JOURNAL OF INFORMATION TECHNOLOGY AND ITS UTILIZATION, VOLUME 2, ISSUE 2, DECEMBER-2019: 29-33
E-ISSN 2654-802X

SENTIMENT ANALYSIS USING RANDOM FOREST ALGORITHM-ONLINE SOCIAL MEDIA BASED

Bahrawi

BBPSDMP Kominfo Makassar, Indonesia
bahrawi@kominfo.go.id

***Abstract*-- Every day billions of data in the form of text flood the internet be it sourced from forums, blogs, social media, or review sites. With the help of sentiment analysis, previously unstructured data can be transformed into more structured data and make this data important information. The data can describe opinions/sentiments from the public, about products, brands, community services, services, politics, or other topics. Sentiment analysis is one of the fields of Natural Language Processing (NLP) that builds systems for recognizing and extracting opinions in text form. At the most basic level, the goal is to get emotions or 'feelings' from a collection of texts or sentences. The field of sentiment analysis, or also called 'opinion mining', always involves some form of the data mining process to get the text that will later be carried out the learning process in the machine learning that will be built. this study conducts a sentimental analysis with data sources from Twitter using the Random Forest algorithm approach, we will measure the evaluation results of the algorithm we use in this study. The accuracy of measurements in this study, around 75%. the model is good enough. but we suggest trying other algorithms in further research**

***Keywords*: sentiment analysis; random forest algorithm; classification; machine learnings.**

I. INTRODUCTION

Sentiment analysis is part of text mining, the dataset that will be analyzed later can be sourced from the comments column, netizens tweets on Twitter, and various sources of uploads from people related to their opinions or sentiment on a matter. For people who work as data science, they may often hear the term about sentiment analysis. Sentiment analysis it's also processed from analyzing various data in the form of views or opinions so as to produce conclusions from various existing opinions. The result of sentiment analysis can be a percentage of positive, negative, or neutral sentiment

Sentiment analysis is useful for various problems of interest to human-computer interaction practitioners and researchers, as well as those from fields such as sociology, marketing and advertising, psychology, economics, and political science [1].

One from several social media which is widely used by society today is Twitter, Twitter has a simple and fast concept because the message is short [2]. Twitter as a social media is widely used by researchers in the field of natural language processing (NLP), in addition, concept simple text data and can be crawled, Twitter also provides an API facility that makes it easy for researchers to retrieve the data.

some previous research has been done with various classification algorithms. here are some of them :

An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis [3], uses six methods for classification namely lexicon-based classifier, NB, Bayesian Network, SVM (Support Vector Machine), C4.5 (Decision Tree), Random Forest and one method called the Ensemble Classifier which combines five methods (NB, Bayesian Network, SVM, C4.5, and Random Forest) to get higher accuracy. This study uses four classes, namely positive class (4288 tweets), negative (35876 tweets), neutral (40987 tweets) and irrelevant (26715 tweets). The accuracy of each when not combined with a two-class dataset (eliminating neutral and irrelevant classes) is Lexicon Based 67.9%, Naïve Bayesian 90%, Bayesian Network 91.4%, SVM 84.6%, Random Forest 89.8%. The Lexicon Based Method did not participate in the combination because its accuracy was at least 67.9%, the acquisition of ensemble accuracy with a two-class dataset was 91.7% while the ensemble's accuracy for the three-class dataset was 84.2%.

Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier [4], two supervised methods are used with two datasets namely film and hotel, the more training data that is entered the better the accuracy obtained in the NB algorithm with the dataset film but for the K-NN method, accuracy is obtained randomly.

Research on presidential candidates examined public opinion on the 2014 Indonesian presidential candidates [5], namely Prabowo-Hatta Rajasa and Joko Widodo-Jusuf Kalla. Research [5] uses NB for the classification of documents, the data in this study were taken in three periods, namely, before the legislative election, when the legislative election was held and after the declaration of the legislative election announcement then from the data the authors grouped public opinion whether positive, negative or neutral. The results are 90% accurate.

Text classification research with the Naïve Bayes algorithm for the Grouping of News Texts and Academic Abstracts [6]. Seven experiments were conducted for news documents and academic abstract documents, in the first experiment with the amount of training data and 9: 1 test data, the highest accuracy was compared with the smallest training data. The use of training data of 50% of the total data obtained an accuracy of

more than 75%.

Opinion Analysis Research on Smartphone Features on Indonesian Language Website Reviews [7]. Data collection is done by means of web scraping, which is taking data review from the target website. From the test results obtained an average value of recall and precision respectively of 0.63 and 0.72 while the accuracy of 81.76%.

Research from Faishol Nurhada, et al [5] dataset used is public timeline tweets taken by period. Using Twitter as a data source by utilizing the API features provided, retrieving data with retrieval techniques based on time periods.

Based on some of the previous studies that have been explained before, this research does the same thing, which is doing sentiment analysis of Twitter data using the Random Forest algorithm approach, we will measure the evaluation results of the algorithm that we use in this research.

II. METHODE

We will follow the typical machine learning pipeline. We will first import the dataset and we will then do exploratory data analysis to see if we can find any trends in the dataset. Next, we will perform text preprocessing to convert textual data to numeric data that can be used by a machine learning algorithm. Finally, we will use machine learning algorithms to train and test our sentiment analysis models.

Datasets

The dataset we used in this study taken from the website kaggle.com with CC BY-NC-SA 4.0 license. This data was originally posted by Crowdflower last February and includes tweets about 6 major US airlines. Additionally, Crowdflower had their workers extract the sentiment from the tweet as well as what the passenger was disappointed about if the tweet was negative. As the original source says, a sentiment analysis job about the problems of each major U.S. airline. Twitter data was scraped from February of 2015 and contributors were asked to first classify positive, negative, and neutral tweets, followed by categorizing negative reasons (such as "late flight" or "rude service").

Preprocessing

Preprocessing is very decisive in the process of determining sentiment, the classification model that is built will be more accurate. Preprocessing is also used to our dataset clean [8]. The preprocessing phase consists of several processes that will be discussed one by one in detail, including *Cleansing data*, Tweets contain many slang words and punctuation marks. We need to clean our tweets before they can be used for training the machine learning model. Cleansing data did reduce noise in the tweet data. Unimportant words will be removed such as URL, hashtag (#), username (@username), email, emoticons (: @,: *,: D), (,), dot (.) and also other punctuation [9].

Case folding, this stage serves to change letters character in the comments into all lowercase letters characters. In social media, especially Twitter, writing tweets, there must be differences in the shape of letters, case-folding stages is a changing process the shape to lowercase letters (lower case) or can also be called uniformity of letters. For example, folding case, input the sentence: "Disappointed with CS services", output the sentence: "disappointed with cs services".

Tokenizing, tokenizing or parsing stage is the cutting stage of the input string based on each word arrange [10]. In principle, this process is to separate every word that composes a document. In general, each word is identified or separated by another word by a space character, so the tokenizing process relies on the space character in the document to do word separations [5].

Stemming is the stage to make the word affixes into basic words. In stemming, conversion of morphological forms of a word to its stem is done assuming each one is semantically related. The stem need not be an existing word in the dictionary but all its variants should map to this form after the stemming has been completed. There are two points to be considered while using a stemmer [11]:

- Morphological forms of a word are assumed to have the same base meaning and hence should be mapped to the same stem.
- Words that do not have the same meaning should be kept separate.

These two rules are good enough as long as the resultant stems are useful for our text mining or language processing applications.

TF-IDF

As defined, TF is the term frequency in a single document. Terms can be words, phrases. For documents, the frequency for each term may vary greatly. Therefore, frequency is an important attribute of the term to discriminate itself from other terms. Sometimes, term frequency is directly used as the value of TF. That is, the TF value of term *i* is

$$TF_i = tf_{ik}$$

where tf_i denotes the frequency of term *i* in document *j*. Since the number of term frequency may be very large, the following formula is also often used to calculate TF value.

$$TF_i = log_2\ (tf_{ij}).$$

As for IDF, various formulas have been proposed. A basic formula was given by Robertson [12]. A later discussion between Spärck Jones[13] and Robertson resulted in the following formula of IDF:

$$IDF_i = log_2\left(\frac{N}{n_j}\right) + 1 = log_2(N) - log_2(n_j) + 1$$

where *N* is the total number of documents in the collection and n_j is the number of documents that contain at least one occurrence of the term *i*.

Random forest algorithm

Ensemble classification methods are learning algorithms that construct a set of classifiers instead of one classifier, and then classify new data points by taking a vote of their predictions. The most commonly used ensemble classifiers are Bagging, Boosting and Random Forest (RF) [14].

Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a

type of learning where you join different types of algorithms or the same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithms of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

RF classifier can be described as the collection of tree-structured classifiers. It is an advanced version of Bagging such that randomness is added to it [15]. Instead of splitting each node using the best split among all variables, RF splits each node using the best among a subset of predictors randomly chosen at that node.

A new training data set is created from the original data set with replacement. Then, a tree is grown using random feature selection. Grown trees are not pruned [15], [16]. This strategy makes RF unexcelled accuracy [17]. RF is also very fast, it is robust against overfitting, and it is possible to form as many trees as the user wants [15], [18].

The random forests algorithm (for both classification and regression) is as follows [19]:

1. Draw n_{tree} bootstrap samples from the original data

2. For each of the bootstrap samples, grow an *unpruned* classification or regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample m_{try} of the predictors and choose the best split from among those variables. (Bagging can be thought of as the special case of random forests obtained when $m_{try} = p$, the number of predictors.)

3. Predict new data by aggregating the predictions of the n_{tree} trees (i.e., majority votes for classification, the average for regression).

III. Result and Discussion

Before carrying out a series of analysis processes on the dataset, a little exploration was done on the dataset used in this study, to see how the distribution structure of the dataset used.

From the results of the description, the total amount of existing tweet data amounted to 14,640 with a total of 15 attributes. The data was divided into 6 airlines, each of which had been polarity labeled positive, negative and neutral sentiments. The following are the results of the description of tweet data based on each airline.

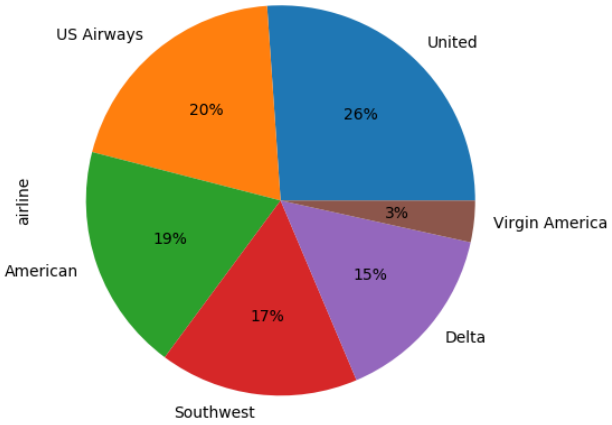

Fig. 1. Percentage of Public Tweet for Airlines

In the output, we can see the percentage of public tweets for each airline. United Airlines has the highest number of tweets i.e. 26%, followed by US Airways (20%), American (19%). For the next description, let's now see the distribution of sentiments across all the tweets.

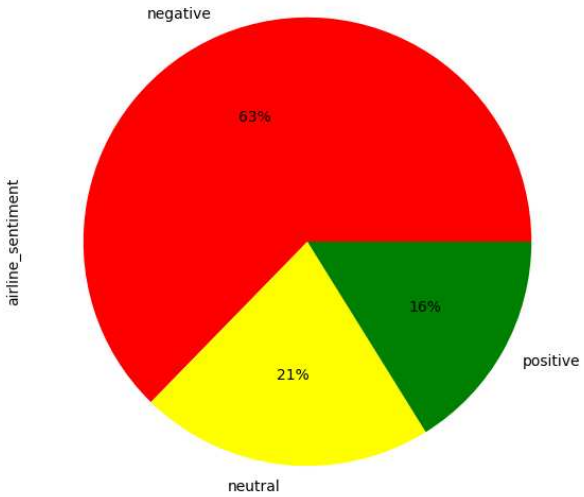

Fig. 2. Distribution of sentiments

From the output, we can see that the majority of the tweets are negative (63%), followed by neutral tweets (21%), and then the positive tweets (16%). To complete the data description, let's see the distribution of sentiment for each individual airline

It is evident from the output that for almost all the airlines, the majority of the tweets are negative, followed by neutral and positive tweets. Virgin America is probably the only airline where the ratio of the three sentiments is somewhat similar.

To see more detail about the dataset that we use in this paper, here following summary dataset table.

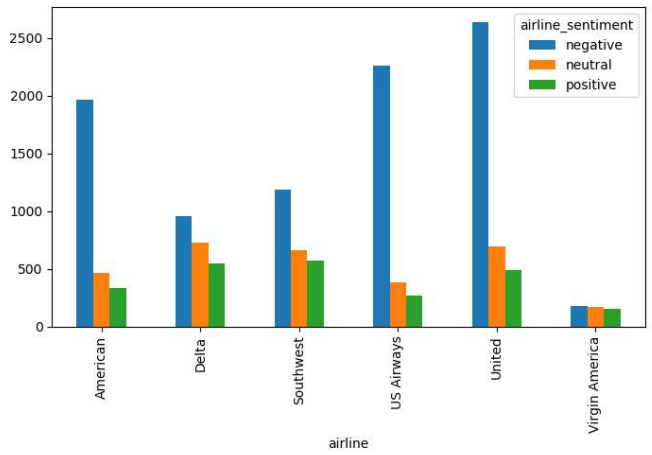

Fig. 3. Distribution of sentiment for each individual airline

TABLE 1.
Summary of datasets

Dataset (Airline)	***Number of Tweets***	***Percentage of***		
		Positive	*Negative*	*Neutral*
Virgin America	504	30%	35%	33%
US Airways	2913	13%	77%	10%
United	2434	12%	68%	18%
Southwest	4841	23%	48%	27%
Delta	2222	24%	43%	32%
American	2760	12%	71%	16%

Setelah kita melihat lebih dalam hasil deskripsi data dataset yang kita gunakan, langkah selanjutnya yaitu melakukan proses cleaning data, lalu kemudian melakukan training model dan terakhir melakukan prediksi dan evaluasi terhadap model.

Perhitungan evaluasi *metrics* klasifikasi yang kita gunakan adalah *confution metrix, F1 measure and accuracy.*

Berikut hasil evaluasi kinerja mesin learning yang kita bangun.

TABEL 2.
Confution matrix

		Aktual		
		Negatif	Neutral	Positif
Prediksi	Negatif	1723	108	39
	Neutral	326	248	40
	Positif	132	58	254

TABEL 3.
Precission, recall and f1-score

	Precision	Recall	F1-score
Negative	0.79	0.92	0.85
Neutral	0.60	0.40	0.48
Positif	0.76	0.57	0.65

From the output, the algorithm achieved an accuracy of around 75.99%. For information, testing is done with Python 3.6 programming language tools with several libraries, mainly sci-kit-learn. library commonly used for sentiment analysis.

IV. CONCLUSION

Sentiment analysis or opinion mining is a field of study that analyzes people's sentiments, attitudes, or emotions towards certain entities. This paper tackles a fundamental problem of sentiment analysis, sentiment polarity categorization. Tweets about six airline data from kaggle.com are selected as data used for this study. We performed sentiment analysis using the random forest algorithm and achieved an accuracy of around 75%. I would recommend you try and use some other machine learning algorithms such as logistic regression, SVM, or KNN and see if you can get better results.

V. ACKNOWLEDGMENT

My gratitude to the institution where *i* work, which has provided the opportunity to always do research. to my research friends in the BBPSDMP Kominfo Makassar Homebase, and other friends who have contributed to this research that *i* did not have time to mention one by one.

VI. REFERENCES

[1] C. J. Hutto and E. E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14).","* *Proc. 8th Int. Conf. Weblogs Soc. Media, ICWSM 2014*, 2014.

[2] N. Bahrawi, "Online Realtime Sentiment Analysis Tweets by Utilizing Streaming API Features From Twitter," *J. Penelit. Pos dan Inform.*, vol. 9, no. 1, pp. 53–62, 2019.

[3] Y. Wan and Q. Gao, "An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis," 2015.

[4] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, "Sentiment Analysis of Review Datasets using Naïve Bayes' and K-NN Classifier."

[5] F. Nurhuda, S. Widya Sihwi, and A. Doewes, "Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier," *J. Teknol. Inf. ITSmart*, vol. 2, no. 2, p. 35, 2016.

[6] A. Hamzah, "Sentiment Analysis Untuk Memanfaatkan Saran Kuesioner Dalam Evaluasi Pembelajaran Dengan Menggunakan Naive Bayes Classifier (NBC," 2014.

[7] D. Setyawan and E. Winarko, "Analisis Opini Terhadap Fitur Smartphone Pada Ulasan Website Berbahasa Indonesia," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 10, no. 2, pp. 183–194, 2016.

[8] I. Zulfa and E. Winarko, "Sentimen Analisis Tweet Berbahasa Indonesia Dengan Deep Belief Network," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 11, no. 2, p. 187, 2017.

[9] D. P. Artanti, A. Syukur, A. Prihandono, and D. R. I. M. Setiadi, "Analisa Sentimen Untuk Penilaian

Pelayanan Situs Belanja Online Menggunakan Algoritma Naïve Bayes," pp. 8–9, 2018.
[10] R. Feldman and J. Sanger, "The Text Mining Handbook," 2006.
[11] M. Anjali and G. Jivani, "A Comparative Study of Stemming Algorithms."
[12] R. Stephen, "Understanding inverse document frequency: on theoretical arguments for IDF," *J. Doc.*, vol. 60, no. 5, pp. 503–520, Jan. 2004.
[13] S. J. Karen, "IDF term weighting and IR research lessons," *J. Doc.*, vol. 60, no. 5, pp. 521–523, Jan. 2004.
[14] Ö. Akar, O. Gungor, and O. Güngör, "Classification of Multispectral Images Using Random Forest Algorithm View project 3D mapping View project Classification of multispectral images using Random Forest algorithm," vol. 1, no. □, pp. 105–112, 2012.
[15] L. Breiman, "RANDOM FORESTS," 2001.
[16] K. Archer and R. Kimes, "Empirical characterization of random forest variable importance measures," *Comput. Stat. Data Anal.*, vol. 52, pp. 2249–2260, 2008.
[17] L. Breiman and A. Cutler, "INTERFACE WORKSHOP-APRIL 2004 RFtools-for Predicting and Understanding Data."
[18] L. B. and A. Cutler, "Random forests - copyright." [Online]. Available: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_papers.htm. [Accessed: 26-Nov-2019].
[19] A. Liaw and M. Wiener, "Classification and Regression by RandomForest," 2002.