

SENTIMENT ANALYSIS OF ELECTRIC CARS USING RECURRENT NEURAL NETWORK METHOD IN INDONESIAN TWEETS

^a Felisia Handayani, ^b Metty Mustikasari

^{a,b} Department of Computer Science and Information Technology Gunadarma University,
Depok, 16424. Indonesia

E-mail: felisiasilalahi97@gmail.com, metty@staff.gunadarma.ac.id

Abstract

Sentiment analysis is computational research of the opinions of many people who are textually expressed against a particular topic. Twitter is the most popular communication tool among Internet users today to express their opinions. Deep Learning is a solution to allow computers to learn from experience and understand the world in terms of the hierarchy concept. Deep Learning objectives replace manual assignments with learning. The development of deep learning has a set of algorithms that focus on learning data representation. The recurrent Neural Network is one of the machine learning methods included in Deep learning because the data is processed through multi-players. RNN is also an algorithm that can recall the input with internal memory, therefore it is suitable for machine learning problems involving sequential data. The study aims to test models that have been created from tweets that are positive, negative, and neutral sentiment to determine the accuracy of the models. The models have been created using the Recurrent Neural Network when applied to tweet classifications to mark the individual classes of Indonesian-language tweet data sentiment. From the experiments conducted, results on the built system showed that the best test results in the tweet data with the RNN method using Confusion Matrix are with Precision 0.618, Recall 0.507 and Accuracy 0.722 on the data amounted to 3000 data and comparative data training and data testing of ratio data 80:20.

Key words: Confusion Matrix, Long Short Term Memory, Recurrent Neural Network, Twitter.

INTRODUCTION

Evolution in information and communication technology have an impact on lifestyle changes in society, internet sites are like an ocean of information resources for many people. The Government of Indonesia was responsive in dealing with demands for information transactions in cyberspace with the enactment of the Law of the Republic of Indonesia Number 11 of 2008 concerning Information and Electronic Transactions (UU ITE). UU ITE consists of several chapters in which to discuss all things related to information through electronic [1]. Today's society tends to be open with social media to write down the events they are going through rather than have a direct talk to their surroundings. According to information from the Ministry of Communication and Information, revealed in 2019 the internet users in Indonesia currently reach 63 million people [2].

With so many internet users, a system was built that could analyze public opinion on a topic called opinion mining. According to Liu, 2012, sentiment analysis or opinion mining is a computational study of the opinions of people, sentiment, and emotions through entities that are compressed in text form. Pang & Lee stated that sentiment analysis would group polarity from the text in sentences or documents to see the opinions expressed in sentences or documents and grouped by positive, negative, or neutral classes [3].

Research conducted by Pengfui (2016) regarding Text Classification with one of the methods in deep learning that answers the most suitable for text is RNN, the study uses three tasks to classify data. This research is done by comparing several models that have been trained and give the results that LSTM architecture can improve the performance of a group of tasks in processing information between layers. The Activation Function is used to flexibly control the flow of information between a shared layer and a specific layer of tasks, thereby obtaining a better sentence representation [4].

According to Research conducted by Brett Duncan and Yanqing Zhang regarding sentiment analysis with Neural Networks using Matlab tools that apply feedforward for pattern data recognition. This research creates a vocabulary list for each unique word

mapping and mapping for each variable. This research model, the Recurrent Neural Network states that with the standard RNN there are problems in terms of memory when iterating, this research also has a small Dataset of 200 tweets [5].

In another study conducted by Ye Yuan, You Zhou regarding research Conducted data sentiment analysis with three different hyperparameters. This research uses 6092 tweets of data. The research analyzed by trying out three activation functions including drop out, ReLu and Tanh and resulted in the drop out being the best activation in the research. The results of the sentiment analysis classification are strongly influenced by the preprocessing stage, the balance of the amount of data for each class and the model used. The research stated that the number of Datasets influenced the results of the analysis [6].

M. Amin and Azhari (2019) have conducted a sentiment analysis of Novel. The data retrieved from the site www.goodreads.com specifically for Indonesia's best Novel which is collected in CSV. Data review then preprocessed and labeled with three classes that are positive, negative and neutral. The research also processes data with the embedding layer, which is intended to study the mapping of each word with a data dictionary into vectors based on models created on Word2vec. According to this study, the use of the RNN method without LSTM can lead to problems in gradient loss in the data analysis process. A comparison of the number of records of each class precisely assessed can affect the score of accuracy, precision, recall, and F-measure gained [7].

Based on the explanation above, this study uses the Recurrent Neural Network method which is capable of storing memory (feedback loop). This method allows us to recognize data patterns well, then used to make accurate predictions. The RNN way to store information is to do looping inside its architecture, which automatically keeps information stored from the past. However, RNN has flaws in gradients that disappear with the increasing length of sequential data that retrained to the RNN. To address the problem, this study used one of RNN architecture, namely Long Short Term Memory. LSTM was created by Hochreiter & Schmidhuber (1997). On LSTM, the cell

contents become more complex than 1 layer of neurons and this is what makes LSTM learn long patterns of sequential data due to the vanishing gradient situation prevented [8]

This study was conducted to analyze public opinion about the usage of electric cars in Indonesia through its tweets. This research is done by retrieving data from Twitter communities that have a public account. The data is processed by the recurrent Neural Network method and Long Short Term Memory architecture to create a positive, negative and neutral sentiment model to measure the classification performance using the accuracy, precision, and recall of the analysis.

MATERIAL

Text Mining

Text Mining is an interesting process of discovering and extracting knowledge from a set of unstructured data. The data sources that are used include documents and from websites as well [9].

Deep Learning

Deep learning can be extracted as a solution to allow computers to learn from experience and understand the world in terms of concept hierarchy, with each concept defined by its relationship to a simpler concept. By collecting knowledge from experience, this approach avoids the need for the human operator to formally or manually determine all Computer-Required knowledge. The concept hierarchy allows the computer to learn a complex concept by building it from a simpler one [10].

Recurrent Neural Network

The recurrent Neural Network (RNN) is a type of neural network that its processing is performed repeatedly to process inputs that are typically sequential data. RNN belongs to the deep learning category because the data is processed through multiple layers. RNN is an algorithm that can recall inputs with internal memory, making it very suitable for machine learning problems involving sequential data. Sequential data has a character that is processed with a sequence of data (e.g. time) and a sample to have a close relationship with each other. In each processing, the resulting output is not only a function of the sample but

also based on the internal state which is the result of the processing sample [11].

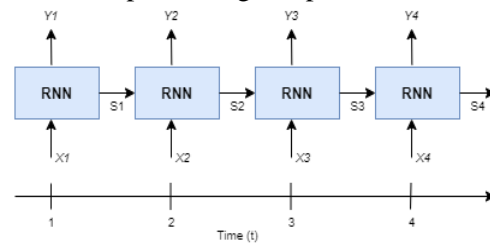


Fig 1. RNN Architecture

Alternative representations of RNN architecture can be seen in Figure 1, where there is only one RNN module. The word recurrent in RNN is because RNN performs the same calculation process repeatedly for the given input. For example, the process of a data x_1 at a certain time segment t can be separated by x_t . Data will be processed by neurons (blue box RNN) to be output values y_t , the results of neuron processing will be stored in the loop connection for use in subsequent data processing x_{t+1} . [10]

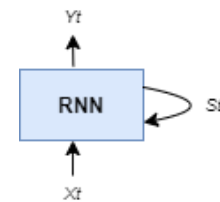


Fig 2. Alternative representations of RNN architecture

Long Short Term Memory

LSTM or Long-Short Term Memory is a type of RNN architecture that is expected to fix a simple RNN weakness. With LSTM, the model can predict a sentence that has a different context by re-entering the calculation results from the previous input on the hidden layer. Thus, the model can know what context is in the sentence and can prefigure out the output sentence to be removed [10].

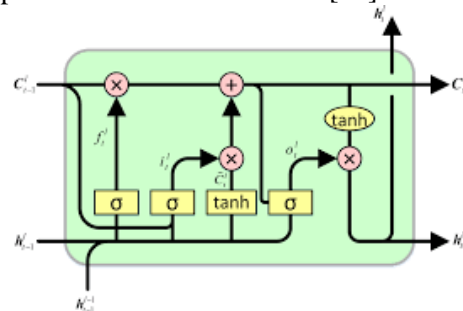


Fig 3. Long Short Term Memory Illustration

The other LSTM advantage is the ability to remove or add information received. That

way, the information used is only useful information for subsequent calculations [10].

To get the output of the sigmoid function on the LSTM is not much different from the simple JST. The difference is only adding new insert i.e. output result on previous time calculation and sigmoid function. [10]

$$f_t = \sigma (W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

To update the condition in the cell LSTM, it takes a new value to change the output result in the previous order state. That is by using the Tanh function. By multiplying the result of the sigmoid function and the result of the Tanh function, it will get a new value that is useful to update the value in the LSTM cell.

$$i_t = \sigma (W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$S_t = (W_s \cdot [h_{t-1}, x_t] + b_s) \quad (3)$$

To get the C_t update value, sum the results of the first sigmoid output with the value result in the previous order with the second sigmoid result multiplied by the result of the Tanh function.

$$C_t = f_t * C_{t-1} + i_t * S_t \quad (4)$$

The result of the LSTM cell calculation will determine whether the information should be forwarded to be output or not [10].

Sentiment Analysis

According to Liu, 2012, Sentiment analysis or opinion mining is one of the areas of Natural Language Processing. Text Mining aims to organize the opinions, sentiments, evaluations, judgments, attitudes and emotions of a person whether he or she is pleased with a particular topic, product, service, organization, individual, or activity [3].

The basic task in sentiment analysis is the grouping of text that exists in a sentence or document, then Specifying the opinion expressed in the sentence or the document whether it is positive, negative or neutral. Sentiment analysis can also illustrate the emotional feelings of sadness, joy, or anger [3].

METHODS

Data sources for this study are taken from a set of tweets taken with a predefined keyword that is "electric car" and comes from tweets that have a public account. The tweets were

retrieved from 5 December 2019 to 3 months later and incorporated into Microsoft Excel with the extension, CSV.

Preprocessing Data

In general, preprocessing data is done by deleting inappropriate data or converting data into simpler forms to be processed by the system. [12]

The preprocessing stages of the study consist of:

1. Cleansing Data

Data cleansing aims to reduce noise in the tweet's data. Words deemed unimportant are omitted such as numbers, URLs, hashtags (#), retweets, usernames (@username), emails, emoticons (:@, :*, :D) and other punctuation marks.

2. Case Folding

Case folding is a stage that changes all letters in a document into lowercase letters. Only letters a through z are accepted.

3. Tokenizing

The tokenizing or parsing is a stage of cutting the input string based on each word that makes it up. In tokenizing, space is used to separate the words.

4. Stopword Removal

Stopword removal is a process of removing words that have no contribution to the document's contents. The words included in the stopword are removed because it gives a bad influence on the process of text mining.

5. Stemming

The stemming process aims to change the words that have been added to the word following the rules of the correct Indonesian language or turn the word into its basic word only.

The language model of the dataset has been created with the Recurrent Neural Network and Long Short Term Memory methods. In this process also occurs word embedding through training data.

Modeling

Modeling of this study using the fastai framework. The Language model will create a model for studying the structure of the language through a hierarchy representation, thereby loading low-level features (Word

representation) and high-level features (semantic meanings).

This can be done because the language model is usually trained on a very large set of unsupervised data and hence this model can learn about the language syntax feature in a much deeper way than the word embeddings. This process can also see some tags applied to the words, which aim to store all the information that can be used to gather an understanding of the vocabulary of the new task. Fast.ai provides an easy-to-use utility (learn.lr_find) to search through different levels of learning to find optimal results for datasets. The learning rate will increase the learning level after each mini-batch. Finally, when the learning rate has a high value results in decreasing the value of the loss. Model training is done with code learning fit_one_cycle. [13]

Confusion Matrix

Confusion Matrix is a model commonly used to present a classification result. With the confusion matrix, the data can analyze how well the classifier can recognize records from different classes. The testing method is done by calculating the Accuracy, Precision and Recall analysis results [14].

Accuracy is testing based on the proximity level between the predicted value and the manually checked values. By knowing the amount of data that is correctly classified, then it can be known as the accuracy prediction result.

Table 1. Confusion Matrix 3 x 3

	Predict		
	Positive	Negative	Neutral
	Positive (TP)	(FPNeg)	(FPNeu)
Current	Negative (FNegP)	(TNeg)	(FNegNeu)
	Neutral (FNeuP)	(FNeuNeg)	(TNeu)

Table 1 is a form of 3 x 3 confusion matrix table which is commonly used in the problems of sentiment analysis [3].

Precision is a test conducted by comparing the amount of relevant information that the system acquired with the sum of all information taken by the system whether it's relevant or not.

Recall is a test with a comparison of the number of relevant information that the system acquired with the sum of all relevant information contained in the collection of

information, either taken or not taken by the system. [14].

Here are formula and description of the use of Accuracy, Precision and Recall [11]:

$$\text{Accuracy} = \frac{TP+TNeg+TNeu}{All} \quad (5)$$

$$\text{Precision} = \frac{TP}{TP+FPNeg+FPNeu} \quad (6)$$

$$\text{Recall} = \frac{TP}{TP+FNegP+FNuP} \quad (7)$$

RESULT AND DISCUSSION

The accuracy testing of tweets in this study discusses the testing of the tweet classification model that has been created with the Recurrent Neural Network method and the architecture of Long Short Term Memory. Tweet classification testing is done by measuring accuracy, precision, recall. The tests conducted with three scenarios datasets used for classification testing consisted of 1000, 2000, and 3000 data. The test was conducted with three variations of the validation test used scenarios data there are 70:30, 80:20 and 90:10.

The summary of the results will be presented in tables. The results table will be used for comparison of each value, therefore it can be analyzed which tests have the best results.

Table 2. Evaluation Result Of 1000 Data

	Precision	Recall	Accuracy
70 : 30	0.478	0.676	0.657.
80 : 20	0.355	0.450	0.592
90 : 10	0.360	0.138	0.641

In Table 2 from test 1000 data, the best test result data is found in the data ratio 70:30 with the Precision value 0.478, Recall 0.676 and Accuracy 0.657

Table 3. Evaluation Result Of 2000 Data

	Precision	Recall	Accuracy
70 : 30	0.559	0.590	0.686
80 : 20	0.502	0.624	0.649
90 : 10	0.477	0.687	0.643

In Table 3 from test 2000 data, the best test result data is in the data ratio 70:30 with the Precision value 0.559, Recall 0.590 and Accuracy 0.686.

In Table 4 from test 3000 the best test result data is found in the data ratio 80:20 with the Precision value 0.618, Recall 0.507 and Accuracy 0.722

Table 4. Evaluation Result Of 3000 Data

	Precision	Recall	Accuracy
70 : 30	0.508	0.668	0.69
80 : 20	0.618	0.507	0.722
90 : 10	0.610	0.521	0.72

CONCLUSION

This study has gained sentiment analysis from tweets with the keyword electric car from Twitter. The research uses Tweet Scraper, one of the Python libraries in the data retrieval process and the stages performed on preprocessing. The method used in creating models is the Recurrent Neural Network with its architecture of Long Short Term Memory. This research conducts the accuracy of the model by using three dataset-sharing

scenarios, there are 70:30, 80:20 and 90:10. Scenario data in the amount of are 1000, 2000, and 3000 data with each having three classes there are positive, negative and neutral. The final result of the test concluded that the best accuracy results occurred on datasets that had 3000 data and had a comparison of 70:30 for training data and data testing, which resulted in Precision 0.618, Recall 0.507 and Accuracy 0.722 which means that the more data used gives better results and the ratio of distribution data also influences the test results.

This sentiment analysis still requires further development, such as data in larger quantities, using different and better algorithms and preprocessing steps, therefore it can improve the performance of the analysis.

REFERENCES

- [1] Utami, Lilyani Asri, 2016, Analysis Of The Public Opinion Sentiment News Of Forest Fires Through Comparison Algorithm Support Vector Machine And K-Nearest Neighbor Based Particle Swarm Optimization. Jurnal Pilar Nusa Mandiri Volume 13 No.1.
- [2] Andri, Delvi. 2017. The Relationship Between The Level Of Interpersonal Trust And The Adolescent Self-Disclosure Of Online Social Media Users At The 2 Holy State High School. Published by AGP Research. ISBN 978-605-82949-3-6
- [3] Liu, Bing. 2012. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publisher. Doi.org/10.2200/S00416ED1V01Y201204HLT016.
- [4] Pengfei. Liu. Qiu, Xipeng. Huang, Xianjing, 2016, Recurrent Neural Network For Text Classification With Multi-Task Learning. Computation and Language
- [5] Brett Duncan And Yanqing Zhang, 2015, Neural Networks For Sentiment Analysis On Twitter. <https://Doi.Org/10.1016/J.Eswa.2017.07.01>
- [6] Ye Yuan, You Zhou, 2015, Twitter Sentiment Analysis With Recursive Neural Networks.
- [7] Nurrohmat, Muh Amin & Sn, Azhari, 2019, Sentiment Analysis of Novel Review Using Long Short Term Memory Method. Indonesian Journal of Computing and Cybernetics Systems. Vol 13, No 3 (2019)
- [8] Lowe, Ryan et al, 2017 Training End-To-End Dialogue Systems With The Ubuntu Dialogue Corpus. DOI: <https://doi.org/10.5087/dad.2017.102>
- [9] Kao, Anne. Poteet, Steve R, 2007. Natural Language Processing And Text Mining. Springer International Publishing Science & Business Media.
- [10] Goodfellow, Ian. Bengio, Yoshua. Courville, Aaron, 2016, Deep Learning, Mit Press.
- [11] Suyanto. Michael. Mandala, Satria. 2019, Deep Learning Modernization Machine Learning For Big Data. Informatics Publisher.
- [12] Hemanth, Jude. 2018. Computational Vision and Bio Inspired Computing Springer, Technology & Engineering.
- [13] Howard et al (2018). "Learner for the text Application" [Website]. Accessed on <https://docs.fast.ai/text.learner.html>, January 2020.
- [14] Ashish Kumar, Luhach, Dharm Singh, Pao-Ann Hsiung, Kamarul Bin Ghazali Hawari, Pawan Lingras, Pradeep Kumar Singh, 2018. Advanced Informatics For Computing Research: Second International Conference, Icaicr 2018, Shimla, India, July.