Sentiment Analysis*: A Comparison of Deep Learning Neural Network Algorithm with SVM and Naïve Bayes for Indonesian Text*

Wahyu Calvin F M[1,3]**, Siti Mariyah**[1,2,3]**, and Setia Pramana**[1,2,3]

[1]Sekolah Tinggi Ilmu Statistik (STIS), Jakarta, Indonesia
[2]Center of Computational Statistics Study, STIS, Jakarta, Indonesia
[3]Badan Pusat Statistik (BPS), Indonesia

calvinrevelation@gmail.com sitimariyah@stis.ac.id setia.pramana@stis.ac.id

Abstract. Deep learning is a new era of machine learning techniques that essentially imitate the structure and function of the human brain. It is a development of deeper Artificial Neural Network (ANN) that uses more than one hidden layer. Deep Learning Neural Network has a great ability on recognizing patterns from various data types such as picture, audio, text, and many more. In this paper, the authors tries to measure that algorithm's ability by applying it into the text classification. The classification task herein is done by considering the content of sentiment in a text which is also called as sentiment analysis. By using several combinations of text preprocessing and feature extraction techniques, we aim to compare the precise modelling results of Deep Learning Neural Network with the other two commonly used algorithms, the Naïve Bayes and Support Vector Machine (SVM). This algorithm comparison uses Indonesian text data with balanced and unbalanced sentiment composition. Based on the experimental simulation, Deep Learning Neural Network clearly outperforms the Naïve Bayes and SVM and offers a better F-1 Score while for the best feature extraction technique which improves that modelling result is Bigram.

Introduction

Text mining is an application of data mining technique that devoted to managing textual data [1]. It refers generally to the process of extracting interesting and non-trivial patterns or knowledge from unstructured data text [2,3]. Basically, text mining is a multidisciplinary field. It involving information retrieval, statistics, mathematics, machine learning, linguistic, and natural language processing. Textual data type is the most natural form on storing information. Based on recent study, it is indicated that 80% of company's information is contained in text documents [2]. The massive use of text-type data demands more intensive use of the knowledge that can be extracted from it. However, text mining is a much more complex task which dealing with unstructured data. One of the most popular domains on text mining is sentiment analysis. This analysis can classify documents based on sentiment contained therein. The sentiment is generally can be classified into two groups (positive and negative classes), or three groups (positive, negative, and neutral classes).

Technically, sentiment analysis can be divided into four types of approach [4]: *Machine learning approach*, it uses a learning algorithm to create a model from training data for supervised classification. (This approach is also used in this paper.). *Lexicon-based approach*, it involves calculation of the

polarity of the sentiment by using the semantic orientation of words or sentences in a document. *Rule-based approach*, it searches for opinion words and then classifies them based on the number of positive and negative words. This approach also considers several different rules such as dictionary polarity, negation words, booster words, idioms, emoticons, and mixed opinions. *Statistical model approach*, it makes any opinion as an aspect and a latent rating. This approach represented by using divisions and then clustered into rank form.

For the machine learning approach, sentiment analysis has some commonly used classification algorithms such as Naïve Bayes and Support Vector Machine (SVM). In terms of modelling paradigm, both algorithms are also usually classified as traditional algorithms. But recently, with the emergence of new method that utilize Deep Learning techniques, now sentiment analysis can perform a new era of text classification tasks. As demonstrated by the recent study, this algorithm has several advantages such as it can work on any text genres with minimal restrictions and with no task-specific or data-specific manual feature engineering [5]. In this study, we aim to measure the ability of that new algorithm (Deep Learning Neural Network) by comparing it with the traditional algorithms (Naïve Bayes and SVM) using several combination of preprocessing and feature extraction techniques for Indonesian textual data. That three algorithms tested with balanced and unbalanced sentiment composition documents.

Related Work

1.1. Deep Learning Neural Network (DLNN)

Deep learning is a new era of machine learning techniques which imitates the structure and function of the human brain on recognizing something. This algorithm has a unique feature that allows to automatically grasps the relevant features required for. Theoretically, the model used in deep learning is a mathematical function with $f: X \rightarrow Y$. Deep learning is the development of a more in-depth Artificial Neural Network (ANN) that uses more than one hidden layer on modelling the given dataset [6]. It consists of three main layers, namely:

1. *Input Layer* : the layer consists of neurons that receive data from variable X.
2. *Hidden Layer* : it consists of neurons that receive data from previous input layer. Each hidden layer is responsible for training the unique set of features. The more number of the hidden layers, the more increasing the complexity and the abstraction.
3. *Output Layer* : the layer consist of neurons that receive data from hidden layer that produces the output value and is the result of the calculation of the variable X to variable Y.

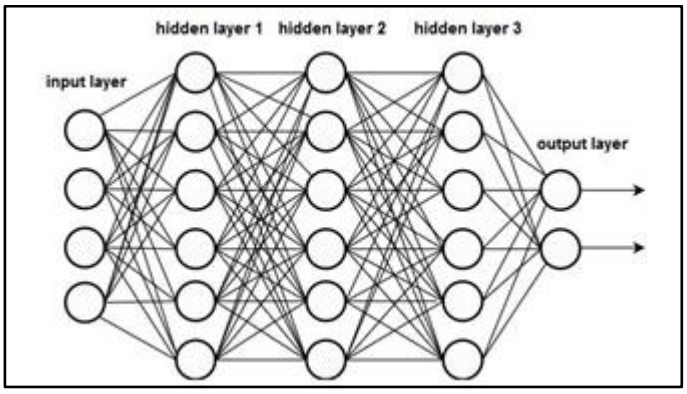

Fig. 1: Deep Learning Neural Network structure

1.2. Evaluation Measurement

For the modelling result comparison for each models, the authors used three types of evaluation measurement namely: Precision, Recall and F-measure.

Precision is the relevancy ratio of the result of the prediction answer. That ratio is a comparison between the prediction accuracy and the total number of predictable answers for a particular class (e.g. positive class). That precision can be formulated as follows:

$$Precision = \frac{true\ postives}{true\ positives + false\ positives} \quad (1)$$

Recall is the ratio of how many truly relevant result are returned. The ratio is a comparison between the accuracy of the predicted answer and the true value. The formula used is as follows:

$$Recall = \frac{true\ postives}{true\ positives + false\ negatives} \quad (2)$$

F-measure or so-called F-1 score is a measure of the accuracy of a predicted result by considering precision and recall values. The result value of this measure is the harmonic average between precision and recall which can be formulated as follows:

$$F1\ Score = 2 \times \frac{precision \times recall}{precision + recall} \quad (3)$$

Methodology

1.3. Dataset Collection Method

Basically, the dataset used in this research is a primary data which is sourced from Twitter (social media). That text data was gained by utilizing web crawling techniques with two search keywords that is the name of one of the government institution in Indonesia and one of the most prominent Indonesian government figures in 2017. Furthermore, the text data gained from the web crawling is still a raw data that has not been labelled, while to do supervised classification we need a labelled data. Therefore the next step process we did is labeling each data one by one manually. The labelled data is used as the base data (or also called corpus) as the training dataset in creating the model.

From the result of labelling process, the data composition used in this study can be detailed as follows:

Table 1. Dataset composition

Keyword Topic	Composition Category	Sentiment		
		Positive	**Negative**	**Neutral**
Government Institution	*Imbalanced*	46	30	1.070
	Balanced	46	30	-
Government Figure	*Balanced*	126	132	-

1.4. Text Mining Method

Text data is an unstructured data type. Therefore, the additional text mining steps are required in order to transforming that text data into a a form that can be processed computationally by the machine. This step also usually called text refining [2]. One of the major limitations on text mining is the natural language used. Each language has it own rules. Because of these limitations, the authors restrict the research is only processing text data that use Indonesian language.

To discover the useful patterns from the text data (sentiment pattern), we use the text mining techniques. That techniques used to create the required model which later used to predict any text with topic about it. In this study, the text mining technique is consist of few stages as follows:

1.4.1 Text preprocessing

Text preprocesing is an early stage of semantic analysis (meaning accuracy) and syntactic analysis (arrangement accuracy). The preprocessing for Indonesian language text steps in this study consists of: case folding, word tokenizing, stop word removal, and stemming [7]. At this stage we also carried out the text formalization and word spell improvement (typo).

- *Case Folding,* is the process of uniforming all the characters of the letters that exist in a text into lowercase. For example, the phrase "*Teknologi Informasi dan Komunikasi*" after the case folding will be "*teknologi informasi dan komunikasi*".

- *Word Tokenizing,* is the process of decomposing the whole text into a words form. For example, the phrase "*teknologi informasi dan komunikasi"* after tokenized will be "*teknologi*", "*informasi*", "*dan*", and "*komunikasi*".

- *Stop Word Removal,* is a process of omitting the meaningless words that often appear in a text such as "*yang*", "*di*", "*ke*", "*atau*", etc.

- *Stemming,* is the process of converting a word into its basic form or by other terms it is the process of omitting affixes from a word. This word conversion should be done to make sure that every words that is same but has different affixes can be recognized as a similar value to avoid bias in the transformation stage.

- *Text Formalization,* is the conversion of an informal word into a standard-formal word. Thus, the words that are supposed to be one thing or the same (semantically) can be equalized when the text transformation stage.

- *Word Spell Improvement (typo),* is the alteration of an incorrect word form (syntactically) which may also be due to the typo errors into the proper arrangement. For example, the word "*beruasha*" after the word spell improvement will be "*berusaha*".

1.4.2. Feature selection

Feature selection is the stage of selecting which feature extraction technique will be used later to transform the text. In this research, the authors use four techniques of feature exraction, among others: Bag of Word, Binary Term Frequency (TF), Term Frequency-Inverse Document Frequency (TF-IDF), and Bigram (N-Grams).

1.4.3. Feature extraction (transformation)

After selecting the feature at the previous stage, the text data then transformed into a matrix form (or also called feature) to make it able to be processed computationaly.

1.4.4. Pattern discovery

This is the main stage of the series of text mining process. In this stage, the important patterns from text data (training) discovered and formed into a model. As descibed earlier, the authors used three machine learning algorithms to be compared which include: Deep Learning Neural Network, Naïve Bayes, and Support Vector Machine (SVM) on modeling the Indonesian language dataset.

1.4.5. Models evaluation

After the pattern recognition stage, for evaluate the formed models, the authors uses the three measurement: Precision, Recall, and F-1 score.

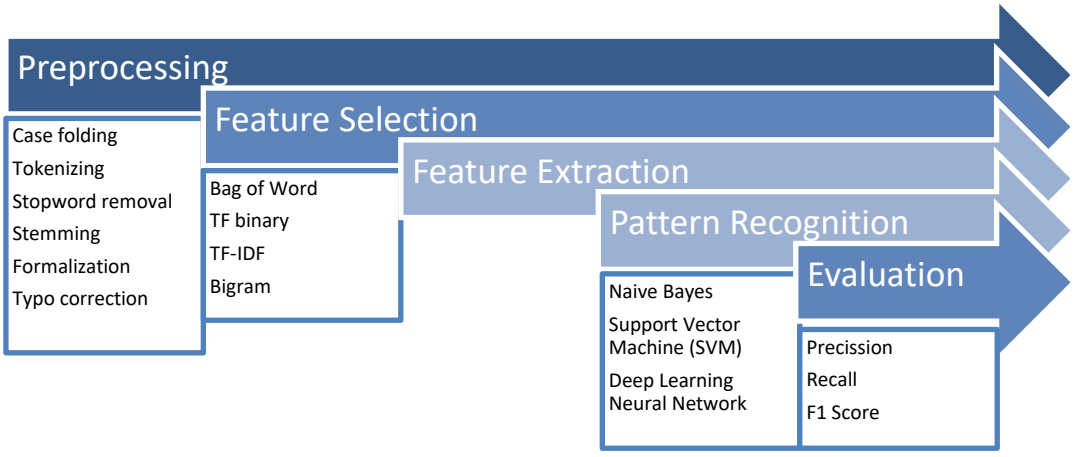

Fig. 2: Text mining process used for *Bahasa Indonesia*

Experiments

For the implementation of this research, the authors uses Python programming language and its library to build up and process these three algorithms. Furthermore, for the Deep Learning Neural Network, we use TensorFlow and Keras as the development tools in building the Deep Learning models. TensorFlow is an open source library used for rapid numerical computation. It is created and maintained by Google and released under the Apache 2.0 license [6]. TensorFlow is the foundation library in making Deep Learning models either directly build or by using a wrappers library such as Keras.

1.5. Deep Learning Classification Architecture

For the Deep Learning Neural Network classification architecture, the authors divide the modelling algorithm into two main parts of classification:

1.5.1 Three Classes of Sentiment (positive, negative, and neutral)

For this three classes classfication, the Keras model used is Sequential with the compile method is Categorical Crossentropy and layers as follows:

- 1 input layer with the number of input dim equal to the length of the vocabulary stored in the transformation stage.
- 2 hidden layer each consisting of 75 neurons and 25 neurons. This hidden layers uses normal init and relu activation.
- 1 output layer with 3 neurons that using normal init and sigmoid activation.

1.5.2 Two Classes of Sentiment (positive or negative)

For this two classes classfication, the Keras model used is Sequential with the compile method is Binary Crossentropy and layers as follows:

- 1 input layer with the number of input dim equal to the length of the vocabulary stored in the transformation stage
- 2 hidden layer each consisting of 60 neurons and 20 neurons. This hidden layers uses uniform init and relu activation.

- 1 output layer with 1 neuron that using uniform init and sigmoid activation that can generate output value in range 0 up to 1 where if the resulting value is less than or equal to 0.5 will be categorized as a negative class and vice versa.

1.6. Research Frame of Mind

To simplify the text mining research in this study, the authors described the process on a frame-of-mind form as follows:

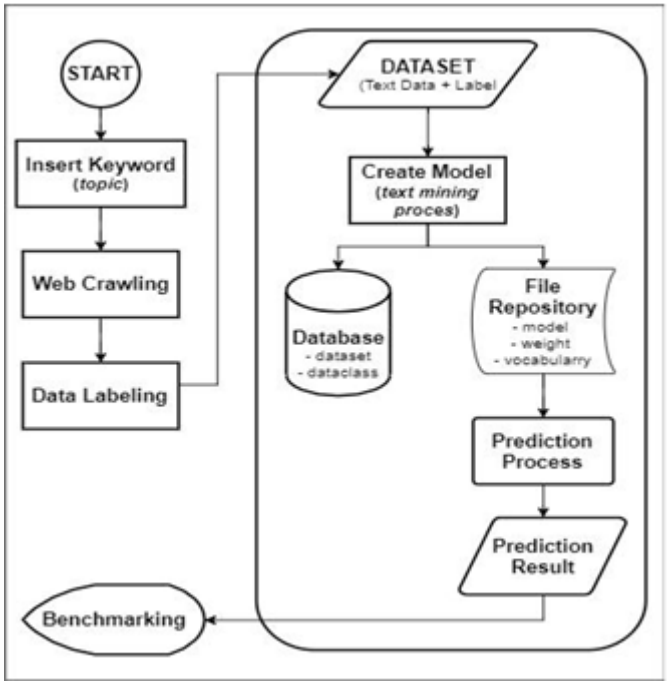

Fig. 3: Mind Frame

The first process is starting from determining the topic of the text data to be studied. As described earlier, we use two topics that is about one of the government institution in Indonesia and one of the most prominent Indonesian government figures in 2017. In applying the web crawling technique, the predetermined topic then used as the text search keyword (in this research is on Twitter). After getting the raw text data needed, the next process is labelling that data one by one manually for further supervised classification process. Now we have got the dataset (data & label) needed. Next step is the main process of text mining. As described earlier, that main process is consist of 5 sub-process, they are: Preprocessing, Feature Selection, Feature Extraction, Pattern Recognition, and Evaluation (Fig.2). The result of the text mining process is the models. After that, each models then evaluated by the three measurement: precission, recall, and F1 Score using the prediction dataset.

1.7. Experimental Result

Before comparing the three algorithm, each dataset has been splitted up into two group of data: 90% part of data for training set (create model) and 10% more for testing set (prediction). There are three set of data used. The first dataset is consist of three classes, they are positive, negative, and neutral. This first dataset is also categorized as imbalanced composition dataset because its neutral class has a very large size while compared to the other two classes. While the second and third class is consist of two classes, positive and negative. They are categorized as balanced composition dataset. Finally, we show the experimental result of our research on the following benchmark tables:

Algorithm	Feature Extraction Technique	Dataset: Government Institution (3 classes) with Imbalanced Composition			Dataset: Government Institution (2 classes) with Balanced Composition			Dataset: Government Figure (2 classes) with Balanced Composition		
		Precision	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Naïve Bayes	Bag of Word	0.85	0.88	0.86	0.75	0.75	0.75	0.85	0.85	0.85
	Binary TF	0.85	0.90	0.87	0.75	0.75	0.75	0.85	0.85	0.85
	TF-IDF	0.85	0.87	0.86	0.60	0.75	0.67	0.82	0.81	0.81
	Bigram (nGram)	0.85	0.76	0.81	0.40	1.00	0.57	0.76	0.69	0.72
Support Vector Machine (SVM)	Bag of Word	0.88	0.90	0.89	0.40	0.50	0.44	0.81	0.80	0.80
	Binary TF	0.88	0.92	0.90	0.40	0.50	0.44	0.78	0.76	0.77
	TF-IDF	0.88	0.91	0.87	0.43	0.67	0.52	0.77	0.77	0.77
	Bigram (nGram)	0.88	0.92	0.90	0.40	0.50	0.44	0.83	0.69	0.75
Deep Learning Neural Network	Bag of Word	0.99	0.84	0.91	0.94	0.93	0.93	**1.00**	**0.92**	**0.96**
	Binary TF	0.99	0.74	0.85	0.94	0.93	0.93	**1.00**	**0.92**	**0.96**
	TF-IDF	0.96	0.15	0.26	0.94	0.95	0.95	**1.00**	**0.92**	**0.96**
	Bigram (nGram)	**1.00**	**0.98**	**0.99**	**0.97**	**0.96**	**0.96**	1.00	0.90	0.95

Conclusion

The sentiment classification performance of the different combination of algorithms and feature extraction techniques gives quite diverse results. Based on the research above, on all of the dataset given, the performance of Deep Learning Neural Network algorithm is over-whelming. At least, one of the feature extraction technique combinations in Deep Learning algorithm for each dataset always gets the top score. In addition, unlike Naive Bayes and SVM, the modelling results produced by Deep Learning Neural Network are also not significantly affected by either balanced or unbalanced data composition conditions. The resulting score is always high. On average, for the combination of Deep Learning Neural Network algorithms, the best feature extraction technique that can improve the final score of the modelling is Bigram (N-Grams). Thus, based on this research can be concluded that the combination of Deep Learning Neural Network algorithm with bigram technique is very suitable for conducting sentiment analysis for Indonesian text data.

References

[1] Feldman, R. & Dagan, I. (1995) Knowledge discovery in textual databases (KDT). In proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada, August 20-21, AAAI Press, 112-117.

[2] Tan, A. H. (1999, April). Text Mining: The state of the art and the challenges. In Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advenced Databases (Vol. 8, pp. 65-70). sn.

[3] Han, J., Kamber, M., & Pei, J. (2006). Data preprocessing. Data mining: concepts and techniques. San Fransisco: Morgan Kaufmann, 47-97.

[4] Collomb, A., Costea, C., Joyeux, D., Hasan, O., & Brunie, L. (2014). A study and comparison of sentiment analysis methods for reputation evaluation. Rapport de recherche RR-LIRIS-2014-002.

[5] Singhal, P., & Bhattacharyya, P. Sentiment Analysis and Deep Learning: A Survey.

[6] Brownlee, Jason. (2017).Deep Learning With Python: Develop Deep Learning Models on Theano and TensorFlow using Keras.

[7] Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. (2010). Text mining: predictive methods for analyzing unstructured information. Springer Science & Business Media.

Acknowledgment

The authors also thanks to institute of statistics (STIS) and The Central Bureau of Statistics (Badan Pusat Statistik or BPS) for the opportunity given in carrying out this research.