ISSN 2355-0082

The Right Sentiment Analysis Method of Indonesian Tourism in Social Media Twitter

Case Study: The City of Bali

Cristian Steven[1], Wella[2]
[1,2] Information System Department, Faculty of Engineering and Informatics, Universitas Multimedia Nusantara,
Tangerang, Indonesia
[2] wella@umn.ac.id

Accepted on 10 August 2020
Approved on 31 October 2020

***Abstract*—The growth of social media is changing the way humans communicate with each other, many people use social media such as Twitter to express opinions, experiences and other things that concern them, where things like this are often referred to as sentiments. The concept of social media is now the focus of business people to find out people's sentiments about a product or place that will become a business. Sentiment Analysis or often also called opinion mining is a computational study of people's opinions, appraisal, and emotions through entities, events and attributes owned. Sentiment analysis itself has recently become a popular topic for research because sentiment analysis can be applied in many industrial sectors, one of which is the tourism industry in Indonesia. To be able to do a sentiment analysis requires mastery of several techniques such as techniques for doing text mining, machine learning and natural language processing (NLP) to be able to process large and unstructured data coming from social media. Some methods that are often used include Naive Bayes, Neural Networks, K-Nearest Neighbor, Support Vector Machines, and Decision Tree. Because of this, this research will compare these four algorithms so that an algorithm can be used to analyze people's sentiments towards the city of Bali.**

***Index Terms*—decision tree, k-nearest neighbor, naive bayes, neural networks, sentiment analysis, social media, support vector machines, twitter**

I. Introduction

At this time social media is certainly not something foreign to talk, social media has changed human lives and ways human interact because social media is currently used by someone as a source of information and entertainment media [1]. The growth of social media itself change the way humans communicate with each other, many people use social media to express opinions, experience or other things that concern them, where things are like this is often referred to as sentiment [2]. The new concept of social media now became a main agenda of business people, be it decision makers or consultants businesses, all trying to identify whether the company can get various benefits [3]. One branch of research which then developed from the information explosion situation on the internet is sentiment analysis [4]. By doing a sentiment analysis, public opinion can be known about a product or service offered or to do a research [5].

Sentiment Analysis or often also called opinion mining is a study computational from people's opinions, appraisal, and emotions through entities, event and attributes owned [4]. Sentiment analysis itself has become a popular topic to be made because sentiment analysis can be applied in many industrial sectors, include the tourism industry [6]. Tourism sector in Indonesia itself always experience improvements every year and this tourism sector now ranked 42nd in the world [7]. One purpose famous tourism both in Indonesia and the world is the city of Bali, no Undoubtedly Bali City has a rich and diverse cultural heritage and the beauty of natural panoramas [8]. technology has become one important factor in improving the tourism industry sector in Indonesia, this is because in this digital age almost all people are connected to it social media as long as they are traveling [7]. In some last year, twitter became one of many social media that used by a lot of people, Twitter is one of the media social and a service microblogging which allows its users to send a message real time [2]. With twitter, public sentiment or opinion on tourism in a city can be known because inside tweet someone often conceives important information from an event that is very valuable to use as a tool to find out public opinion about these attractions, in addition that by using twitter can be seen the event or topic of the discussion currently popular with regard to tourism in the city by using hashtags [7].

To be able to do a sentiment analysis mastery of several techniques such as techniques to perform are required text mining, machine learning and natural language processing (NLP) to be able to process large and unstructured data that comes from social media [9]. Several algorithms that are often used to do sentiment analysis’s are Support Vector Machine, Naïve Bayes, K-Nearest Neighbors, and Decision tree. Research conducted by Mardiana make a comparison of methods for carrying out a classification among them are algorithms Support Vector Machine, Naïve Bayes, K-Nearest Neighbors, and Decision tree mention that the highest accuracy is obtained by using the SVM algorithm [10]. Other studies conducted by Romadloni conducted an algorithmic comparison Naïve Bayes, KNN, and Decision Tree mention that the highest accuracy is obtained with using an algorithm Decision Tree [11]. While other studies were conducted by Aulianita who compares algorithms SVM and KNN mention that the highest accuracy is obtained by using an algorithm KNN [12].

From the discussion above it was found that social media has been used by more than half of the world's population and social media alone have many the benefits. One of the benefits of social media is that it can do sentiment analysis or public opinion on something because of social the media is currently used by the public to express opinions or experience that concerns them. There are several algorithms that is usually used in sentiment analysis, i.e. the algorithm Naive Bayes, Support Vector Machine, K-Nearest Neighbors, and Decision Tree. Surely every method or algorithm used for analyzing sentiments will require different methods and producing results different. Therefore, this study will compare the four algorithms in conducting tourism sentiment analysis in Indonesia, especially the city of Bali, using social media twitter as data source.

II. Basic Theory

A. *Sentiment Analysis*

Sentiment analysis is a technique or method used for identify how a sentiment is expressed using text and how these sentiments can be categorized as positive sentiments negative sentiment [2]. Sentiment analysis includes the detection, analysis and evaluation of states of mind people to various events, problems, services or other interests [13]. The purpose of sentiment analysis itself is to find opinions, identify the sentiments they express, and then classifies its polarity [14]. Sentiment analysis itself can be divided into 2 parts, opinion mining relating to expressions and opinions and emotional mining related to with one's emotions in pronunciation or articulation [13].

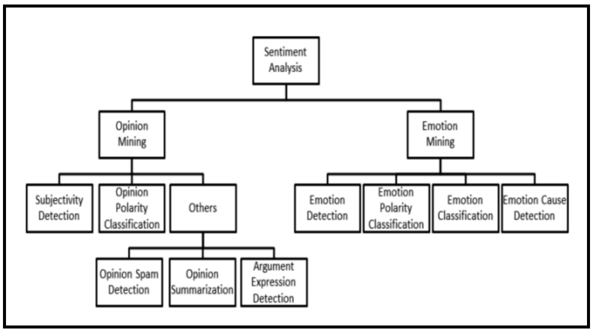


Fig. 1. Taxonomy of sentiment analysis

Opinion Mining more towards the concept of opinion expressed in the text which can be categorized into positive, negative or neutral expressions, while emotion mining more toward someone's emotions (happy, sad, angry) which is poured into a text [13].

B. *Support Vector Machine*

Support vector machine is a set of guided learning methods (supervised learning) which analyzes data and recognizes patterns, is used for classification and regression analysis [2]. Support Vector Machine is one of the best methods that can be used in the problem of classification, the concept of SVM stems from the problem of classification two class so that requires training set positive and negative [15]. The concept of classification is done by maximizing boundaries hyperplane that separating a data set or class, ability Support Vector Machine in finding hyperplane make this algorithm has a high level of generality and makes it the algorithm with the best level of accuracy compared to the others algorithm [1].

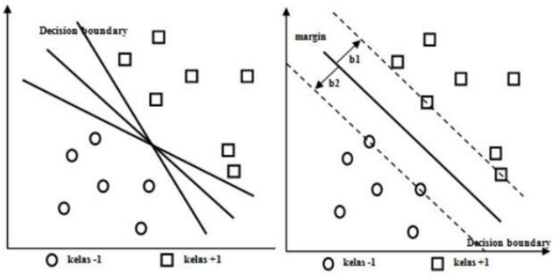


Fig. 2. Illustration of support vector machine

The picture above explains the concept of SVM classification, shown in figure (a) some data with circles as class -1 and squares as class +1, at the picture also exists hyperplane which is possible for data sets [16]. Figure (b) is hyperplane the maximum, calculation hyperplane done by calculating the distance of the margin to the nearest data from each class, the closest data is called Support Vector Machine [16].

C. *K- Nearest Neighbors*

Algorithm K-Nearest Neighbors (k-NN) is one of the most popular algorithms in machine learning, this is because the process is easy and simple, other than that k-NN also one of the algorithms supervised learning with process learning based on the value of the target variable associated with the variable value predictor [17]. Simple principle adopted by the algorithm NN is "if an animal is walking like a duck, then the animal may be duck", the closer the test data location is, it can be said that the training data these are seen more closely by the test data [12]. In the algorithm k-NN all data owned must have a label, so that when there is new data provided is then compared with existing data and the most similar data is taken and see the label of the data [17]. Learning data is projected into a multi-dimensional space, where each dimension represents a feature of the data, this space is divided into sections based on the classification of learning data, the best k values for this algorithm depends on the data, in general, a high k value will be reduce the effects of noise on classification, but create boundaries between each the classification becomes more blurred [11].

D. Decision Tree

Decision tree is a classification method of representation of a tree or decision tree, where an attribute is represented as node, and as the value is the branch of the tree, and the class is presented as a class [18]. In this case node the very top of a decision tree named as root, as in the picture [18].

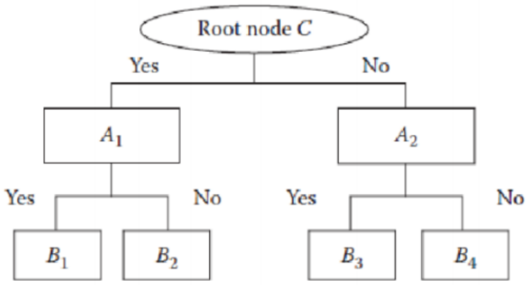


Fig. 3. Overview of the decision tree structure

One decision tree algorithm is C4.5, this algorithm is used for produces a decision tree and is commonly used for classification, therefore this is C4.5 often referred to as a statistical classifier [19].

E. Naïve Bayes

Algorithm Naïve Bayes classifier is an algorithm used for look for the highest probability value to classify test data in categories the most appropriate, this algorithm is one method machine learning which uses probability calculations [11]. This algorithm works with how to classify classes based on simple probabilities where in this case it is assumed that each attribute that exists is separate from each other, as for The formula equation of the method is as follows (1) and (2) [18].

$$P = (Y_k | x_1, x_2, \dots x_a)$$
$$P = (Y_k | x_a) = \frac{P(Y_k) + P(X_a | Y_k)}{P(X_a)} \qquad (1)$$

Where:

- $Y_k P(X_a | Y_k)$ = class category
- $P(Y_k)$ = class probability
- $P(X_a)$ = probability of document appearing

Based on the results obtained, then the class selection process is carried out optimal so that the greatest opportunity value of each probability is chosen existing classes [18].

F. Confusion Matrix

Confusion matrix performs tests to estimate objects that are true and false, the test sequence is tabulated in confusion matrix Where the predicted class is displayed at the top and the class observed next left [22]. Confusion Matrix illustrated by a table that states the amount of test data is correctly classified and the amount of test data that is incorrect classified [23].

TABLE I. CONFUSION MATRIX

Correct Classification	*Classified as*	
	Predicted "+"	**Predicted "-"**
Actuall "+"	True Positives	False Negative
Actual "-"	False Positives	True Negatives

Based on the table above:

- True Positives are the number of positive data records classified as a positive value.
- False Positives are the number of negative data records classified as a positive value.
- False Negatives are positive data records classified as positive value.
- True Negatives are negative data records classified as negative value.

G. ROC Curve

ROC curve widely used in data mining research in assessing results prediction, technically ROC Curve divided into two dimensions, where level True Positive put on the Y axis and level False Positive put on the axis X [22].). Chart Receiver Operating characteristic (ROC) is techniques for describing, organizing and choosing classifiers based on their performance, this curve is used to measure value Area Under Curve (AUC) [24]. he curve ROC show accuracy and comparing classification visually with presenting confusion matrix, whereas AUC is calculated to measure the difference in method performance used [25]. The guidelines for classifying accuracy testing using the AUC value are as follows:

a) 0.90 - 1.00 = Excellent Classification

b) 0.80 - 0.90 = Good Classification

c) 0.70 - 0.80 = Fair Classification

d) 0.60 - 0.70 = Poor Classification

e) 0.50 - 0.60 = Failure

III. RESEARCH METHODOLOGY

The research method used in this study was adapted from Kusumawati with a number of adjustments [26]. This research consists of data crawling, data-pre-processing, data labeling, data sharing, sentiment classification, and result & validation. The flow of research stages can be seen in Figure below.

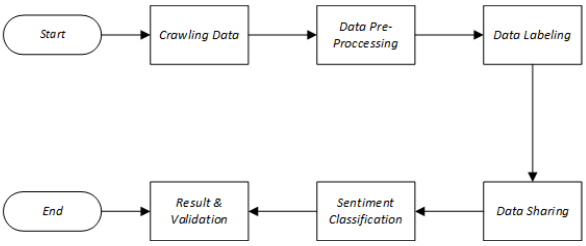


Fig. 4. Flow of work

A. *Crawling Data*

Tweets is collected with use API which has been provided by twitter, this API allows users to retrieve data from the application twitter a maximum of 7 days before. To be able to retrieve data, a tweet is used by the R Studio application. The R version used is version 3.6.2 and library used for data retrieval is rtweet library. In this research, tweet taken is which contain keyword "Bali" to get an opinion or experience Twitter users regarding tourism in the city of Bali. Tweets which are already collected is written in English and stored in CSV format.

B. *Data Pre-Proccessing*

At this stage and still using R Studio as tools. This stage will cleanse tweet collected and make a few changes.

- Change all words to lowercase, all words are changed to lowercase letters. This is done because library which exists in the R application it will read word for word in lowercase letters.
- Remove punctuation, all punctuation in tweet will be deleted.
- Delete the username, if in a tweet call or mention user Otherwise, the username will be deleted.
- Remove hashtag, delete all the "#" symbols in the data tweet that has been collected.
- Remove URL, tweet taken from twitter contain many urls, whether its http or https or link picture so it needs to be removed to simplify the labeling process.
- Removing RT or Retweet, all RT words or retweets will delete because it was judged to have no meaning.
- Remove emoticons, remove emoticons or existing emoji in the tweet.
- Clear data repetition, tweet taken identified has some data repetition that needs to be deleted for simplify the sentiment analysis process.

C. *Data Labeling*

After making a data crawling and data pre-processing, then the next step is giving label to every tweet you get. Data given label is data that contains user opinions on tourism in the city of Bali. Data tweets that contain advertising and not relevant to tourism in the city of bali will be removed. Data labeling used to provide positive or negative labels for each tweet which exists. In this study, the manual labeling process will be carried out by interviewees. Selected interviewees are people who have good understanding of English (English Literature Graduates). Interviewees can give positive, negative or neutral labels to tweets relating to tourism in the city of Bali. Label given by resource persons based on their interpretation of the contents of the tweet.

D. *Data Sharing*

At this stage the data that has been given a label will be shared into 2 parts, namely data training and data testing. Algorithm which will be used is an algorithm Supervised Learning, that is learning methods with training and trainers. In this approach, to find the decision function, separator function or regression function, training data is needed as an example of data that has output or label during the training process [27]. In this research data training and data testing will use the 70:30 rule.

E. *Sentiment Classification*

In this research the sentiment classification algorithm will used is an algorithm Naive Bayes, Support Vector Machine, K-Nearest Neighbors, and Decision Tree by using tourism in Bali city as a research object. This is done to find out the right algorithm is used to do sentiment analysis community towards the city of Bali. The model will be trained with data that has already been collected and determined sentiments. Next is the testing data whose sentiment has not been determined will be used to show whether the model created can represent data that has been be trained.

F. *Result & Validation*

After classifying sentiments and the results have been obtained prediction of sentiment analysis using the fourth algorithm, that is Naive Bayes, Support Vector Machine, K-Nearest Neighbors, and Decision Tree then the next result will be validated to find out how much accuracy each produced each algorithm in classifying tourism sentiments in the city Bali. Score Area Under Curve obtained will be the basis comparison of these four algorithms to determine for the right algorithm is used in classifying sentiments in Indonesia's tourism industry.

IV. Result & Discussion

A. Crawling Data

Crawling data carried out in stages, starting from 1 February 2020 until March 21, 2020. Data collection is done using the R Application Studio with the help of the API provided by twitter. Data during the pandemic of COVID-19 was not used because it could give different results for the sentiment analysis.

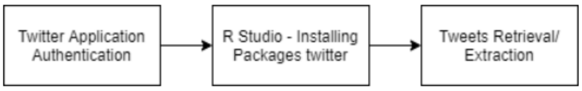

Twitter Application Authentication
R Studio - Installing Packages twitter
Tweets Retrieval/ Extraction

Fig. 5. Data crawling stages

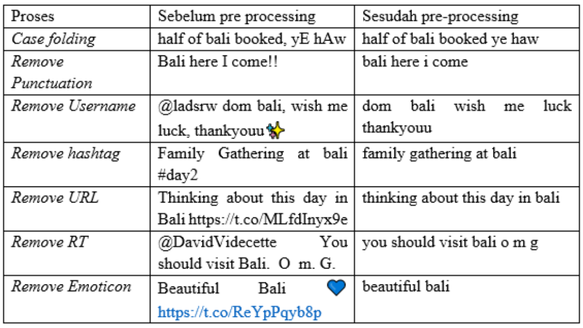
The stages of data collection are divided into 3 stages, namely:

1) Twitter Application Authentication, Twitter requires all data requests to use O Auth for user authentication. Application developers are asked to register the application on the site https: //apps/twitter.com for can access features data crawling which can pull tweet data from Twitter social media.

2) Installing R Package, there are several packages that are used to be able to do interaction with Twitter API and do Twitter data crawling. Some of the packages include packages 'rtweet', 'twitteR', 'streamR' and 'RtwitterAPI'. In this study packages were used 'rtweet' which is used to connect by including tokens obtained inside the script and use functions data crawling by including keywords which is desired.

3) Extraction of tweets, Tweets collected in this study are tweets contain the keyword "Bali". Total tweets that have been collected with these keywords reaching 16,000 tweets accompanied by several attribute or variable attached on the tweet.

B. Data Pre-Proccessing



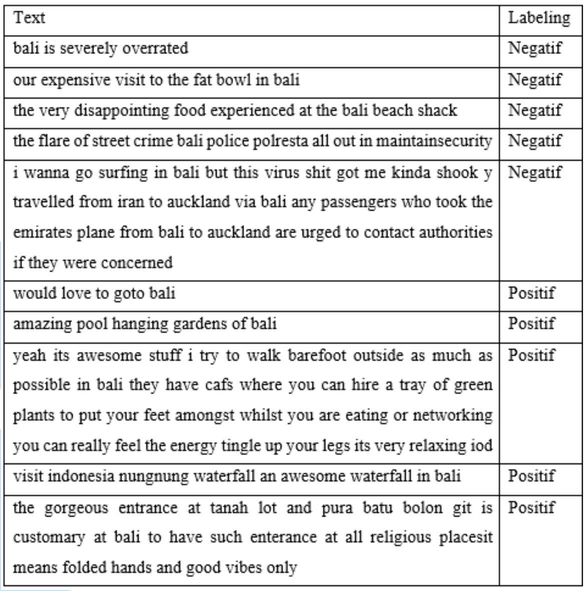
Proses	Sebelum pre processing	Sesudah pre-processing
Case folding	half of bali booked, yE hAw	half of bali booked ye haw
Remove Punctuation	Bali here I come!!	bali here i come
Remove Username	@ladsrw dom bali, wish me luck, thankyouu✨	dom bali wish me luck thankyouu
Remove hashtag	Family Gathering at bali #day2	family gathering at bali
Remove URL	Thinking about this day in Bali https://t.co/MLfdInyx9e	thinking about this day in bali
Remove RT	@DavidVidecette You should visit Bali. O m. G.	you should visit bali o m g
Remove Emoticon	Beautiful Bali 💙 https://t.co/ReYpPqyb8p	beautiful bali

Fig. 6. Result of the pre-proccessing stages

Figure 6 shows the results of the data pre-processing process performed before conducting the sentiment analysis process. This is done to make it easier the analysis process by balancing the format of the data to be analyzed.

C. Data Labeling

Overall, data that has been collected and has gone through the stages of data pre-processing totaling 4000 tweet. Data labeling is done manually with two criteria, namely positive and negative. The following results have been given data label.



Text	Labeling
bali is severely overrated	Negatif
our expensive visit to the fat bowl in bali	Negatif
the very disappointing food experienced at the bali beach shack	Negatif
the flare of street crime bali police polresta all out in maintainsecurity	Negatif
i wanna go surfing in bali but this virus shit got me kinda shook y travelled from iran to auckland via bali any passengers who took the emirates plane from bali to auckland are urged to contact authorities if they were concerned	Negatif
would love to goto bali	Positif
amazing pool hanging gardens of bali	Positif
yeah its awesome stuff i try to walk barefoot outside as much as possible in bali they have cafs where you can hire a tray of green plants to put your feet amongst whilst you are eating or networking you can really feel the energy tingle up your legs its very relaxing iod	Positif
visit indonesia nungnung waterfall an awesome waterfall in bali	Positif
the gorgeous entrance at tanah lot and pura batu bolon git is customary at bali to have such enterance at all religious placesit means folded hands and good vibes only	Positif

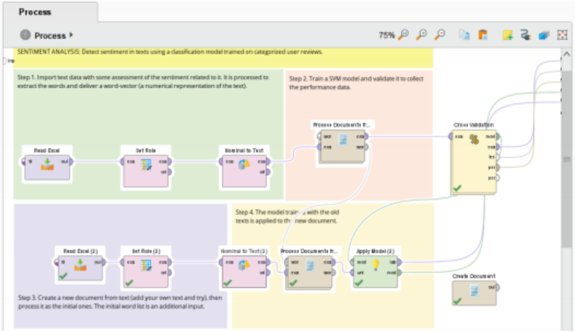
Fig. 7. Examples of tweet data labeling

D. Data Sharing

The data that has been collected will be divided into two parts, namely data training and also testing data. Training data is used to train machines that are made to find out how a tweet can be categorized as a tweet which is positive or negative. While testing data is used for test whether the classification has a high level of accuracy or low. Accuracy results will certainly be influenced by training data has been used to train the system that was made. the data divided into 2 parts with the proportion of training and testing data which is 70:30. The proportion then produces 2,273 tweets which will be made training dataset training dataset training dataset training dataset training dataset training dataset training dataset and 973 tweets will be testing dataset.

E. Sentiment Classification

Rapidminer this time will be used as a tool to classify public sentiment towards the city of bali. Rapidminer has a variety of operators that can be used for various purposes of analysis. Here is a big picture of the process sentiment classification is done using the operators of the application Rapidminer.



SENTIMENT ANALYSIS: Detect sentiment in texts using a classification model trained on categorized user reviews.

Step 1. Import text data with some assessment of the sentiment related to it. It is processed to extract the words and deliver a word-vector (a numerical representation of the text).

Step 2. Train a SVM model and validate it to collect the performance data.

Step 3. Create a new document from text (add your own text and try), then process it as the initial ones. The initial word list is an additional input.

Step 4. The model trained with the old texts is applied to the new document.

Fig. 8. Sentiment analysis process

In the initial stages the excel operator is used to enter data tweet which have been manually labeled to determine the sentiment of each twet, whether positive or negative. Data entered into rapidminer is data that has been carried out with pre-processing stages using the R Studio application. In making the sentiment classification process, it needs to be determined attribute which one will be classified. This can be done using operator set role on the rapidminer application. This operator works for determine the function of an attribute to be processed. Other than set role operation, the nominal to text operator is used too for changes everything nominal attribute becomes a string attribute. Each nominal value is only used as the string value of the new attribute. If the value is missing in the nominal attribute, the value new will also disappear.

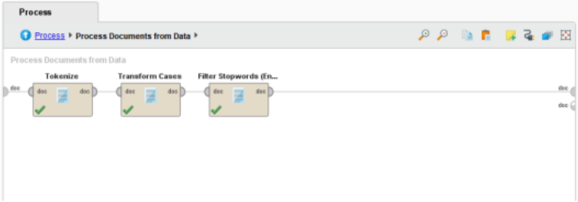


Fig. 9. Document process

In Figure 9 shows the operators used in processing documents or text into a vector with the method Term Frequency-Inverse document frequency. This method combines two concepts for weight calculation, i.e. the frequency of occurrence of a word inside a specific document and inverse the frequency of documents containing the words. There are several sub operators created in the process operator document to assist the process sentiment classification, such as the tokenize operator used to separate Data string becomes a word and stop words operators used to eliminate general or irrelevant words like the, for, of, and, so, forth, so that a collection of texts that has meaning and is related is produced with sentiment classification.

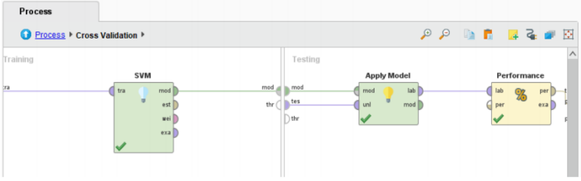
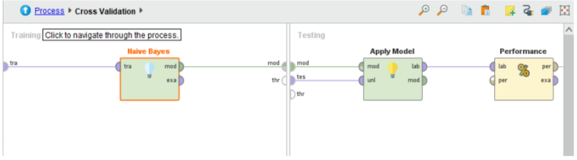


Fig. 10. Modeling with the SVM method

Cross validation is a resampling procedure used to evaluate machine learning models on data samples limited. This procedure has a single parameter called k which refers on the number of groups where the data sample will be shared. Because of this, procedure this is often called k-fold cross-validation. total fold used in the model made is 10 fold. In this case there are several operators available in the operator Cross Validation. SVM operator or Support Vector Machine used because of the algorithm this is one of the algorithms used in making models sentiment classification. This algorithm is one of the best methods you can used in classification problems. The key idea of SVM is to find the surface of the decision (Hyperlane) the maximum from each point data, to carry out machine training supported by vectors or commonly called Support Vector Machine. The SVM operator in the Rapidminer application supports various types of kernels including point, radial, polynomial, neural, anova, epachnenikov, a combination of gaussian and multiquadric. Explanation of this type of kernel is given in the parameters section.

Furthermore, the algorithm is used to do the classification sentiment is an algorithm Naive Bayes, This algorithm makes use of the method probabilities and statistics that predict future probabilities based on experience or data in the past.



Click to navigate through the process.

Fig. 11. Modeling with the Naive Bayes method

Figure 11 shows the use of the Naïve Bayes algorithm for do sentiment classification. Based on the documentation carried out by Rapidminer, this algorithm can build a good model even with small data sets, and are easy to use and don't require resource which is great for doing computing.

The third algorithm that will be used to classify sentiments the community against the city of Bali is an algorithm Decision Tree. Method Decision tree is a classification method of the representation of a tree decision, where an attribute is represented as a node, and as its value is the branch of the tree.

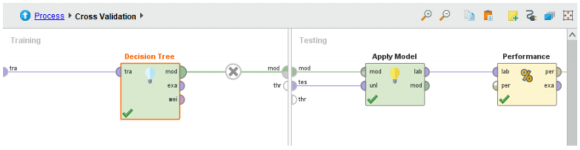


Fig. 12. Modeling with the Decision Tree method

Figure 12 shows the process of making a model using method decision tree. Operator decision tree in the Rapidminer application is a decision tree that contains collections node meant for make decisions about the value of affiliation to a class or estimated value numeric target. Every node in the decision tree represents the rules separation for one attribute certain, in the case of classification these rules are used to separate the values that have class different.

The final algorithm that will be used to do the classification people's sentiment towards the city of Bali is an algorithm K-Nearest Neighbors (KNN). This algorithm classifies based on similarity of data with other data.

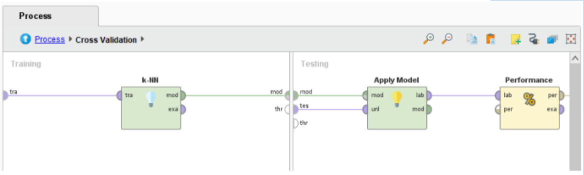


Fig. 13. Modeling with the k-NN algorithm

In Figure 13 shows the stages of the model making process with use operator k-nearest neighbors y which is in the Rapidminer application. This operator works based on comparing examples that do not known (data testing) with training examples k (data training) which is nearest neighbors from unknown examples.

There are several performance operators that exist in the Rapidminer application, such as performance (Classification), performance (Binominal Classification), Performance (Regression), or general performance that can be used for all types of learning assignments. In this study performance used is binominal classification performance. Performance binominal classification makes predictions where the result has two values, namely positive and negative. Additionally, the predictions for each example might be correct or wrong, leads to confusion matrix 2x2 with 4 classifications, namely trues positive, false positive, true negative, and false negative.

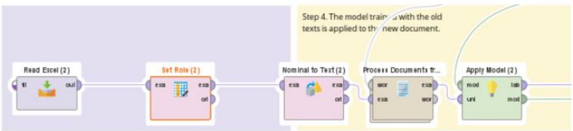


Fig. 14. Testing process

Figure 14 shows the repetition of the process used to do testing of the model that has been made.

F. Result & Validation

Based on the data collection process that has been carried out, obtained frequency of words that appear most often in searches with the keyword "Bali" using the R Studio application.

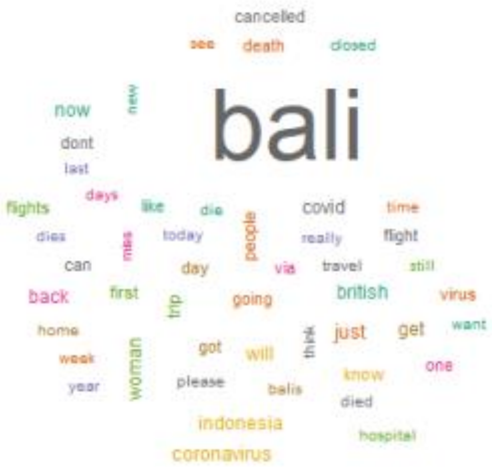


Fig. 15. Positive worldcloud

In Figure 15 is the frequency of words that most often appear from the data given is labeled "negative".

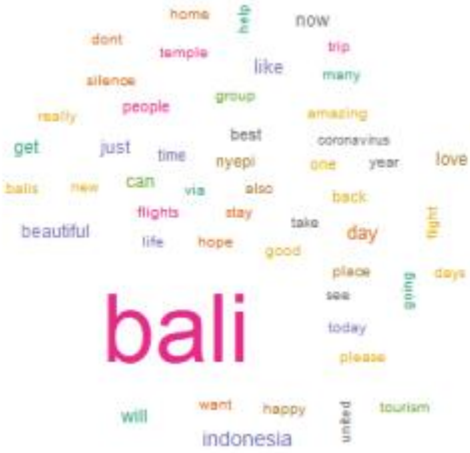


Fig. 16. Negative worldcloud

Figure 16 shows the frequency of words that appear most often from the data that is labeled "positive". Can be seen words like beautiful, love, like, happy and others are positive words.

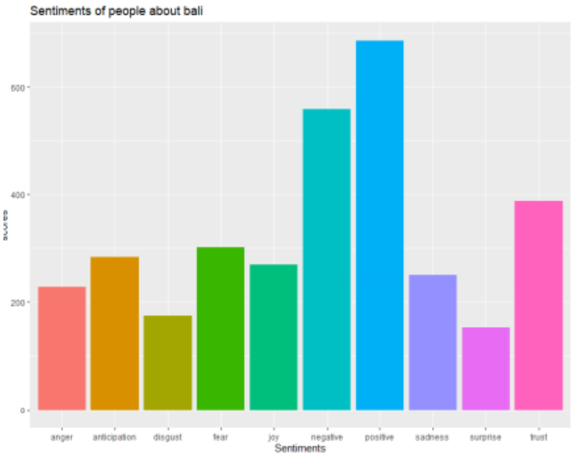
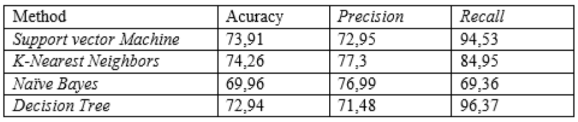


Fig. 17. Sentiments about Bali

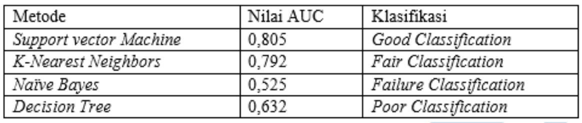
Figure 17 is a visualization of emotions or sentiments from the data people's sentiments about the city of Bali. Visualization is made using ggplot2 package in the R.



Method	Acuracy	*Precision*	*Recall*
Support vector Machine	73,91	72,95	94,53
K-Nearest Neighbors	74,26	77,3	84,95
Naïve Bayes	69,96	76,99	69,36
Decision Tree	72,94	71,48	96,37

Fig. 18. Comparison tables for accuracy, precision and recall

Figure 18 is a comparison table of accuracy values, precision, and recall generated from each method used. Highest accuracy value obtained by using an algorithm k-Nearest Neighbors with value accuracy of 74.26%. This value has a difference of 0.35% with the algorithm SVM which obtained a value of 73.91%.



Metode	Nilai AUC	Klasifikasi
Support vector Machine	0,805	*Good Classification*
K-Nearest Neighbors	0,792	*Fair Classification*
Naïve Bayes	0,525	*Failure Classification*
Decision Tree	0,632	*Poor Classification*

Fig. 19. Comparison of AUC values

Figure 19 shows the comparison of values AUC each method gets used. Algorithm Support Vector Machine be the most algorithm appropriate to be used in analyzing public sentiment towards the city of Bali with value AUC of 0.805 and included in Goof Classification. Next is the algorithm K-Nearest Neighbors get an AUC value of 0.792 and go inside fair classification. Although the resulting accuracy value the algorithm K-NN slightly larger than the SVM algorithm but overall the algorithm K-NN get a smaller AUC value than the algorithm SVM. Apart from that algorithm Naïve Bayes who get an AUC value of 0.525 and Decision Tree get an AUC value of 0.632. Both algorithms are entered into classification Failure classification and Poor Classification. This matter indicates the two algorithms are not appropriate to be used in doing sentiment analysis with tourism objects in the city of Bali.

V. Conclusion

This research succeeded in analyzing public sentiment towards the city of Bali based on data available on social media Twitter. Model data training conducted using the algorithm Naive Bayes, Neural Network, K-Nearest Neighbor, Support Vector Machines, and Decision Tree has a level of accuracy, precision recall that different. Based on the AUC value generated, the algorithm Support Vector Machine is the most appropriate algorithm used in analyzing tourism sentiments in Indonesia, especially in Bali AUC value of 0.805 and entered into Good Classification. Based on the values generated, the SVM algorithm is appropriate in analyzing sentiments in the tourism industry and through the process labeling that has been done, it is also known that people's sentiments towards the city of Bali which is more positive than sentiments that are negative.

For further research, it could measure the most frequently visited tourist attractions in the city of Bali based on the data that has been collected. This is done to obtain information on the ranking of tourist attractions in Bali that are currently popular with tourists. In addition, it is hoped that further research can automate tourism sentiment in Indonesia which is implemented into an application, which aims to provide information to the public about tourist attractions in Indonesia that have positive sentiments and are worth visiting.

References

[1] M. R. A. Nasution and M. Hayaty, "Perbandingan Akurasi dan Waktu Proses Algoritma K-NN dan SVM dalam Analisis Sentimen Twitter," *J. Inform.*, vol. 6, no. 2, pp. 226–235, 2019, doi: 10.31311/ji.v6i2.5129.

[2] N. Muchammad Shiddieqy Hadna, P. Insap Santosa, and W. Wahyu Winarno, "Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen Di Twitter," *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2016, no. Sentika, pp. 2089–9815, 2016.

[3] A. Pourkhani, K. Abdipour, B. Baher, and M. Moslehpour, "The impact of social media in business growth and performance: A scientometrics analysis," *Int. J. Data Netw. Sci.*, vol. 5, no. 1, pp. 223–244, 2019, doi: 10.5267/j.ijdns.2019.2.003.

[4] M. Murnawan, "Pemanfaatan Analisis Sentimen Untuk Pemeringkatan Popularitas Tujuan Wisata," *J. Penelit. Pos dan Inform.*, vol. 7, no. 2, p. 109, 2017, doi: 10.17933/jppi.2017.070203.

[5] J. Murphy *et al.*, "Social media in public opinion research," *Public Opin. Q.*, vol. 78, no. 4, pp. 788–794, 2014, doi: 10.1093/poq/nfu053.

[6] M. Ciric, A. Stanimirovic, N. Petrovic, and L. Stoimenov, "Comparison of different algorithms for sentiment classification," *2013 11th Int. Conf. Telecommun. Mod. Satell. Cable Broadcast. Serv. TELSIKS 2013*, vol. 2, pp. 567–570, 2013, doi: 10.1109/TELSKS.2013.6704442.

[7] D. T. Hermanto, M. Ziaurrahman, M. A. Bianto, and A. Setyanto, "Twitter Social Media Sentiment Analysis in Tourist Destinations Using Algorithms Naive Bayes Classifier," *J. Phys. Conf. Ser.*, vol. 1140, no. 1, pp. 0–8, 2018, doi: 10.1088/1742-6596/1140/1/012037.

[8] I. W. Budiasa, "Konsep Dan Potensi Pengembangan Agrowisata Di Bali," *dwijenAGRO*, vol. 2, no. 1, 2011.

[9] A. L. F. Alves, C. De S. Baptista, A. A. Firmino, M. G. De Oliveira, and A. C. De Paiva, “A comparison of SVM versus naive-bayes techniques for sentiment analysis in tweets: A case study with the 2013 FIFA confederations cup,” *WebMedia 2014 - Proc. 20th Brazilian Symp. Multimed. Web*, pp. 123–130, 2014, doi: 10.1145/2664551.2664561.

[10] T. Mardiana, H. Syahreva, and T. Tuslaela, “Komparasi Metode Klasifikasi Pada Analisis Sentimen Usaha Waralaba Berdasarkan Data Twitter,” *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 267–274, 2019, doi: 10.33480/pilar.v15i2.752.

[11] N. T. Romadloni, I. Santoso, and S. Budilaksono, “Perbandingan Metode Naive Bayes , Knn Dan Decision Tree Terhadap Analisis Sentimen Transportasi Krl,” *J. IKRA-ITH Inform.*, vol. 3, no. 2, pp. 1–9, 2019.

[12] R. Aulianita, “Komparasi Metode K-Nearest Neighbors dan Support Vector Machine Pada Sentiment Analysis Review Kamera,” *J. Speed – Sentra Penelit. Eng. dan Edukasi*, vol. 8, no. 3, pp. 71–77, 2016.

[13] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, “Current state of text sentiment analysis from opinion to emotion mining,” *ACM Comput. Surv.*, vol. 50, no. 2, 2017, doi: 10.1145/3057270.

[14] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014, doi: 10.1016/j.asej.2014.04.011.

[15] A. Pratama, R. C. Wihandika, and D. E. Ratnawati, “Implementasi Algoritme Support Vector Machine (SVM) untuk Prediksi Ketepatan Waktu Kelulusan Mahasiswa,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. March, pp. 1704–1708, 2018.

[16] N. Neneng, K. Adi, and R. Isnanto, “Support Vector Machine Untuk Klasifikasi Citra Jenis Daging Berdasarkan Tekstur Menggunakan Ekstraksi Ciri Gray Level Co-Occurrence Matrices (GLCM),” *J. Sist. Inf. Bisnis*, vol. 6, no. 1, p. 1, 2016, doi: 10.21456/vol6iss1pp1-10.

[17] F. Tempola, M. Muhammad, and A. Khairan, “Perbandingan Klasifikasi Antara Knn Dan Naive Bayes Pada Penentuan Status Gunung Berapi Dengan K-Fold Cross Validation Comparison of Classification Between Knn and Naive Bayes At the Determination of the,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, pp. 577–584, 2018, doi: 10.25126/jtiik20185983.

[18] O. Somantri and D. Dairoh, “Analisis Sentimen Penilaian Tempat Tujuan Wisata Kota Tegal Berbasis Text Mining,” *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 2, p. 191, 2019, doi: 10.26418/jp.v5i2.32661.

[19] N. S. B. Kusrorong, D. R. Sina, N. D. Rumlaklak, J. I. Komputer, and U. N. Cendana, “Kajian Machine Learning Dengan Komparasi Klasifikasi Prediksi Dataset Tenaga Kerja Non-Aktif,” vol. 7, no. 1, pp. 37–49, 2019.

[20] N. I. Widiastuti, E. Rainarli, and K. E. Dewi, “Peringkasan dan Support Vector Machine pada Klasifikasi Dokumen,” *J. Infotel*, vol. 9, no. 4, p. 416, 2017, doi: 10.20895/infotel.v9i4.312.

[21] M. Nurjannah and I. Fitri Astuti, “PENERAPAN ALGORITMA TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF) UNTUK TEXT MINING Mahasiswa S1 Program Studi Ilmu Komputer FMIPA Universitas Mulawarman Dosen Program Studi Ilmu Komputer FMIPA Universitas Mulawarman,” *J. Inform. Mulawarman*, vol. 8, no. 3, pp. 110–113, 2013.

[22] I. Menarianti, “Klasifikasi data mining dalam menentukan pemberian kredit bagi nasabah koperasi,” *J. Ilm. Teknosains*, vol. 1, no. 1, pp. 1–10, 2015.

[23] M. F. Rahman, D. Alamsah, M. I. Darmawidjadja, and I. Nurma, “Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN),” *J. Inform.*, vol. 11, no. 1, p. 36, 2017, doi: 10.26555/jifo.v11i1.a5452.

[24] E. Indrayuni, “Analisa Sentimen Review Hotel Menggunakan Algoritma Support Vector Machine Berbasis Particle Swarm Optimization,” *J. Evolusi Vol. 4 Nomor 2 - 2016*, vol. 4, no. 2, pp. 20–27, 2016.

[25] T. Rosandy, “PERBANDINGAN METODE NAIVE BAYES CLASSIFIER DENGAN METODE DECISION TREE (C4.5) UNTUK MENGANALISA KELANCARAN PEMBIAYAAN (Study Kasus : KSPPS / BMT AL-FADHILA,” *J. Teknol. Inf. Magister Darmajaya*, vol. 2, no. 01, pp. 52–62, 2016.

[26] R. Kusumawati, A. D’Arofah, and P. A. Pramana, “Comparison Performance of Naive Bayes Classifier and Support Vector Machine Algorithm for Twitter’s Classification of Tokopedia Services,” *J. Phys. Conf. Ser.*, vol. 1320, no. 1, pp. 0–10, 2019, doi: 10.1088/1742-6596/1320/1/012016.

[27] M. Ridwan, H. Suyono, and M. Sarosa, “Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier,” *Eeccis*, vol. 7, no. 1, pp. 59–64, 2013, doi: 10.1038/hdy.2009.180.