Offensive Language Detection using Artificial Neural Network

Meredita Susanty
Computer Science
Universitas Pertamina
Jakarta, Indonesia
meredita.susanty@universitaspertamina.ac

Sahrul
Computer Science
Universitas Pertamina
Jakarta, Indonesia
hssahrul@null.net

Ahmad Fauzan Rahman
Computer Science
Universitas Pertamina
Jakarta, Indonesia
ahmadfauzan1198@gmail.com

Muhammad Dzaky Normansyah
Computer Science
Universitas Pertamina
Jakarta, Indonesia
dzakynormansyah@gmail.com

Ade Irawan
Computer Science
Universitas Pertamina
Jakarta, Indonesia
adeirawan@universitaspertamina.ac.id

***Abstract*— Governments and social media providers put an effort to tackle offensive, abusive, and profanity in social media as an abuse of speech freedom. Considering the number of Internet user in Indonesia and the conflict caused by offensive content about religion, race, and inter-group issues in Indonesia, there is an urge to develop offensive content detection for posts written in Bahasa. This paper uses an artificial neural network model for not only classifying the words as (non)offensive words but also considering the structure of the sentence to get its context. The challenges are informal grammar and word abbreviation used in social media. Hence, there are noise elimination and normalization processes to address these challenges. The computer simulation results show excellence accuracy of 99.18% training, 94.28% validation, and 96.8% testing, only by utilizing the sigmoid activation function. This model can assist government enforcing the information and electronic transaction law and decreases the number of disputes due to aspiration freedom abuse in social media.**

Keywords—offensive content detection, artificial intelligent, neural networks, machine learning, natural language processing

I. Introduction

According to the Indonesian Internet Service Providers Association's report (APJII), the participation of netizens in the country reached 143 million in 2016 [1]. In other words, 54.68% of Indonesia's population are internet users. Furthermore, the report shows that 87.13% of Indonesian internet users use it for accessing social media. Social media becomes an inseparable part of Indonesian digital lifestyle.

Unfortunately, the social freedoms, that the sites allow us to have, increase the number of offensive posts. The offensive post that might cause chaos are categorized into ethnicity, religion, race, and inter-group issues, cyberbullying and body shaming. Similar with other countries, Indonesia government seriously tackle the problem through the enactment of laws related to information and electronic transactions in Undang Undang No.19 of 2016. Align with the governments, social media provider put some efforts to tackle the issues [2]. However, the efforts made were only limited to the provision of reporting channels for offensive, abusive or profanity content and deals with it by simply filtering out a post when it is flagged as offensive.

Another way to tackle the issue is by exploiting the computer to classify a sentence to an offensive or non-offensive post, known as the sentiment analysis method. Natural Language Processing (NLP) enables a computer to process or understand human natural language very well [3]. NLP has two aims; (1) scientific perspective, modeling cognitive mechanisms to improve computer understanding of natural human language, (2) engineering perspective, developing an application to facilitate computer and human interaction [3]. Therefore, NLP can be utilized in sentiment analysis.

This research used the derivation of sentiment analysis named sentiment classification. Sentiment classification is used to classify data sets into a class. To do classification in the form of text, data is pre-processed before being used on the Artificial Neural Network (ANN) models. Pre-processing data in the form of text includes noise elimination, normalization, and segmentation [4].

ANN is a computational system inspired by the biological network of neurons that make up a human brain [5]. Technically, ANN is a computational learning system that uses a network of function to understand and translate a certain form input data into the desired output, usually in a different form [6]. The ANN is not an algorithm. Instead, it is a framework for many different machine learning algorithms to work together in processing complex input data [6]. ANN models have been widely applied to daily life, for example, to recognize sounds or images, classify documents, medical diagnoses, and continue to be applied to many other fields. In this research, we use ANN models for sentiment classification for a collection of text data.

Several social media such as Facebook and Instagram have an automatic word detection feature [7] [8]. Unfortunately, it only detects words without considering the context of the sentences which considered not effectively addressing the issues. There are several researches on offensive detection using Naïve Bayes, Support Vector Machine, Semantic Method [9] [10] and Obfuscation Methods [11]. However, these methods also do not consider the context of sentences or phrases. Another research that considers the context of the sentences detects the offensive content and give an alternative non-offensive version is using an unsupervised approach [12].

All previous work addressing the offensive language in a post or content written in English. Considering the number of Internet user in Indonesia and the conflict caused by offensive content about religion, race, and inter-group issues in Indonesia, this research aims to develop offensive content detection for post written in Bahasa. This research introduces a supervised approach to tackle offensive context in Bahasa by using artificial neural network model. The challenges in detecting offensive language in Bahasa are informal grammar and the form of the word which informally abbreviated.

978-1-5386-8448-1/19/$31.00 ©2019 IEEE

II. Methodology

This research follows activities as shown in Fig. 1 which consist of collecting data, pre-processing the data, labelling the data, input data into the model, then perform the simulation.

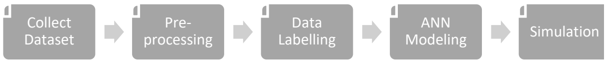


Fig. 1. Research Activities

A. Dataset

Web crawling using web scraping is used to obtain data from various sites on the Internet. Collected data are in the text format. Dataset consists of both positive and negative sentiments. Data are divided into three; training data, validation data and testing data.

B. Pre-processing

Before the data are inputted into the ANN model, we need to perform noise elimination, normalization, and segmentation. Noise elimination eliminates incompatible characters in the data. Once noise has been eliminated, the data are normalized. The goals of this step are correcting any typographical errors and reducing any words redundancy which might lead to ambiguity in the model. The last step in the pre-processing data is segmentation. Dataset can be a paragraph consist of more than one sentence or phrase. Sentences or phrases that build a paragraph is separated using a punctuation such as; period (.), question mark (?), exclamation mark(!), enter, and more than one period (…). Then, each data instance eithers, sentence or phrase, are divided into separate segments (words) in the form of a matrix.

C. Data Labeling

Data labeling is performed manually based on cultural custom where label 1 indicates that the text is offensive and 0 indicates non-offensive.

D. ANN Modeling

Fig. 2 show the ANN model to process the text input. The model consists of 2 layers: one input layer and one hidden layer. For large amounts of data, the ANN model uses more than one hidden layer. This model is known as deep neural networks or deep learning.

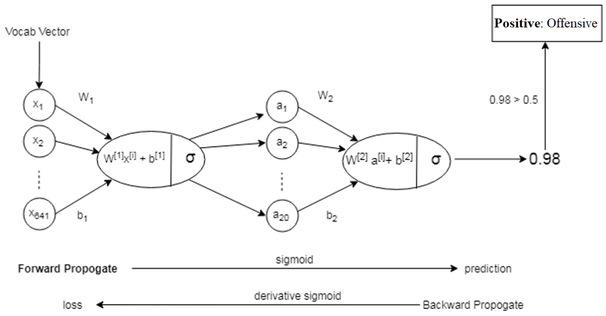


Fig. 2. Artificial Neural Network Model

The activation function in the forward-propagation process is the sigmoid function as shown in equation (2). Equation (1), where W defines the weight, x specifies the input word vector, and b represents the bias, is the input for the sigmoid function is the sum of weighted inputs of that particular neuron. We add a bias in the formula so that the ANN model does not lose its generality. Because we classify data into two classes; non-offensive (0) or offensive (1), we use the sigmoid function which is considerably good for binary classification.

$$z = Wx + b \tag{1}$$

$$\sigma(z) = \frac{1}{1+e^{-z}}\ , 0 \le \sigma(z) \le 1 \tag{2}$$

Derivative sigmoid function, shown in equation (3), is used for back-propagation. It works from the last layer to the input layer.

$$\frac{dy}{dx}\sigma(z) = \frac{1}{1+e^{-z}} \cdot \left(1 - \frac{1}{1+e^{-z}}\right) \tag{3}$$

The output from the model is calculated by equation (4) for the input layer or first layer and equation (5) for the hidden layer, where y defines the output, W specifies the weight, x specifies the input vector, a is a word vector resulted from the former layer, and b specifies the bias.

$$y = \sigma\,(Wx + b),\ 0 \ \le y \le 1 \tag{4}$$

$$y = \sigma\,(Wa + b),\ 0 \ \le y \le 1 \tag{5}$$

III. Experiments

This research uses 504 data; 307 as training data, 70 validation data, and 63 testing data. Data is gathered from Twitter using API Twitter and Tweepy library from Python. The output is file in jsonl extension. Jsonl file is converted into json then converted again into csv.

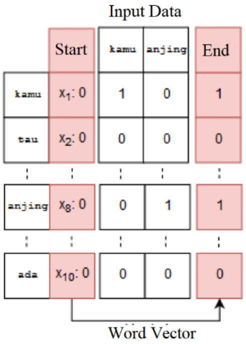
In pre-processing, first, all punctuations and tags were removed from the data. For example, eliminate punctuation and tags, "`<p> kamu tau babi gak? </p>`" was changed into "`kamu tau babi gak`". Next, in normalization, any typographical errors were corrected, e.g, "`liat ada njing`" to "`liat ada anjing`", and any words listed in a file were translated into its synonym, e.g, "`habis kesabaran gua anjing`" was changed into "`habis kesabaran saya anjing`". Lastly, segmentation, separating an instance into words by identifying a space between words. Example of instance segmentation as follows:

```
{"kamu tau babi gak", "habis kesabaran saya
    anjing", "liat ada anjing"} into,
{"kamu": 0, "tau": 1, "babi": 2, "gak": 3,
 "habis": 4, "kesabaran": 5, "saya": 6,
   "anjing": 7, "liat": 8, "ada": 9}
```

The segmentation result is a word vector in a matrix `(n,1)`, where n is the total number of unique words. The total vocabulary from the preprocessing steps is 641 words. This matrix then became the input for the ANN model.

In segmentation, each column represents a vocabulary. When we provide input data; "`kamu anjing`", means that there are 2 vocabularies `['kamu', 'anjing')`. Initially (before training), all vocab vector elements were initiated with 0, when it found the same word, the vector contents incremented as shown in Fig 3.

Fig. 3. Dataset conversion to vocab vector



Manual data labelling result is shown in table 1 below:

TABLE I. DATA LABELLING

Teks	Label	Category
`<p>kamu tau babi gak?</p>`	0	Non-offensive
Habis kesabaran gua anjing	1	Offensive
Liat ada njing	0	Non-offensive

In the ANN model, the size of the input layer (the number of neurons/nodes) is the same as the total number of the vocabulary vector; 641 neurons, 20 hidden neurons, and 1 neuron output. All 641 neurons (input layers) are extracted by sigmoid activation function into 20 neurons (hidden layers). Similarly, the output neuron extracts neurons from hidden layers using the sigmoid activation function.

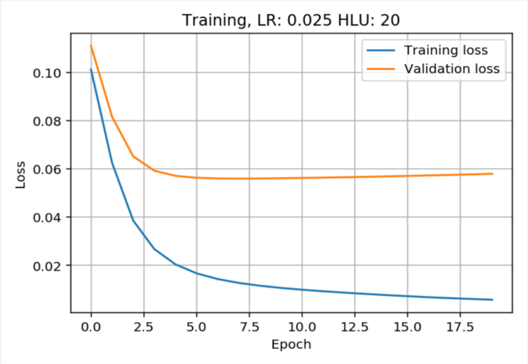


Fig. 4. Simulation Result

In table 2, we compare the training, validation, and testing result. From the table, we can conclude that the ANN model with 1 hidden layer and sigmoid function as an activation function can produce high accuracy result. Fig.4 shows that the gap between the result and the desired result (loss) decreases with the higher number of iteration (epoch).

TABLE II. RESULT ACCURACY

Dataset	Accuracy
Training	99.18%
Validation	94.28%
Testing	96.8%

We perform hyperparameter tuning in this research to optimize the result. First, we change the learning rate. It gives impact to the lost and accuracy as shown in Fig.5 and Fig 6. For the same number of iterations, the learning rate ≥ 0.025 show faster learning process compare to < 0.025. However, the testing accuracy performance in this region is similar (stagnant at around 94%). Another hyperparameter tuning that we change is epoch. Fig.7 shows that the larger number of iterations the higher training accuracy is. However, validation and testing accuracy become smaller which indicates overfitting.

Based on the hyperparameter tuning result, the optimum parameter for learning rate is 0.025 with the epoch 20. Better accuracy might be obtained by increasing the number of data, being covered in further research. Another improvement is in the method for collecting and labeling data to make it scalable.

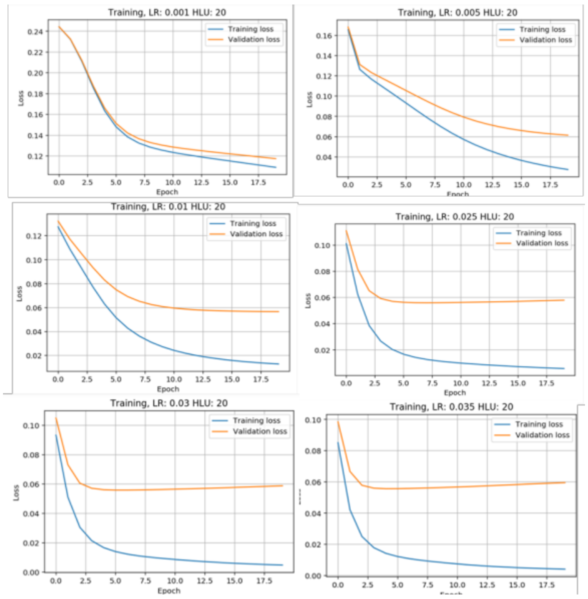


Fig. 5. Learning Rate Tuning Result – Lost

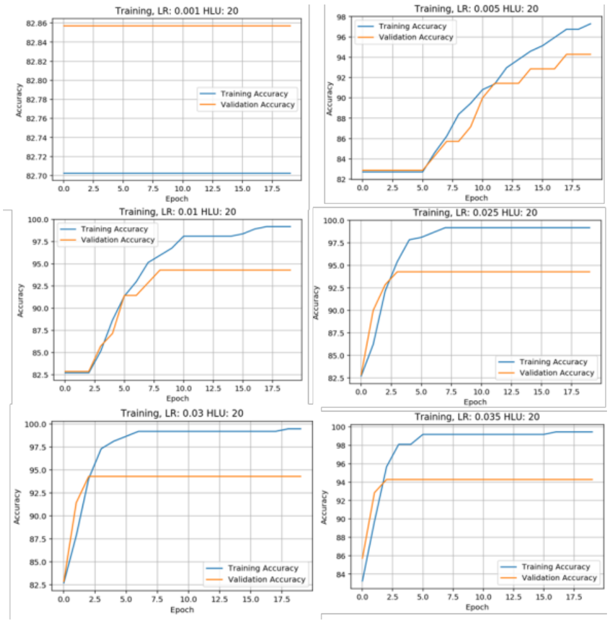


Fig. 6. Learning Rate Tuning Result - Accuracy

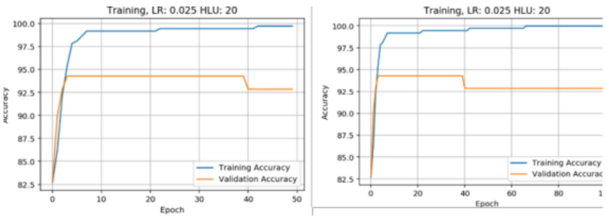


Fig. 7. Epoch Tuning

IV. Conclusion

Based on empirical result that shows 96.8% accuracy, ANN using a sigmoid function as the activation function effectively classifies an instance in Bahasa into offensive or non-offensive sentiment. Furthermore, the challenges such as informal grammar and abbreviation successfully tackled in the pre-processing step. This work is the first step in helping the government combating any offensive post in social media. The further improvements on the proposed method might be achieved by increasing the number of training data, adding hidden layer in the model, and changing the activation function using ReLU. These methods are left as the future work.

References

[1] A. P. J. I. Indonesia, "Penetrasi dan perilaku pengguna internet Indonesia Survei 2017," 2017. [Online]. Available: https://www.apjii.or.id/survei2017/. [Accessed 18 December 2018].

[2] A. Breland, "Social media fights back against fake news," The Hill, 27 May 2017. [Online]. Available: https://thehill.com/policy/technology/335370-social-media-platforms-take-steps-to-protect-users-from-fake-news. [Accessed 14 December 2018].

[3] S. Sood, J. Antin and E. F. Churchill, "Profanity use in online communities," in *SIGCHI Conference on Human Factor in Computing Systems*, Austin, Texas, 2012.

[4] M. S. Seidenberg and L. A. Petitto, "Communication, symbolic communication, and language: comment on savage-rumbaugh, McDonald, Sevcik, Hopkins, and Rupert (1986)," *Journal of Experimental Psychology General ,* vol. 116, no. 3, pp. 279-287, 1987.

[5] G. Laboreiro and E. Oliveira, "What we can learn from looking at profanity," in *11th International Conference PROPOR 2014*, San Carlos, Brazil, 2014.

[6] C. N. d. Santos, I. Melnyk and I. Pandhi, "Fighting offensive language on social media," in *56th Annual Meeting of the Association for Computational Linguistics* , Melbourne, Australia, 2018.

[7] L. Deng and Y. Liu, Deep learning in natural language processing, Singapore: Springer, 2018.

[8] M. Mayo, "KDnuggets," 2017. [Online]. Available: https://www.kdnuggets.com/2017/12/general-approach-preprocessing-text-data.html. [Accessed 18 December 2018].

[9] B. Vandersmissen, Automated detection of offensive language behavior on social networking sites, Ghent: Faculteit Ingenieurswetenschappen en Architectuur, 2012.

[10] "DeepAI," DeepAI, 2018. [Online]. Available: https://deepai.org/machine-learning-glossary-and-terms/neural-network. [Accessed 28 December 2018].

[11] H. Agrawal, "Facebook page feature : block words and profanity blocklist," Shout Me Loud, 11 August 2018. [Online]. Available: https://www.shoutmeloud.com/facebook-page-profanity-blocklist-spam-words.html. [Accessed 14 December 2018].

[12] A. Carman, "Instagram is now letting everyone filter abusive words out of their comments," The Verge, 16 September 2016. [Online]. Available: https://www.theverge.com/2016/9/12/12887514/instagram-comments-abusive-words-filter-section. [Accessed 14 December 2018].

[13] M. van Gerven and S. Bohte, "Editorial: artificial neural networks a models of neural information Processing," *Frontiers in Computational Neuroscience,* 19 December 2017.