential social media in political sector. One of the results showed that public put their high attention and wait for the breakthrough of Nadiem Makarim as the Minister of Education and Culture (republika.co.id).

In principle, the classification methods on sentiment analysis are divided into two, namely machine learning-based and lexical-based approaches. Machine learning approach depends on the training data of several researchers such as Support Vector Machine [17], Naïve Bayes [18], Random Forest [17], and CNN [19]. The lexicon approach is an approach by using sentiment lexicon containing words of opinion and comparing them to data to identify a value of a word [7].

Several researchers have done the comparison of several algorithms on some datasets. The study done by Bansal, B [8] compared the algorithms of SVM, Naïve Bayes, Logistic Regression, and Random Forest. The examination result showed that Random Forest acquired the most superior accuracy. The most proper model for Indonesian-language sentiment analysis has yet to be found of all research results that have been done. This research aimed to seek the best algorithm from Twitter's sentiment analysis in categorizing Indonesian-language positive, negative, and neutral sentiments. Therefore, the author will compare some machine learning algorithms such as Decision Tree, K-NN, Naïve Bayes, and Random Forest as the input of national policy planning based on public opinions. The magnitude of sentiments aimed toward Ministry of Education and Culture can be made as an consideration alternative within the reshuffle done by the Indonesia President.

2. THE MATERIAL AND METHOD

The methodologies used in this research were experimental and case study researches.

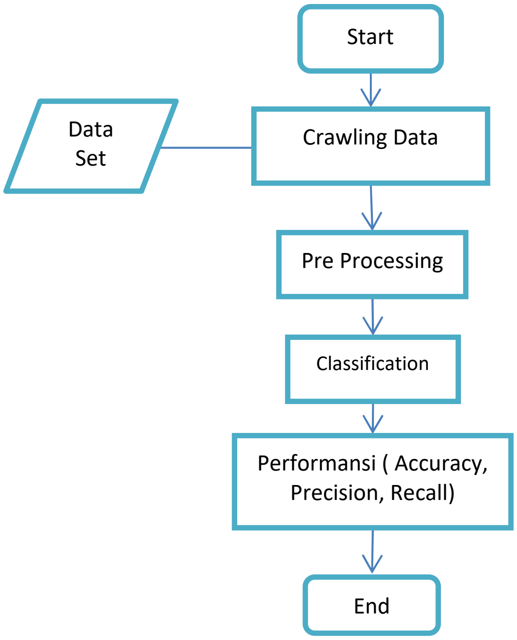


Figure 1: Flow Chart of Sentiment Analysis

2.1 Data Collection

Data were collected through Application Program Interface (API) of Twitter from October 23 to November 23 2019 that amounted to 569 tweets. Nadiem Makarim was used as the keyword. Of 569 tweets; 485 neutral, 39 negative, and 45 positive were obtained.

Table 1: Example Of The Label Settings

N0	Text	Sentiment
1	RT @smpn3bayat: *Empat program prioritas Mendikbud Nadiem Makarim. Semoga berhasil dan merata di seluruh Indonesia.* https://t.co/8IBgUcadGx	Neutral
2	RT @blogger_eksis: *Kenapa Nadiem Makarim yang jadi Menteri Pendidikan? Bukannya* @AdamasBelva *aja yang udah punya jejak di* @ruangguru	Negative
3	RT @NCLYS: *Jadi mungkin meletakkan Bung Nadiem buat jadi ujung tombak pendidikan, atas dasar pengalamannya di industri digital, nggak ngaco…?*	Negative
4	@WikiDPR #kom10 Rano @PDI_Perjuangan #Banten3: *jgn khawatir bang Nadiem kita semua pasti mendukung anda, krn ketika…* https://t.co/zqk3MiS0xB	Positive
5	*Ini baru Mentri yg hebat..berilmu tp tak sombong..bkn ky yg ono udh bnyk gaya banyak bacot tp ngga ada hasil...ma…* https://t.co/0IGMIwIg3q	Positive

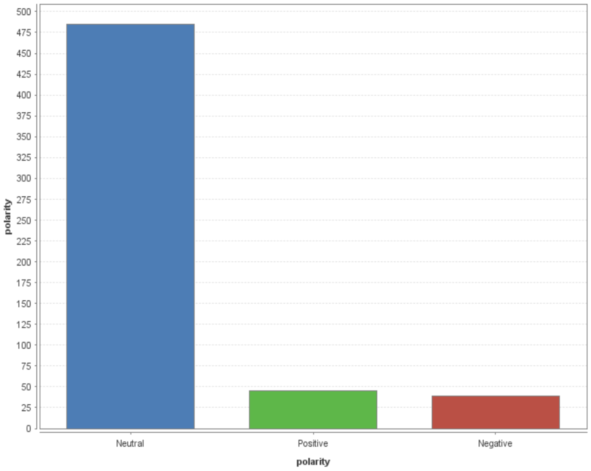


Figure 2: The bar chart of tweet sentiments

2.2 Preprocessing

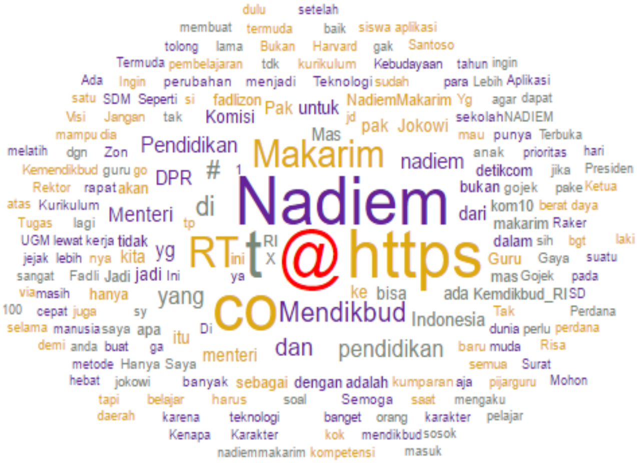


Figure 3: The world-cloud of Nadiem Makarim

Before the learning process was done, pre-processing was performed prior to the classification process for the dimension of vector space model to be narrower. By narrowing the dimension of vector space model, the classification process will be faster. This preprocessing was done to equate words and reduce the volume of words. The stages in pre-processing phase were

2. 2. 1 Document Filtering

The process of document filtering aimed to clean tweets from unnecessary words or symbols to reduce the noise during the classification process. The eliminated words were:

a.□ Twitter Hashtag (#)
b.□ Twitter username (@username)
c.□ Site address (url)
d.□ Query as the search keyword

2. 2. 2 Case Folding

Texts exist within the document were changed into lower case

2. 2. 3 Tokenization

This process will divide a group of characters in a text into a unit of word. This action is done by differentiating specific characters which will be treated as the divider or not. For instance, characters of whitespace, such as enter, tabulation, and space which are considered as the word divider.

2. 2. 4 Stopwords Elimination

Stopwords can add the dimension of data on the classification process. The data dictionary of Stopwords that is generally used (consisted of yang (which, that, who, whom), di (in, at, on), ke (to), dari (of, from), etc.) will be added with twitter special stopwords, such as "wkkwk, "hihihi", "xoxoxox", and others. Complete data dictionary of twitter and stopwords of Bahasa are provided in the appendix. The data dictionary of twitter's stopwords of Indonesian Language/Bahasa was collected manually from twitter.

2. 2. 5 Stemming

Stemming is the transformation process of the word form into the basic word. This method which changes the word form to basic word is adjusting the structure of language used in the stemming process.

2.3 Data weighting

The terms that have gone through the stemming process then be calculated on their weights by using TF-IDF. Weights are aimed to provide scores on the frequency of the occurrence of a word. Term Frequency is a weighting concept by searching how frequent (the frequency) a term occurs within a document. Because every document has different lengths, a word is possible to occur more often in a long document compared to a shorter document. Therefore, term frequency is often divided by the length of the document (total words exist within the document). Whereas, Document Frequency is the number of document where a term occurs. The smaller occurrence frequency will make the weight to be smaller as well. All words within the calculation process of term frequency are as importance as others.

$$tf(i) = \frac{freq\ (t_1)}{\sum freq\ (t)}$$

................(1)

tf(i): the score of Term Frequency in a word within a document.
freq (ti): the occurrence frequency of a word in a document.
$\sum$ freq(t) : total number of words in a document

2.4 Classification of Sentiment Analysis

Classification is defined as a form of data analysis to extract a model which will be used to predict a class label [1]. The class in a classification is an attribute in a most unique dataset as the independent variable statistically [9]. Data classification consisted of two process, namely learning and classification stages. The learning stage was a stage in the formulation of classification model, whereas the classification stage was the stage of the utilization of classification model to predict the class label of a datum. There are many algorithms that can be used in classifying data, however, this study only compared four algorithms, i.e. Decision Tree, K-NN, Naïve Bayes Classifier, and Random Forest [21].

1) □Decision Tree

Decision tree is one of the most popular classification methods because it is easy to be interpreted by humans. Decision tree is a prediction model that uses tree structure or hierarchy structure.

$$Entropy = \sum_{v=0}^{1} -P.\log(P)$$

……......................(2)

□□

2) □K-NN

K-Nearest Neighbor (K-NN) is a method to classify objects or data tested into a class where the objects stand and which one is the closest to the test objects. If K is more than 1, the closest member of the learning set is selected and the tested object will be classified into the majority class through voting system [6].

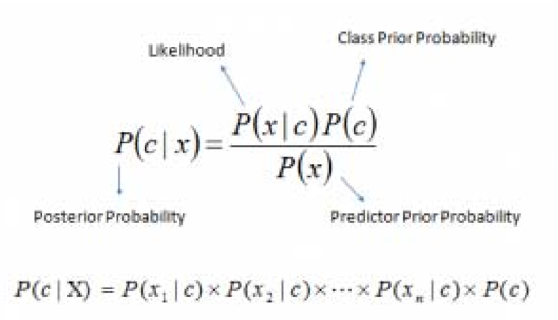
$$DE(x_i, x_j) = \sqrt{(x_i - x_j)^2 + (y_{xi} - y_{xj})^2}$$

……..(3)

□□ □□

3) □Naïve Bayes

Naïve Bayes is a classification method (supervised learning) through a probabilistic approach. This approach generates an assumption regarding how data can be produced by applying probabilistic model to embody them. Naïve Bayes classification is a simple method by assuming that whole attributes on data are not depending on each other based on the context of class [3]. The formula of Naïve Bayes's implementation on the text classification is as follows.



$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

P(c|x) is the posterior probability of class (target) given predictor (attribute).
P(c) is the prior probability of class.
P(x|c) is the likelihood which is the probability of predictor given class.
P(x) is the prior probability of predictor.

4) □Random Forest

Random forest is the development of Decision Tree by using some groups of Decision Tree in which each of them has been performed with drillings through the application of individual sample and each attribute is distributed to the selected tree between random subset attributes. And during the classification process, the individuals are based on the most votes in a group of population tree .

CART (Classification and Regression Tree) is a data exploration method that is based on decision tree technique. The classification tree is produced when the response changer is in categorical data, while the regression tree is produced when the response variable is in numeric data. The tree is formed through the process of binary recursive selection on a data cluster, thus, the score of response variable on each data cluster of the selection result will be higher.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \quad ....(5)$$

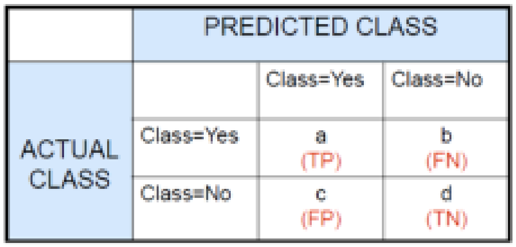
D. Evaluation of Sentiment Analysis

The measurement of performance for the classification was divided into several formulas, namely:

1) Precision is the comparison between accuracies among classifications requested by users with the answer given by the system.
2) Recall is the success rate in rediscovering the information.
3) Accuracy is the comparison between the number of correct classifications and the number of items that supposed to be in the class. The accuracy rate can be defined as the proximity level between actual and prediction values of the system.

This formula is used for the accuracy measurement in the classification

Table 2: Confusion Matrix



		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+TN+FP}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

3. RESULTS AND DISCUSSION

The conducted research used a computer with the following specifications; Intel Core i7 2.9 GHz CPU, 8Gb of RAM, and 64-bit Windows 10 Professional as the operating system. This study used RapidMiner 9.1 software. The data of this study used Twitter's data acquired from October 23 to November 23 2019 amounted to 569 tweets. After the stage on text preprocessing was done, the classification was performed through the algorithms of Decision Tree, K-NN, Naïve Bayes, and Random Forest. The next process was performing cross-validation. In this case, the cross-validation was performed to avoid overlapping on data testing. In this research, the author used a standard validation i.e. 10 folds cross-validation in which this process divided data randomly into 10 parts. The testing process was initiated by forming a model with the data in the first part. The established model will be tested to the rest nine parts. The next process after the testing was measuring the performance of the applied classification algorithms. In this study, the performance was measured through three methods, namely Accuracy, Precision, and Recall.

Table 3: Performance Of Different Classifiers.

Methods	Accuracy	Precision	Recall
Decision Tree	91%	64%	55%
K-NN	87%	82%	40%
Naïve Bayes	99%	94%	99%

Table II is the summary of the comparison results of classification algorithms in which Naïve Bayes acquires the best score. The classification on sentiment analysis is highly depending on the tested data.

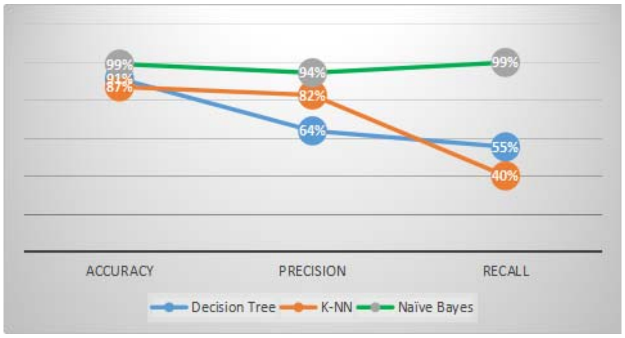


Figure 4: Visualization Performance classifier

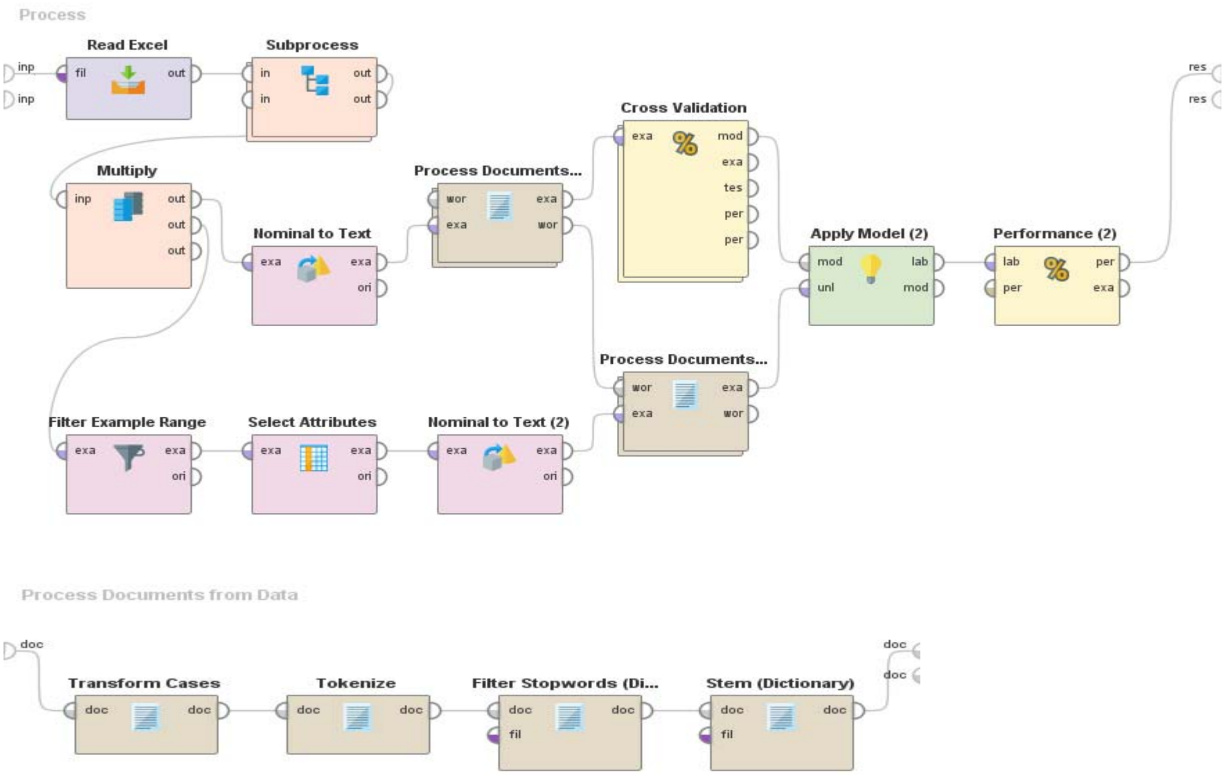


Fig. 5 Main and Preprocessing on Rapidminer

4. CONCLUSIONS

Based on the comparison results of the classification algorithms between Decision Tree, K-NN, Naïve Bayes, and Random Forest, Naïve Bayes acquired the best result with 99% accuracy, 94% Precision and 99% Recall. Naïve Bayes was the best classifier to be used with the dataset of Indonesian-Language social media because it gave the most accurate and correct prediction. According to the conducted study, a conclusion has been taken in which not all algorithms or classification results can classify text data properly. The selection of data amount, features, or attributes is also affecting the efficiency and effectiveness of the applied method. It is expected that the preprocessing performed on other studies can be optimized in further to acquire more quality data. Indonesian language gives many challenges due to its complex structures; various dialects, emoticons, and slangs. This opens new challenges for the researchers in using larger and more complex datasets with improvements on the number of label and the extent of public figure in standard and non-standard Indonesian language.

REFERENCES

[1] Heikal, M. (2018). ScienceDirect Sentiment Analysis of Arabic Tweets using Deep Learning Sentiment Analysis of □ Arabic Tweets using Deep Learning. Procedia Computer Science, 142, 114–122. https://doi.org/10.1016/j.procs.2018.10.466

[2] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications : A survey. Ain Shams Engineering Journal, 5(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

[3] Bansal, B., & Srivastava, S. (2018). ScienceDirect Sentiment classification of online consumer reviews using word vector representations. Procedia Computer Science, 132, 1147–1153. https://doi.org/10.1016/j.procs.2018.05.029

[4] Singh, A., Agarwal, A., & Dimri, P. (2018). ScienceDirect ScienceDirect Comparative Study of Machine Learning Approaches for Amazon Reviews. Procedia Computer

Science, 132, 1552–1561. https://doi.org/10.1016/j.procs.2018.05.119

[5] Potdar, A., Patil, P., Bagla, R., Pandey, R., & Prof, N. J. (2016). SAMIKSHA - Sentiment Based Product Review Analysis System. Procedia - Procedia Computer Science, 78(December 2015), 513–520. https://doi.org/10.1016/j.procs.2016.02.096

[6] Das, S., & Behera, R. K. (2018). ScienceDirect Real-Time Sentiment of Streaming for Stock Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction. Procedia Computer Science, 132(Iccids), 956–964. https://doi.org/10.1016/j.procs.2018.05.111

[7] Luca, E. De, Fallucchi, F., Giuliano, R., Incarnato, G., & Mazzenga, F. (2019). Analysing and Visualizing Tweets for U . S . President Popularity. 9(2), 692–699.

[8] Bansal, B., & Srivastava, S. (2018). ScienceDirect On On predicting predicting elections elections with with hybrid hybrid topic topic based based sentiment sentiment analysis analysis of tweets of tweets. Procedia Computer Science, 135, 346–353. https://doi.org/10.1016/j.procs.2018.08.183

[9] Akhtar, N., Zubair, N., Kumar, A., & Ahmad, T. (2017). ScienceDirect ScienceDirect Aspect based Sentiment Oriented Summarization of Hotel Reviews. Procedia Computer Science, 115, 563–571. https://doi.org/10.1016/j.procs.2017.09.115

[10] Aliandu, P. (2015). Sentiment Analysis to determine Accommodation , Shopping and Culinary Location on Foursquare in Kupang City. Procedia - Procedia Computer Science, 72, 300–305. https://doi.org/10.1016/j.procs.2015.12.144

[11] Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Computers in Human Behavior Sentiment analysis in Facebook and its application to e-learning. Computers in Human Behavior, 31, 527–541. https://doi.org/10.1016/j.chb.2013.05.024

[12] Sagadevan, S., Hashimah, N., Hassain, A., & Husin, M. H. (2015). Sentiment Valences for Automatic Personality Detection of Online Social Networks Users Using Three Factor Model. Procedia - Procedia Computer Science, 72, 201–208. https://doi.org/10.1016/j.procs.2015.12.122

[13] Ling, J., Hui, O., Hoon, G. K., Mohd, W., & Wan, N. (2018). ScienceDirect ScienceDirect Effects of Word Class and Text Position in Sentiment-based News Classification. Procedia Computer Science, 124, 77–85. https://doi.org/10.1016/j.procs.2017.12.132

[14] Khan, S. N., Nawi, N. M., Imrona, M., Shahzad, A., & Ullah, A. (2018). Opinion Mining Summarization and Automation Process : A Survey. 8(5), 1836–1844.

[15] Alhasani, H., Saad, S., & Kassim, J. (2018). Classification of Encouragement (Targhib) And Warning (Tarhib) Using Sentiment Analysis on Classical Arabic. 8(4), 1721–1727.

[16] Bilal, M., Israr, H., Shahid, M., & Khan, A. (2016). Sentiment classification of Roman-Urdu opinions using Naı ̈ ve Bayesian , Decision Tree and KNN classification techniques. Journal of King Saud University - Computer and Information Sciences, 28(3), 330–344. https://doi.org/10.1016/j.jksuci.2015.11.003

[17] Amrani, Y. A. L., Lazaar, M., Eddine, K., & Kadiri, E. L. (2018). ScienceDirect ScienceDirect Random Forest and Support based Hybrid on Vector Intelligent Machine Approach to Sentiment Analysis Random Forest and Support Vector Machine based Hybrid Approach to Sentiment. Procedia Computer Science, 127, 511–520. https://doi.org/10.1016/j.procs.2018.01.150

[18] Ullah, A., Khan, R., Khan, M., & Khan, M. B. (2016). Naïve Multi-label classification of YouTube comments using comparative opinion mining. Procedia - Procedia Computer Science, 82(March), 57–64. https://doi.org/10.1016/j.procs.2016.04.009

[19] Liao, S., Wang, J., Yu, R., Sato, K., & Cheng, Z. (2017). ScienceDirect ScienceDirect CNN for situations understanding based on sentiment analysis of twitter data. Procedia Computer Science, 111(2015), 376–381. https://doi.org/10.1016/j.procs.2017.06.037

[20] Sartika, D., Sensuse, D. I., Indo, U., Mandiri, G., & Komputer, F. I. (2017). Perbandingan Algoritma Klasifikasi Naive Bayes , Nearest Neighbour , dan Decision Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian. 1(2), 151–161.

[21] Bayhaqy, A. (n.d.). Sentiment Analysis about E-Commerce from Tweets Using Decision Tree , K-Nearest Neighbor , and Naïve Bayes. 2018 International Conference on Orange Technologies (ICOT), 1–6.