Identify Abusive and Offensive Language in Indonesian Twitter using Deep Learning Approach

Muhammad Okky Ibrohim¹, Erryan Sazany², and Indra Budi³

Faculty of Computer Science, Universitas Indonesia, Depok, 16424, Indonesia

E-mail: muhammad.okky71@ui.ac.id¹, erryan.sazany71@ui.ac.id², indra@cs.ui.ac.id³

Abstract. Indonesia has a huge number of Twitter users and a lot of them often communicate using abusive language. Not only in the context of jokes, many Indonesian netizens often use abusive language to curse (offense) someone. Research about abusive language detection in Indonesian Twitter has been done using classical machine learning approach. However, the performance was still not too good, especially in differentiating whether the tweet is an abusive but not offensive or an offensive language. This paper implements a deep learning approach to enhance the performance when identifying abusive but not offensive or an offensive language. We use Long Short-Term Memory (LSTM) with word embedding because our literature study found that LSTM with word embedding is good for text classification (both for English or Indonesian text classification). The experiment result shows that LSTM with word embedding can increase the $F_1$-Score from the previous work until 19.44%, from 70.06% to 83.68%.

1. Introduction

Indonesia is one of the countries that have a huge number of social media users. In the social media, people are free to get any hot news, advertising something, establish communication with a new friend, sharing their activities, or just share what they were feeling at the time. However, the freedom of use of social media often leads many netizens to communicate using abusive language [1].

In Indonesia, the abusive language (expression that contains abusive/dirty word/phrase) are commonly come from unpleasant condition such as mental disorder, sexual deviation, physical disability, lack of modernization, condition where someone does not have etiquette, conditions that are not allowed by God or religion, and other conditions that related to unfortunate circumstances; animals that have bad characteristic, disgusting, and forbidden in certain religion; astral beings that often interfere with human life; a dirty and bad smell object; a part of the body and an activity that related to sexual activity; and low-class profession that forbade by religion [2-4].

Nowadays, in a modern conversation, not every abusive language is used to offense someone (use as offensive language). The abusive language is often used to express astonishment, wonder, amazement, etc. [2]. However, although not every abusive language is used to offend someone and just used in the context of jokes, the use of abusive language on social media can lead to misunderstandings and trigger conflict among netizens [5]. Furthermore, the use of abusive language that is not in place on social media also can cause children and adolescent learn the inappropriate language for their age [6].

To avoid conflict among netizens and protect the children and adolescent from learning the inappropriate language because of the use of abusive language that is not in place, the abusive language in the social media is must to be filtered. However, identify the abusive language in the social media is
not easy because it cannot just use word matching [7]. When speech an abusive language in the social media, netizens often use a very informal spelling and grammar. Especially on Twitter, the limitation character of a tweet is lead netizens to use abbreviations when posting a tweet. In Indonesia, the informal words often used by netizens when posting a tweet are words that show feelings, character repetition, changing vowels to numbers, and using slang words [8].

In recent years, many researches have been done on abusive language identification in various social media with the various approach. Many researchers are using several machine learning algorithms such as Naive Bayes (NB), Random Forest Decision Tree (RFDT), Support Vector Machine (SVM), k-Nearest Neighbor (kNN), etc.; with several features such as word n-grams, character n-grams, Term Frequency-Inverse Document Frequency (TFIDF), linguistic features, etc. [1, 6, 7]. In order to address the machine learning issue where heavily depends on the feature engineering and enhance the performance of abusive language identification system, some researchers implement several deep learning algorithms in building an abusive language identification system [5, 9, 10].

For the Indonesian language, research in abusive language detection is still very rare (based on our knowledge). We just found one previous work [4] about abusive language detection in the Indonesian language. In [4], they used a machine learning approach with simple word n-grams and character n-grams feature to identify the abusive language in the Twitter dataset. Their approach is good enough to identify whether a tweet is an abusive language or not. Unfortunately, their approach is poor enough when trying to classify the tweet into three labels which are not abusive language, abusive but not offensive, and offensive language. Here, their approaches faced difficulties to differentiate whether the tweet is an abusive but not offensive (abusive language in the context of jokes or a vulgar conversation) or offensive language.

In this paper, we use a deep learning approach to identify the abusive language in Indonesian Twitter dataset in order to address the issue in [4]. We use Long Short-Term Memory (LSTM) [11] as deep learning method with word embedding. LSTM is one of deep learning approach that can give an excellent performance when used for text classification [5, 10], including when used for Indonesian text classification [12].

The rest of this paper is organized as follows. Section 2 discusses related work, Section 3 discusses the dataset and proposed method, Section 4 presented the experiment result, and Section 5 discusses the conclusions and future works.

2. Related Work
Many researches have been done on abusive language detection in several social media with various language using various approach. In [6], they used NB and SVM with word n-grams (1-5 grams), appraisal approach, and lexical syntactical features to identify abusive language in English YouTube comment board.

Research about abusive language detection also has been done in the Thai language by [1] in Facebook comment dataset using several machine learning based-classifier such as kNN, NB, SVM, RFDT, etc.; with word n-grams (word unigram and word bigrams) and TFIDF for feature extractions. Their maximum performance reaches 86.01% of $F_1$-Score when using Discriminative Multinomial Naive Bayes (DMNB) as the classifier with IDF for feature extraction.

In recent years, many researchers used deep learning approach for text classification in order to address the machine learning issue where heavily depends on the feature engineering and enhance the performance [12]. In [5], they used Bi-Directional LSTM (BLSTM) and standard LSTM for inappropriate text detection on English web search queries and conversation. For inappropriate web search queries detection, they create a dataset that consist of 79,041 unique web search queries annotated by human judges (inappropriate/clean); while for inappropriate conversation detection they used Xbox conversations and Zo conversations\textsuperscript{1} dataset. Their experiment result shows that BLSTM can give slightly better performance with 81.49% of $F_1$-Score compared to standard LSTM with 78.50% of $F_1$-

\textsuperscript{1}https://www.zo.ai/
Score for web search queries dataset. For Zo conversations dataset, BLSTM also can give slightly better performance with 89.30% of $F_1$-Score compared to standard LSTM with 87.90% of $F_1$-Score. However, for Xbox conversations dataset, BLSTM cannot give better performance compared to the LSTM (78.60% of $F_1$-Score while using BLSTM and 78.90% while using LSTM).

Another researcher also implements LSTM for offensive language detection. In [10], they build an architecture to classify English Twitter dataset into three labels (neutral, sexism, and racism). They used multiple LSTM with special word embedding (word embedding with additional features for each class) that implemented using Keras\textsuperscript{2}. From the experiment result, multiple LSTM with special word embedding can give slightly better performance (93.20% of $F_1$-Score) compared with single LSTM with standard word embedding (no additional feature, 91.96% of $F_1$-Score).

For the Indonesian language, [4] have done a preliminaries research on abusive language identification in social media using Indonesian Twitter dataset. They used several traditional machine learning algorithms which are NB, SVM, and RFDT with simple word n-grams (unigram, bigrams, and trigrams) and character n-grams (4, 5 grams) features to classify the Twitter dataset. Their experiment results show that their approach is good enough when classifying tweet just into two labels (not abusive language and abusive language) with 86.43% of $F_1$-Score for maximum performance when using NB with word unigram feature. However, all their approach difficulties to differentiate whether the tweet is an abusive but not offensive or offensive language. They just reach 70.06% of $F_1$-Score for the maximum performance when classifying the tweet into three labels (not abusive language, abusive but not offensive, and offensive language) using NB with the combination of word unigram and word bigrams features.

In order to enhance the research result of [4], we use LSTM with word embedding for abusive language identification in Indonesian Twitter dataset. Some research results have shown that LSTM is good for resolving text classification [5, 10, 13]. Not only for text classification in English, but some research results also have shown that LSTM is good for Indonesian text classification [12]. In [12], they used LSTM with word embedding for Indonesian hate speech identification. They used Indonesian Twitter dataset provided by [14]\textsuperscript{3}. Their experiment results show that the LSTM with word embedding can give excellent performance with 95.39% of $F_1$-Score.

3. Dataset and Method
In this research, we use the LSTM network with word embedding, as done in [12]. The LSTM \textsuperscript{11} is one of Recurrent Neural Network (RNN) variant, which falls into deep learning approach. This method specializes in prediction with the sequential type of data. The LSTM model was adapted such that it could be applied for the multiclass classification task.

We used two pre-trained word embeddings\textsuperscript{4} built on FastText method introduced in [15] and Word2Vec method introduced in [16], both as the reflections of the word distributions. These could be used as a semantic representation which supports the learning process. In general, this research flowchart can be seen in Figure 1.

![Flowchart](image)

Figure 1. The research flowchart

\textsuperscript{2} https://github.com/keras-team/keras
\textsuperscript{3} https://github.com/ial_na/id-hatespeech-detection
\textsuperscript{4} https://github.com/Kyubyong/wordvectors
In order to compare the research result of [4], in this research we used their dataset that is open for public\(^5\). They crawled the Twitter data using Twitter API and Tweepy Library\(^6\) with a set of the abusive word as the query. For more detail about the dataset that we used including how they annotate the dataset, the reader is referred to [4].

We also applied a sort of cleansing processes to all tweets as done in [4], such as lower casing, retweet marks (RT) removal, Twitter username removal, URL removal, and excessive white spaces and line breaks handling. After the cleansing processes were done, a transformation was applied to change some non-formal words into formal ones. We used an Indonesian non-formal word dictionary given by [4].

After the data set was ready, we performed some preparations on the normalized data set and the word embedding, similar to what was done in [12]. Tokenization by space was the first thing applied to the data set. This process produced a dictionary which mapped token to incremental index. Then, we converted the tokens in the tweets and the words in the word embedding into an index using the dictionary mentioned previously. However, one issue that we encountered was the gaps in term of the vocabulary between the tweets and the word embedding. Since the words in the tweets were noisy and full of non-formal and slang words, it was possible to have some words in the tweets unavailable in the word embedding and vice versa. To handle the issue, we expanded the word-to-index dictionary with the words in the word embedding which did not exist in the tweets. From this point, both the data set and the word embedding were represented by numbers. Finally, the data set and the embedding were fed into the model.

To evaluate our classifier configuration, we used the 10-fold cross validation strategy [17]. This strategy will divide the dataset into 10 parts where 9/10 from the dataset will become the training data the others will become the testing data. The training and testing process will be repeated 10 times (10-fold) and each data will alternately become training data and testing data. For every fold, we use \(F_1\)-Score as the metric evaluation [18]. To avoid bad classification result because of the unbalanced data [19], we use Synthetic Minority Over-sampling Technique [20] that implemented using the SMOTE library for handling the unbalanced data problem.

4. Experiments and Discussions
In this experiment, we run our LSTM implementation that implemented using Keras deep learning framework\(^7\) and we set the backend to Tensorflow [21]. We use two alternate word embedding that are from Word2Vec and FastText and the results with the baseline from the previous work [4]. We configured the LSTM layer to 100, epoch count to 1, and batch size to 32, as done in [12]. At the final layer, we set the output final layer size into 3 as we wanted to classify the text to three labels as described before. Then, we set Adam [22] as the model optimizer and categorical cross entropy of Keras as the loss function. Our experiment result can be seen in Table 1.

From Table 1, we can see that our proposed method can outperform the result of [4]. Our LSTM implementation can increase the \(F_1\)-Score from previous work until 19.44%, from 70.06% to 83.68%. The LSTM with word embedding that we used successfully can differentiate whether a tweet is abusive but not offensive or offensive language better that all approach that used in [4].

\(^5\) [https://github.com/okkyibrohim/id-abusive-language-detection](https://github.com/okkyibrohim/id-abusive-language-detection)

\(^6\) [http://www.tweepy.org/](http://www.tweepy.org/)

\(^7\) [https://keras.io/](https://keras.io/)
Table 1. The $F_1$-Score of experiment results.

| Features                       | Baseline (NB) [4] | Proposed (LSTM) |
|--------------------------------|------------------|-----------------|
| word unigram                   | 69.50%           | -               |
| word bigrams                   | 50.75%           | -               |
| word trigrams                  | 27.33%           | -               |
| word unigram + bigrams         | 70.06%           | -               |
| word unigram + bigrams + trigrams | 69.64%     | -               |
| char trigrams                  | 66.67%           | -               |
| char quadgrams                 | 69.55%           | -               |
| Word2Vec                       | -                | 81.34%          |
| FastText                       | -                | 83.68%          |

5. Conclusions and Future Works

Research in abusive language identification in social media is very important to avoid conflict which caused by the used of abusive words/phrases that not in place on social media and also to create more safety social media for children and adolescent. We have discussed several previous studies working on the same problem with classical machine learning and then we compared the result of our implementation against the baseline method. In this research, we used the LSTM method which falls into a deep learning approach and equipped with pre-trained word embedding built on Word2Vec and FastText containing vocabulary in Indonesian. The result shows that our implementation gives 83.68% of $F1$-Score, increasing around 19.44% from the baseline that just reached 70.06% at maximum. Between the two pre-trained word embedding, the FastText one gave slightly better classification than Word2Vec one, specifically when distinguishing abusive but not offensive and offensive language.

For future work, we suggest using embedding built on the large-sized social media corpus to capture more semantics in the learning process. This might be needed to handle the high noise or variety of writing in the dataset. Another type of embedding such as emotion embedding also one of the potential areas to be researched. Besides, the dataset itself should be increased in quantity or tested to other datasets to see if our model generalizes well to unseen data.

Acknowledgment

The authors acknowledge the PITTA research grant 1881/UN2.R3.1/HKP.05.00/2018 from Directorate Research and Community Services, Universitas Indonesia.

References

[1] Turaob S and Mitrpanont J 2017 Automatic discovery of abusive thai language International Conference on Asia-Pacific Digital Libraries pp 267-278
[2] Wijana I D P and Rohmad M 2010 Sosiolinguistik: Kajian, Teori, dan Analisis (Pustaka Pelajar, Yogyakarta)
[3] Triadi R 2017 Penggunaan makian bahasa indonesa pada media sosial (kajian sosiolinguisik) Jurnal Sasindo Unpam vol 5(2) pp 1-26
[4] Ibrohim M O and Budi I 2018 A dataset and preliminaries study for abusive language detection in Indonesian social media Procedia Computer Science vol 135 pp 222-229
[5] Yenala H, Jhanwar A, Chinnakotla M K and Goyal J 2017 Deep learning for detecting inappropriate content in text International Journal of Data Science and Analytics
[6] Chen Y 2012 Detecting offensive language in social media to protect adolescent online safety Expert Systems with Applications vol 36 pp 71-80
[7] Nobata C, Tetreault J, Thomas A, Mehdad Y and Chang Y 2016 Abusive language detection in online user content International World Wide Web Conference Committee (IW3C2) pp 145-153
[8] Hanafiah N, Kevin A, Sutanto C, Fiona, Arin Y and Hartanto J 2017 Text normalization algorithm
on twitter in complaint category Procedia Computer Science vol 116 pp 20-26

[9] Park J H and Fung P 2017 One-step and two-step classification for abusive language detection on twitter Proceedings of the First Workshop on Abusive Language Online (Association for Computational Linguistics) pp 41-45

[10] Pitsilis G K, Ramampiaro H and Langseth H 2018 Detecting offensive language in tweets using deep learning CoRR

[11] Hochreiter S and Schmidhuber J 1997 Long short-term memory Neural Computation vol 9 pp 1735-1780

[12] Sazany E and Budi I 2018 Deep Learning-Based implementation of hate speech identification on texts in Indonesian: Preliminary study 2018 International Conference on Applied Information Technology and Innovation (ICAITI 2018)

[13] Badjatiya P, Gupta S, Gupta M and Varma V 2017 Deep learning for hate speech detection in tweets International World Wide Web Conference Committee

[14] Alfina I, Mulia R, Fanany M I and Ekananta Y 2017 Hate speech detection in the indonesian language: A dataset and preliminary study International Conference on Advanced Computer Science and Information Systems (ICACSIS) pp 233-238

[15] Joulin A, Grave E, Bojanowski P and Mikolov T 2016 Bag of tricks for efficient text classification arXiv preprint:1607.01759

[16] Mikolov T, Sutskever I, Chen K, Corrado G S and Dean J 2013 Distributed representations of words and phrases and their compositionality Advances in Neural Information Processing Systems 26 ed Burges C J C, Bottou L, Welling M, Ghahramani Z and Weinberger K Q (Curran Associates, Inc.) pp 3111-3119

[17] Kohavi R 1995 A study of cross-validation and bootstrap for accuracy estimation and model selection Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2 IJCAI'95 (San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.) pp 1137-1143

[18] Hossin M and Sulaiman M N 2015 A review on evaluation metrics for data classification evaluations International Journal of Data Mining Knowledge Management Process vol 5 pp 1-11

[19] Ganganwar V 2012 An overview of classification algorithms for imbalanced datasets International Journal of Emerging Technology and Advanced Engineering vol 2(4) pp 42-47

[20] Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 SMOTE: Synthetic minority over-sampling technique Journal of Artificial Intelligence Research vol 16 pp 321-357

[21] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M et al. 2016 Tensorflow: a system for large-scale machine learning. OSDI vol 16 pp 265-283

[22] Kingma D and Ba J 2015 Adam: A method for stochastic optimization International Conference on Learning Representations