Sentiment Analysis on Social Media to Detect Toxic Comments

*

1st Vincent
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
vincent053@binus.ac.id

2nd Edric Pratama Widjaja
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
edric.widjaja@binus.ac.id

3rd Wanda Safira
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
wanda.safira@binus.ac.id

***Abstract*— Nowadays, the internet and social media have data that contains information that can be used to conduct research. The data taken are comments on social media divided into two types, namely toxic and non-toxic comments. Several sentiment analysis methods can be used to classify this type of comment, using LSTM, Naive Bayes, Decision Tree, KNN, and MLP. The dataset used is the cyberbullying dataset from the data.mendeley.com and kaggle.com sites. This case study aims to try sentiment analysis using the same data but with different methods. After that, it will be determined which method is the best for conducting sentiment analysis. This process shows that the LSTM approach is the most effective because it has an average accuracy rate of 79.5%.**

Keywords—LSTM, social media, toxic, Classification, sentiment analysis

I. INTRODUCTION

Social media is a part of the internet that facilitates users to express themselves, collaborate, interact, share, and communicate with other users. Thus, forming social bonds in the virtual world [11]. In 2020, Hootsuite stated that half of the world's population is already using social media. Start from 5,2 billion who have mobile phones, 4,5 billion who are connected to the internet, and 3,8 billion people who use social media actively. Meanwhile, the value of 3,8 billion equals 49% of the earth's population and an increase of 9% during 2020 [12].

Over time, social media has become a platform for complaining or expressing feelings in the form of typing. Usually, it is in the form of comments or tweets that are uploaded to their social media timeline. However, some users are not wise enough to use social media. So, there are often comments or tweets that contain negative traits such as blaspheming each other or saying rude things in making these comments or tweets. Even the use of unwise words often contains elements of SARA and pornography. Kominfo (Kementerian Komunikasi Republik Indonesia) has taken many ways to reduce negative actions on social media, and they have recognized how difficult it is to eradicate negative content on social media. Therefore, we created a journal by raising the topic of sentiment analysis on social media related to negative comments. One of the methods we will use is Long Short-Term Memory (LSTM). LSTM algorithm is a type of architecture of the Recurrent Neural Network (RNN), which is commonly used in problems related to deep learning [13]. LSTM was created by Hochreiter and Schmidhuber in 1997 and was developed and popularized by many researchers. LSTM can remember a collection of information that has been stored for a long time and delete information that is no longer relevant. This algorithm is more efficient and can process, predict, and classify data based on a certain time sequence. Using this algorithm, we can quickly identify words that fall into the negative category and ultimately eliminate those comments on several social media. Besides that, there are also several methods that can be used for sentiment analysis such as, Naive Bayes, Decision Tree, KNN, and MLP.

II. LITERATURE REVIEW

Social Media is a medium that is used to interact with each other and present themselves for an unlimited time to encourage the value of social interaction. However, many people cannot separate good and bad language in presenting themselves. So many people also feel shunned by the bad language given to others on social media. Therefore, there is a need for research to classify language on social media using sentiment analysis. Sentiment analysis is a branch of text mining research that works through several stages, namely the understanding stage, the extraction stage, and the automatic data processing stage, to obtain the sentences contained in the opinion sentences [7]. It is estimated that 80% of the world's data is unorganized in a predetermined way. Most of the data come from text, for example, reviews, emails, surveys, or chats on social media. The data is difficult to understand, so many companies use sentiment analysis to understand these unstructured texts. So, companies can automate business processes, save manual processing time, and gain actionable insights.

Many previous research papers have been used as a reference for conducting sentiment analysis on people who express themselves on social media. So, from the previous research, machine learning methods emerged to classify a sentence that people uploaded to social media to be positive or negative. There are several methods used in conduction sentiment analysis, namely Naive Bayes, DNN, CNN, RNN, SVM, TFID, KNN, and MLP. Of the many machine learning methods selected, the LSTM method is one method that has a fairly high accuracy of 85% in Dr. Gorti Satyanarayana Murty with the title "Text-Based Sentiment Analysis using LSTM" [1]. It can be compared with several other methods, such as the MLP method, namely the accuracy of 67,45% for the train test set and the test set is about 52,60%. As for using

the technique from the support vector machine, based on the results of a test conducted on the analysis of public sentiment towards the relocation of the Indonesian capital, which consists of 2 labels, namely positive and negative, using the RapidMiner. It shows an accuracy of 77,72% with True Positive 1 data, True Negative 770 data, False Positive 1 data, and False Negative 220 data [8]. So, to find sentiment analysis on social media, we can use LSTM as a settlement step.

III. METHODOLOGY

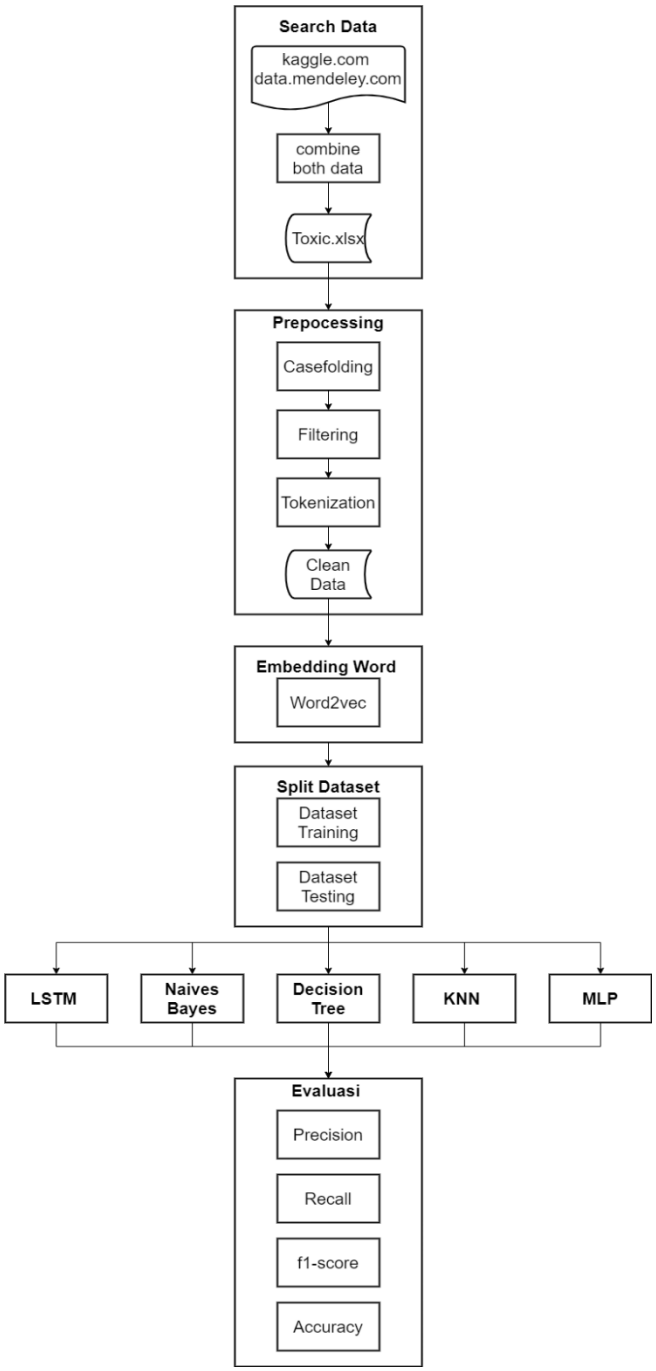

Fig. 1. Flowchart Sentiment Analysis

A. *Dataset*

	Text	Label
0	Does N.e.bodyelse Hear her Crazy ass Screamin ...	0
1	There are so many things that are incorrect wi...	0
2	3:26 hahah my boyfriend showed this song to me...	1
3	dick beyonce fuck y a ass hole you are truely ...	1
4	DongHaeTaemin and Kai ;A; luhansehun and bacon...	0
...	...	...
5024	::It shows that Dawkins is a liar.	1

Fig. 2. Dataset of toxic comment.

Our system's first phase is to collect comments from social media for our training dataset. We have used separate polarity datasets, each social media that fell into two categories: Nontoxic and Toxic, Which toxic labels with one and nontoxic labels with zero. The comments in the datasets were taken from Mendeley and Kaggle. Figure 2 shows that the dataset has 5025 comments from social media platforms, consisting of 2711 positive comments and 2314 negative comments.

B. *Pre-processing*

The second phase is to preprocess the stored comments. Comments may contain ambiguous data, making it difficult for the system to determine. Therefore, we much preprocess the retrieved data to normalize it. Case Folding is a process that converts all letters to lowercase. Meanwhile, other characters that are not letters and numbers, such as punctuation marks and spaces, are considered delimiters. This delimiter can also be removed or ignored. Tokenizing is removes numbers and punctuation, such as symbols and punctuation that are unimportant, and removes whitespace [16]. Filtering is the process of eliminating words that do not have a specific meaning.

	Text	Label
0	does nebodyelse hear her crazy ass screamin ho...	0
1	there are so many things that are incorrect wi...	0
2	326 hahah my boyfriend showed this song to me ...	1
3	dick beyonce fuck y a ass hole you are truely ...	1
4	donghaetaemin and kai a luhansehun and bacon x...	0
...	...	...
5024	it shows that dawkins is a liar	1

Fig. 3. The Dataset shown above has already been processed.

C. *Embedding Word*

Word embedding is converting terms or words that are alphanumeric into vector form. Each term or word is a vector that represents a point in space using a predetermined dimension [15].

Word2Vec is a method for embedding words that are used to represent words as vectors. Word2Vec uses a neural network to convert words into vectors. Word2Vec works by entering the existing text content as input and producing a vector representation of each word in the text content as output [15].

After doing Embedding Word with Word2Vec, we get 55.451 for the vocab, 50 for the size, and 0.025 for the alpha.

D. *Split data set*

To get a proportional dataset and maximum results, we divide a dataset into 80% training data and 20% test data. This makes the training data have 4,021 data and 1,005 test data.

E. *LSTM*

Long Short-Term Memory (LSTM) is a machine learning strategy used to solve classification problems. Furthermore, LSTM is an RNN (recurrent Neural Network) variation comprising cell, input gate, output, and forget. The input is taken and stored by the LSTM cell. Meanwhile, the input gate determines how far the value will travel into the cell, and the forget gate determines how long the value will remain in the cell. The output gate determines how far the value in the cell is utilized to calculate the LSTM unit's activation output [13].

```
model = Sequential()
model.add(w2vc_model.wv.get_keras_embedding(True))
model.add(Dropout(0.2))
model.add(LSTM(50, return_sequences=True))
model.add(GlobalMaxPooling1D())
model.add(Dropout(0.2))
model.add(Dense(1))
model.add(Activation('sigmoid'))
model.summary()
```

Fig. 4. Architecture model of LSTM.

To compile the LSTM model, we use binary cross-entropy, optimizer Adam, and metrics accuracy with iterations 10, 30, and 50. For the LSTM architecture model in figure 4, we use a sequential model consisting of Word Two Vector, LSTM functions, Global Max Pooling, Dropout, and Dense as outputs

F. *Naïve Bayes*

Naïve Bayes is a classification method based on simple probabilities and is made with the assumption that one attribute and another attribute are not dependent on each other or is usually called independent. The advantages of the Naïve Bayes method are that it can use small amounts of data when classifying and is efficient and easy to make. The disadvantage of the Naïve Bayes method is that the independent nature between attributes makes accuracy weak and does not apply if the probability value is zero (0) [3].

$$P(Sentiment \mid Sentence) = \frac{\text{P(Sentiment)P(Sentence|Sentiment)}}{\text{P(Sentence)}}$$

P(Sentiment | Sentence) = Posterior Probability
P(Sentence | Sentiment) = Likelihood
P(Sentiment) = Class Prior Probability
P(Sentence) = Predictor Prior Probability

G. *Decision Tree*

Decision tree is a method used for data mining. Data mining is a solution that is often applied to classifying. Decision Tree uses the supervised machine learning method, which means that new data will be classified based on the existing training sample. A decision tree has a root, an internal node, and a leaf. Each node represents an attribute, and an existing branch will represent the attribute's value, while the leaves are used to represent the class. The node at the top is called the root [14].

H. *K-Nearest Neighbor*

K-Nearest Neighbor (KNN) is a machine learning algorithm with the simplest supervised learning. KNN is a nonparametric algorithm. Therefore, KNN does not make assumptions about the data. The advantages of using KNN are that it is easy to understand, easy to apply, and can be used in many different classes. The disadvantage of using KNN is high computation cost, long processing time, and sensitive data with a lot of noisy data, missing data, and outliers [6].

We use GridSearchCV to determine the best parameters in the KNN method to get good accuracy. So we get that the best KNN neighbor is 7.

I. *MLP*

A Multiplayer perceptron (MLP) is a neural network that connects many layers in a directional graph. Where the signal path through the nodes only goes in one direction. The advantage of using MLP is that it can solve problems that cannot be solved by linear solutions and gives a good prediction for different values. The disadvantage of using MLP is that MLP cannot be used for sequence data modeling [10].

To get good accuracy, we use GridSearchCV to determine the best parameters in the MLP method. So, we use solver Adam and iteration 20.

IV. Result and Analysis

To prove that the model is one of the best classification methods, we compared it with four other machine learning methods, namely Naive Bayes, Decision Tree, KNN, and MLP. So that the following results are obtained:

TABLE I. MODEL COMPARISON

Word Embedding	Model	Accuracy	Precision	Recall	F1Score
Word2vec	LSTM (10 Iterations)	77,8%	72,8%	82,7%	77,5%
	LSTM (30 Iterations)	80,1%	85,8%	68%	75,9%
	LSTM (50 Iterations)	80,5%	81,6%	74,5%	77,9%
	Naive Bayes	53,2%	49,6%	94,2%	65%
	Decision Tree	66,6%	63,4%	64,8%	64,1%
	KNN	62,6%	56,1%	87%	68,2%
	MLP	61,8%	58,7%	57,7%	58,2%

The comparison table shows that the LSTM method is the best method for classifying a comment as toxic or non-toxic. LSTM has the lowest accuracy at iteration 10 of 77.8%, and the highest accuracy at iteration 50 is 80.5%. If we calculate the average accuracy of the LSTM method with the iterations of 10, 30, and 50, then we get an average of 79.5%. In the experimental model, if we want to increase the accuracy of the LSTM method, then we can increase iterations on the model. For the Naive Bayes method, Decision Tree, KNN, and MLP have an adequate accuracy with about 53% - 62%.

V. CONCLUSION

After analyzing the 5 machine learning methods, we draw the following conclusions:

1. LSTM is one of the best methods for classifying text. This can be seen from the high LSTM accuracy, which is 82% compared to other methods.
2. For the Naive Bayes method, Decision Tree, KNN, and MLP can still be used for classification even though they have low accuracy.
3. To improve accuracy in the model, we can change iterations and parameters.

VI. FUTURE WORK

In further research, we can add other methods that can be used for sentiment analysis and determine whether these methods have a higher level of accuracy than LSTM. After trying several of these methods, we can compare which ways are easy and good to use in sentiment analysis.

REFERENCES

[1] Murthy, Allu, Shanmukha Andhavarapu, Bhargavi, Bagadi. "Text based Sentiment Analysis using LSTM", International Journal of Engineering Research, vol 9, 2020.

[2] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection", in IEEE Access, vol. 6, pp. 13825-13835, 2018.

[3] S. Suparyati, F. Agus, "Analisis Sentimen Dengan Klasifikasi Naive Bayes pada Review Hotel Tripadvisor", 2020.

[4] A. Gaydhani, V. Doma, Shrikant kendre and L. Bhagwat, "Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TF-IDF base Approach", 2018.

[5] Dang, N. C, G. Morneo, M. N, De la Prieta, F, "Sentiment analysis based on Deep Learning: A Comparative Study", Electronics, 9(3), 483, 2020.

[6] Septian, Jeremy, Fahrudin, Maulana.T, Nugroho, Aryo," Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor", Journal of Intelligent System and Computation, 1, 43-49, 2019.

[7] T.H.F. Tezza, Garnoa, dan R. Azhari, "Analisis Sentimen Pemindahan Ibu Kota Pada Twitter Dengan Metode Support Vector Machine", Jurnal Ilmu Komputer, vol. 14, No. 2, 2021.

[8] T. Elghazaly, Mahmoud, A. Hefny, "Political sentiment analysis using Twitter data", Proceedings of the International Conference on Internet of Things and Cloud Computing, 2016.

[9] J. Yadla, S. K , M. Ravilla, S. Madhu, "Sentiment analysis using logistic regression algorithm", European Journal of Molecular & Clinical Medicine, ISSN 2515-8260 ,vol 7, Issue 4, 2020.

[10] A. M. Ramadhani and H. S. Goo, "Twitter sentiment analysis using deep learning methods", 2017 7th International Annual Engineering Seminar (InAES), 2017.

[11] Nasrullah, Rulli. "Media Sosial Perspektif Komunikasi, Budaya dan Sosioteknologi. Bandung: Simbiosa Rekatama Media", 2015.

[12] Kemp, Simon, "Digital 2020, Global Digital Overview", 2020.

[13] X. Song, J., Huang., dan D. Song, "Air quality prediction based Long Short-Term Memory (LSTM) neural network model, "J. Pet. Sci. Eng., vol 186, 2020.

[14] A. Bayhaqy, S. Sfenrianto, K. Nainggolan, & E. R. Kaburuan, "Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes", International Conference on Orange Technologies,2018.

[15] D. Jatnika, M. A. Bijaksana, & A. A. Suryani, "Word2Vec Model Analysis for Semantic Similarities in English Words.", Procedia Computer Science, 157, 160–167, 2019.

[16] B. M. Akcay & K. Oguz, "Speech Emotion Recognition: Emotional Models, Databases, Features, Preprocessing Methods, Supporting Modalities, and Classifiers." Speech Communication, 2682. 2019.