Hate Speech Detection using Convolutional Neural Network Algorithm Based on Image

Bagas Prakoso Putra[1], Budhi Irawan[2], Casi Setianingsih[3],
Annisa Rahmadani[4], Farradita Imanda[5], Izzu Zantya Fawwas[6]

School of Electrical Engineering, Telkom University, Bandung, Indonesia

[1]imbagas@student.telkomuniversity.ac.id, [2]budhiirawan@telkomuniversity.ac.id,
[3]setiacasie@telkomuniversity.ac.id, [4]annisarahmadani@student.telkomuniversity.ac.id,
[5]farraditai@student.telkomuniversity.ac.id, [6]izzuzantyaf@student.telkomuniversity.ac.id

2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE) | 978-1-6654-0544-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISMODE53584.2022.9742810

Abstract— **Hate speech is words behavior that can cause an attitude of violence and anarchy against other individuals or groups. The internet has become necessary in this day and age, so internet morals need to be considered. However, several parties deviate from using the internet to spread hate speech, such about race, ethnicity, and religion. Nowadays, hate speech detection systems are usually through text but hate speech detection through images tends to be rare. For that reason, this study is aimed to detect whether there is hate speech or not in the selected image. This project uses the Convolutional Neural Network (CNN) algorithm and Deep Learning method to classify the aspect of hate speech contained in an image and recognize any hate speech on the image through the existing text. After this application is developed, the machine learning system can detect some hate speech on an image that contains the text. It achieves about 95.89% accuracy and 94.43% precision. After that, the authors hoped that the authorities could reduce hate speech in the community and follow up more quickly.**

Keywords: ***Convolutional Neural Network (CNN), Deep Learning, Hate Speech, Text Classification.***

I. INTRODUCTION

On August 25, 2017, three people from Saracens who are perpetrators of hate speech were arrested by the Criminal Investigation Police. They spread hate content directed at a particular group of people, religious conflicts, and concepts that cause the denigration of the related person through 800,000 accounts starting from 2000 social media accounts [1]. Under article number 27 in paragraph 3 of ITE Law, "Any person intentionally and without right or transmitting or making accessible electronic information or electronic documents containing contempt and/or defamation". Through social media that can spread photos, such as Facebook, Twitter, and Instagram, some individuals irresponsibly spread hate speech images on various social media because the image becomes attractive and accessible media by society [2].

In this system, using the input in the form of images that will be converted into text, the text will be analyzed, i.e., the text that hates speech or not, by implementing the Deep Learning method with Convolutional Neural Network (CNN) algorithm. Deep Learning is excellent in vision, sentiment analysis, etc. Sentiment analysis is already in many studies using Deep Learning models because it has connections and parameters that are easier and easier to process; CNN is a compelling model for recognizing image contacts. Section II provided related work from previous research to explain the related study and compare it with current research.

II. RELATED WORK

There are quite a few studies that talk about Convolutional Neural Networks (CNN) for text and sentence classification purposes. Yoon Kim (2014); the researcher trained the CNN on pre-trained word2vec for doing sentence-level classification tasks. The outcomes reveal that excellent results on multiple benchmarks can be attained with small hyperparameter tuning and static vectors on simple CNN. The outcomes increase the strong evidence that unsupervised pre-trained word vectors are a precious component in Deep Learning for Natural Language Processing purposes [3].

In their research, Ye Zhang and Byron Wallace (2015) tried to search the aftermath of architecture elements on model performance handled by sensitivity analysis of one-layer CNN. The author intends to distinguish between crucial and relatively inconsequential design decisions for sentence classification. Because of their comparative simplicity and solid empirical performance, the author's focal point is on one-layer CNNs meant to separate from more complex models, making it a modern standard baseline method akin to Support Vector Machine (SVMs) and logistic regression. This paper has done a comprehensive experimental analysis of CNNs for sentence classification. It then can deploy CNNs in real word sentence classification scenarios by summarizing significant discoveries and getting practical guidance [4].

Another Deep CNN for doing sentiment analysis inside a short text that utilizes character in sentence-level information was proposed by C´ıcero Nogueira dos Santos and Ma´ıra Gatti (2014). The latest results of their research are 85.7% accuracy for single sentence sentiment prediction in both binary positive/negative classification and 48.3% accuracy for fine-grained classification. The sentiment prediction accuracy is approaching 86.4% for the STS corpus [5].

A new semi-supervised framework for text categorization purposes that utilize Convolutional Neural Networks (CNNs) was developed by Rie Johnson and Tong Zhang (2015). The method we use is different from previous studies that used word insertion. We integrate small-text insertion of unlabeled data into a supervised CNN from studying it. The intended scheme for embedding learning is based on the idea of semi-supervised two-sighted learning, which is proposed to be helpful for interesting tasks even if

Authorized licensed use limited to: University of Plymouth. Downloaded on February 19,2024 at 14:41:13 UTC from IEEE Xplore. Restrictions apply.

the training is carried out on unlabeled data. The model we have created works better for sentiment and topic classification tasks, and it produces better results than the previous studies. Section III provided a research method that was used in this paper [6].

III. RESEARCH METHOD

The central concept of this research is to obtain text information from an image then classify the text into hate speech or not. The data is collected from Twitter about 1,223 data as dataset and training, about 645 hate speech and 578 non-hate speech data, which Dinas Balai Bahasa Kota Bandung has verified.

A. *Hate Speech*

Hate speech is a form of harassment, insult, or incitement from an individual or group to another individual or group, which can be in the form of words or actions in terms of numerous aspects such as gender, citizenship, religion, color, race, disability, sexual orientation, and others.

B. *Data Mining*

Data mining is data searching by the data owners to discover advantageous knowledge and information [7]. Data mining combines several existing technologies such as database systems, machine learning, statistics, and visualization that are considered a fast-growing technology.

C. *Optical Character Recognition (OCR)*

Optical Character Recognition (OCR) is a frequently used module for converting data from image to text. It works to scan photos, recognize if it has text or not, and create editable text. The text scanned by OCR can be searched, manipulated, replaced, or given barcode line by line. OCR can support the scanner as an additional application. Using OCR, images like handwritten images, typewriter images, or computer text can be editable and manipulated [8].

D. *Word2vec*

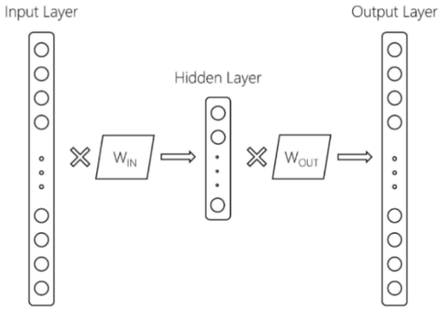


Fig 1. Word2vec representation

Word2Vec is a model used for representing words into vectors. Word2vec model, windows size, and vector dimensions configuration can be produced by using several features. In some previous studies, Windows size and vector dimensions configuration is frequently used for creating the Word2vec model [9]. Fig. 1 shows the illustration of Word2vec representation.

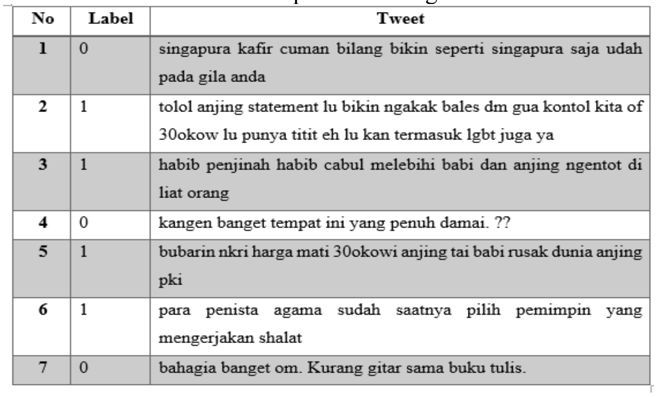
E. *Convolutional Neural Network*

Convolutional Neural Network (CNN) is the most frequently used computer vision and image processing model. They are designed like that to mimic the animal's visual cortex structure. In particular, three dimensions arrange neurons: width, height, and depth. In addition, in specific layers, neurons are only connected to a small area of the previous layer [6].

F. *Example Convolutional Neural Network, Study Case Hate Speech*

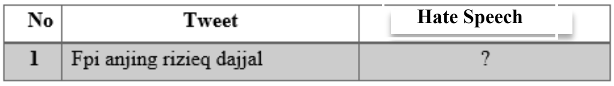
In the table below are examples of the training and testing data. Label 0 is non-hate speech, and label 1 is hate speech.

TABLE I. Example of Training Data CNN



No	Label	Tweet
1	0	singapura kafir cuman bilang bikin seperti singapura saja udah pada gila anda
2	1	tolol anjing statement lu bikin ngakak bales dm gua kontol kita of 30okow lu punya titit eh lu kan termasuk lgbt juga ya
3	1	habib penjinah habib cabul melebihi babi dan anjing ngentot di liat orang
4	0	kangen banget tempat ini yang penuh damai. ??
5	1	bubarin nkri harga mati 30okowi anjing tai babi rusak dunia anjing pki
6	1	para penista agama sudah saatnya pilih pemimpin yang mengerjakan shalat
7	0	bahagia banget om. Kurang gitar sama buku tulis.

TABLE II. Example of Testing Data CNN



No	Tweet	Hate Speech
1	Fpi anjing rizieq dajjal	?

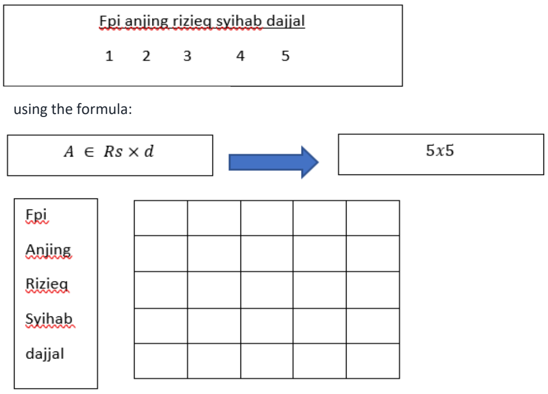


Fig 2. Input to Word2vec

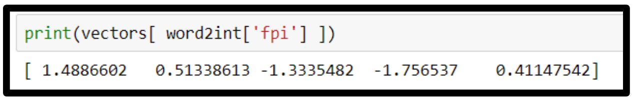
A detailed explanation of each step of word formation until it is classified by CNN to solve the problem is in Table 2 in Fig. 2. Before getting the value of the input layer, we must specify the size and contents of the matrix using Word2vec. From table 2, there are 5 words, so the dimension embedding matrix used is 5.

Separating each word to get the respective value of the word can be done by using Word2vec. Then "fpi", "anjing" "rizieq" "syihab" "dajjal", where 5 words are found in the sentence. Use the formula: $A \in Rs \times d$.

Where A defines the input layer, Rs represents the number of words contained inside the sentence, and d represents the number of dimensional layers used in the system. With a dimension of 5 and 5 words in the sentence, the input layer matrix was 5 x 5. Each word processed by

Authorized licensed use limited to: University of Plymouth. Downloaded on February 19,2024 at 14:41:13 UTC from IEEE Xplore. Restrictions apply.

Word2vec generates an output value inserted into the previous input layer. For each word, we get the following matrix values shown in Fig. 3. Obtained vector value: Fpi = [1.4886602 0.51338613 -1.3335482 -1.756537 0.41147542] For the words "anjing", "rizieq", "syihab", and "dajjal" performed the same process using Word2vec.



```
print(vectors[ word2int['fpi'] ])
[ 1.4886602   0.51338613 -1.3335482   -1.756537    0.41147542]
```

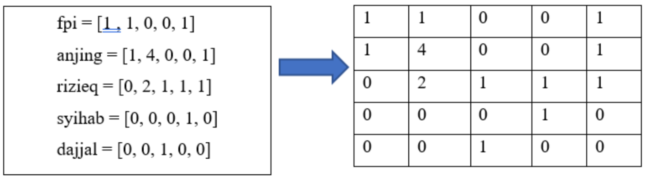
Fig 3. Input to Word2vec

TABLE III. Result of the sentence from Word2vec

Fpi	[1.4886602 0.51338613 -1.3335482 -1.756537 0.41147542]
Anjing	[1.1676364 3.5387237 -0.872862 -0.2513935 0.7912404]
Rizieq	[-0.21822342 2.3067977 1.0291293 0.65628636 0.6740474]
Syihab	[-1.6249902, -1.1351361, -1.4234633, 1.2607635, 0.38661343]
dajjal	[-0.07134825 -1.0066274 0.48204803 -0.03595227 1.8018485]

A simplifying calculation process is needed before the value is inserted into the input layer matrix because it will help speed up the convolution process [10]. The end result of rounding for the input layer vector value. Then the final result is then inserted into the CNN input layer for processing as a preliminary step [11] before filtering can be seen in Fig. 4 [12] and CNN variables in Fig. 5. Region size defines the range of filters to be used, while the filter size determines the number of filters used in this convolutional neural network method. With the size of the specified region size are [3, 4, 5] and filter size of 2 then the number of filters used in this method is as follows: using the formula in Fig. 3 it is defined that F represents the number of filters, f means the size of the filter and s is the region size. It was found that the filter size was 6.

Explained that there are 6 filters, then each input is mapped into 6 pieces of filter. Then each filter has a different size matrix. This value is based on the region size value previously discussed [3][4][5]. The implementation is shown in Fig. 6.



fpi = [1, 1, 0, 0, 1]
anjing = [1, 4, 0, 0, 1]
rizieq = [0, 2, 1, 1, 1]
syihab = [0, 0, 0, 1, 0]
dajjal = [0, 0, 1, 0, 0]

1	1	0	0	1
1	4	0	0	1
0	2	1	1	1
0	0	0	1	0
0	0	1	0	0

Fig 4. The Preliminary Step

Parameter	Value
Region size	3, 4, 5
Filter size	2
Stride	1
Padding	0
Total Filter	F = f x s = 2 x 3 = 6

Fig 5. CNN Variables

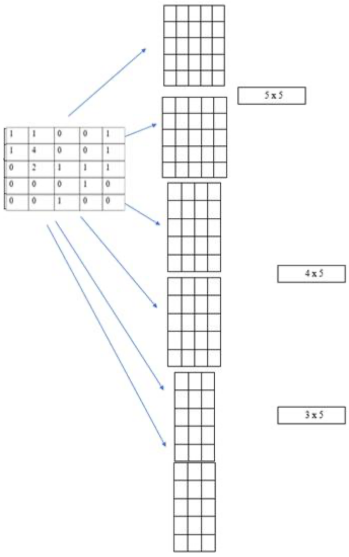


Fig 6. CNN Filters

After each filter has a multiplier of the wise element value, For the calculation of convolution layer process using sliding window because padding and stride size value 0, there is no moving process on the sliding window then the multiplier is reset. Fig.7 shows the convolution process to a predetermined filter [13].

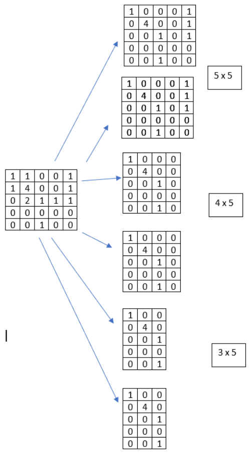


Fig 7. The Convolution Process

The next stage is the activation function; at this stage, every result layer that has been processed on the convolution layer will be on the map again. Decide the size of the output layer on the activation function requires the formula: $\frac{N+2P-F}{S}+1$

After obtaining the output layer matrix size, the next step is to move the value of the convolution layer into our output layer so the process activation function runs. The activation function is the process of activating the neuron [14]. The following functions: $ci = f\ (oi + b)$ Ci represents activation function f is for the filter oi for the output layer and b for bias. The bias value is added when one of the neurons does not meet. If the value of the neuron is 0, then it needs additional bias, but if it meets more than 0, it is unnecessary. Furthermore, after the activation function is obtained, the next step is to map the results by making a feature map of the activation function to be processed to the next stage. Max-pooling for the feature map implementation is shown in Fig 8.

Authorized licensed use limited to: University of Plymouth. Downloaded on February 19,2024 at 14:41:13 UTC from IEEE Xplore. Restrictions apply.

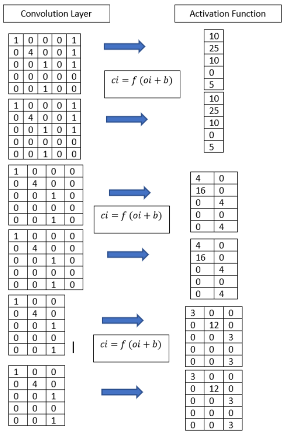


Fig 8. Feature Map Implementation

The largest or maximum max pooling is done on the next layer of the max-pooling layer, the largest value is taken so that the neural network can make the selection of the value of each layer that can be taken to the next layer to do the process of weighting and get the probability value on softmax regularization. Here is the function of the max-pooling layer: $c = max\{c\}$.Where c is the max-pooling layer and max is the fungi to retrieve the largest or maximum value of the layer activation function of each layer already in the filter [15]. Where c is the max-pooling layer and max is the fungi to retrieve the largest or maximum value of the layer activation function for each layer already in the filter.

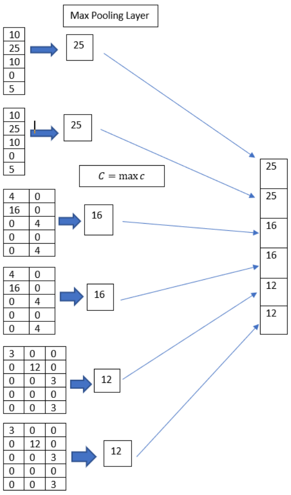


Fig 9. Max Pooling Layer

For this softmax, the total probability is always close to 1. Each value is 0.5, 0.3, and 0.2 of the 3 vectors. The largest value is at 0.5 vectors 1, and the implementation is shown in Fig.10.

The most significant value inside vector 1 is 0.5, that value is classified as 1, and the others are classified as 0. According to the class category that has been created, for hate speech, class is worth = [1] [0], for non-hate speech = [0] [1], then the value is close to hate speech. The last sentence, "fpi dog rizieq syhihab dajjal" is a sentence that is categorized as hate speech. Session IV explained the system design of this paper.

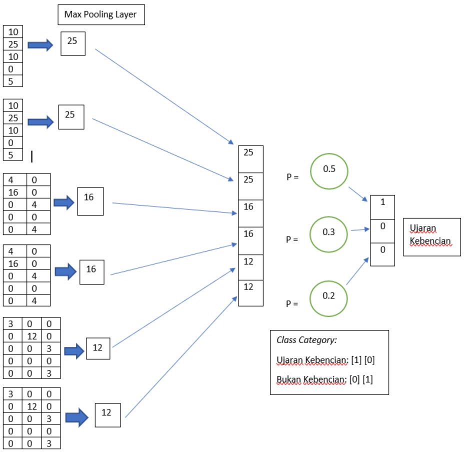


Fig 10. The Output of CNN

IV. SYSTEM DESIGN & OVERVIEW

A. System Overview

In this exploration, Convolutional Neural Network (CNN) algorithm will be used for the data mining classification. Input is as an image, and it will be processed by Optical Character Recognition (OCR) to make the output as text, and the CNN classification will process the textbook. The description of this examination is as follows:

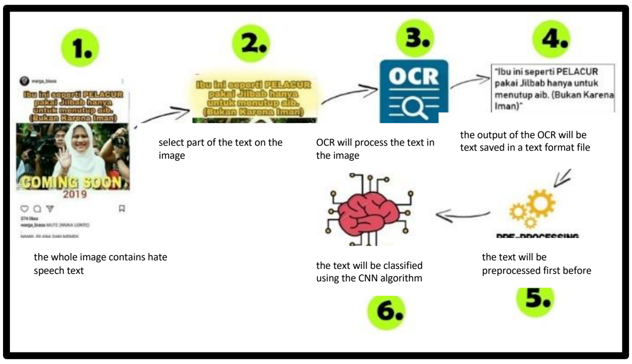
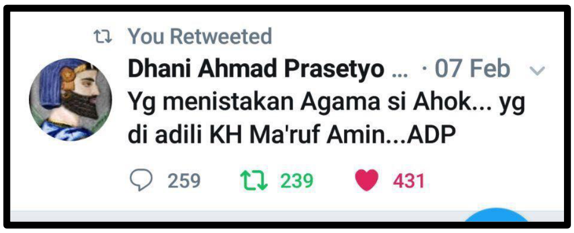

1.
2.
3.
4.
5.
6.
OCR
"Ibu ini seperti PELACUR pakai Jilbab hanya untuk menutup aib. (Bukan Karena Iman)"
COMING SOON 2019
select part of the text on the image
OCR will process the text in the image
the output of the OCR will be text saved in a text format file
the whole image contains hate speech text
the text will be classified using the CNN algorithm
the text will be preprocessed first before
PRE-PROCESSING

Fig 11. The Twitter data collection process

B. Data Twitter

Data were taken from different accounts that spread hate speech tweets in this research. The number of data that the West Java Provincial Language Center has verified is 1223 data. Fig.12 shows an example of training data that will be processed.



You Retweeted

Dhani Ahmad Prasetyo ... · 07 Feb

Yg menistakan Agama si Ahok... yg di adili KH Ma'ruf Amin...ADP

259 239 431

Fig 12. An example of a tweet that will be used for the data train

1. Initial Input

The first thing to do is select the image to be tested for recognition to figure out if there is any hate speech on the image; the initial input to this imaging system is text. The text on the image is expected to be legible because it will

Authorized licensed use limited to: University of Plymouth. Downloaded on February 19,2024 at 14:41:13 UTC from IEEE Xplore. Restrictions apply.

facilitate identification later. Fig. 13 shows the input image of the initial system.

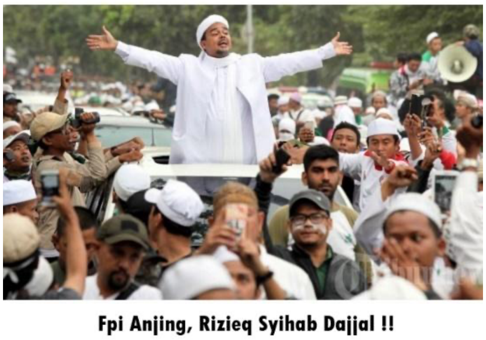


Fig 13. Example tweet account @TMCPoldaMetro (2)

2. Translating Image into Text

After getting the first input from the system, the next step is to convert the image into text or Optical Character Recognition (OCR). The process carried out at this stage is to select the image first by separating the image area (cropping) from the text's part. This is done to simplify the recognition of text in images so that the time required for text output from the Optical Character Recognition (OCR) reading process is not too long. Fig. 14 is an example of an image that has been separated from the text.

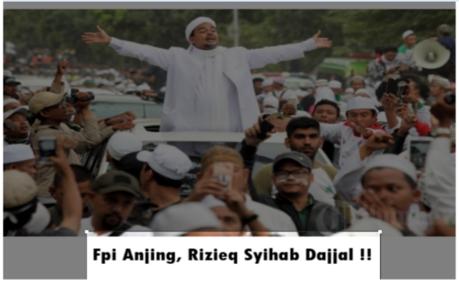


Fig 14. The image is separated from the text

After the image selection is made, the OCR process will be done. By the several phases described in Chapter II. Later on, the final result will get the full text stored in a text storage file of .txt.

3. Pre-Processing Data

The next step after the OCR outcome is complete is data preprocessing. Data preprocessing is an initial technique to arrange data to be processed. The processed data is in the form of username and tweet information. Pre-processing is carried out in 3 stages: case folding, cleaning, and stemming. The following is an explanation of each preprocessing step:

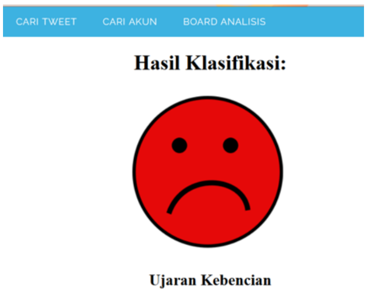
1. Case folding is the stage to change the input to lowercase and characters other than letters are omitted and are considered delimiters, including dots. This will simplify the tokenizing process.
2. Cleaning is used to remove unnecessary symbols in tweet posts. For example, "@, RT, #, https://etc" is done to make it easier for the system to recognize tweets that contain hate speech.
3. Stemming is making a retrieval system, and this process is carried out after filtering. The stemming process will keep the term in the filtering table as the root word and remove all affixes such as: put on, say, di-, i, pe, peng-, a-, etc.

Session V explained the results obtained from the test.

V. THE RESULT

A. *Visualization Convolutional Neural Network classifies the result.*

After calculating with the CNN algorithm, the classification results are obtained by checking on the label results whether the data has a value of (1), meaning that it contains hate speech, and (0) does not include hate speech. Below is a display of the results of the CNN classification. Fig 15 is the result obtained that the image contains hate speech.



Hasil Klasifikasi:

Ujaran Kebencian

Fig 15. The screenshot of the web application

B. *The Classification Using CNN Algorithm*

1. Data Preparation

There are 1482 tweets from preprocessing that will be carried out in the next step, namely performance testing, where the data is divided into two parts, namely training data and testing data. Finally, entropy and max gain that will be the foot of each attribute will be calculated by the C.45 algorithm.

2. Performance Testing

This test serves to determine the work of the CNN algorithm in classifying data into a predetermined space. Testing data will be used to test the decision table that has been made. Create a confusion matrix by calculating the test results' recall, precision, and accuracy.

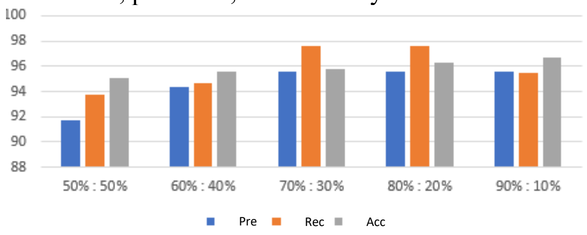


Fig 16. Precision results, recall, and data accuracy

In the fifth test with 90% training data and 10%, the dataset obtained Precision, Recall, and Accuracy values of 95.58%, 95.49%, and 96.71%, and an average accuracy value of 95.89%

3. Epoch & batch size testing

To obtain a higher accuracy value, perform testing with epoch and batch size parameters corresponding to the data. From the highest Precision, Recall and Accuracy values will be retested using epoch and batch size parameters to see any changes in accuracy values. It was found that Epoch 350 and

Authorized licensed use limited to: University of Plymouth. Downloaded on February 19,2024 at 14:41:13 UTC from IEEE Xplore. Restrictions apply.

batch size 512 get the highest accuracy value, equal to 98.6%, and the total of the 4 tests has an average value of accuracy of 95.5%.

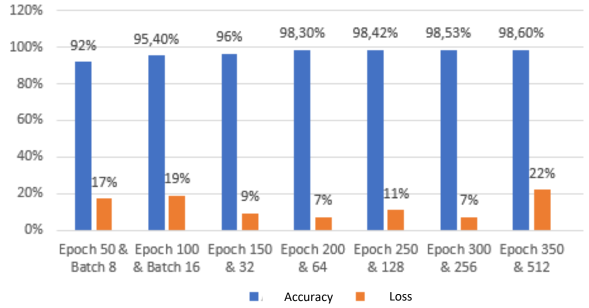


Fig 17. Data accuracy value

4. Testing Summary

After CNN classifies the images containing the tweet, the next step will be the process of classification test results. The confusion matrix technique will test the accuracy of the classification results by dividing the data partition. The accuracy test results obtained by the system with a data partition of 90%: 10% have the highest accuracy rate compared to others. This is because the more the amount of data training, the more the system recognizes the more vocabulary learned so that the system becomes more intelligent recognizes the classification. After testing the system by partitioning data, it was found that partition 90%: 10% has a high accuracy value worth 98.30% then; the second test is to change epoch and batch size parameters. After testing for four stages, it was found that in the fourth stage of epoch 350, batch size 512 reached the highest value. Session IV explains the conclusions and accuracy obtained from this system.

VI. CONCLUSION

This study concludes that there are two types of classes in the process of classifying hate speech, namely, hate speech or non-hate speech using the convolutional neural network (CNN) algorithm. The method of testing the partition data obtained 94.43% precision, 95.33% recall, and 95.89% accuracy. The test result of the epoch system and batch size got an average accuracy of 96.8%; further research can use other methods to make the system faster.

REFERENCES

[1] Doly, D. (2017). Pengaturan penyebaran ujaran kebencian dan isu sara ditinjau dari hukum konstitusi. Info Singkat Hukum, 9(17), 1-4.

[2] Indonesia, Kepolisian Negara Republik. (2015). Surat Edaran Nomor: SE.

[3] Kim, Y., Chiu, Y. I., Hanaki, K., Hegde, D., & Petrov, S. (2014). Temporal analysis of language through neural language models. arXiv preprint arXiv:1405.3515.

[4] Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820.

[5] Dos Santos, C., & Gatti, M. (2014, August). Deep convolutional neural networks for sentiment analysis of short texts. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers (pp. 69-78).

[6] Johnson, R., & Zhang, T. (2015). Semi-supervised convolutional neural networks for text categorization via region embedding. Advances in neural information processing systems, 28, 919.

[7] Kumar, S., Morstatter, F., & Liu, H. (2014). Twitter data analytics (pp. 1041-4347). New York: Springer.

[8] Schantz, H. F. (1982). History of OCR, optical character recognition. Recognition Technologies Users Association.

[9] Saksesi, A. S., Nasrun, M., & Setianingsih, C. (2018, December). Analysis Text of Hate Speech Detection Using Recurrent Neural Network. In 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC) (pp. 242-248). IEEE.

[10] Ashari, D. S., Irawan, B., & Setianingsih, C. (2021, October). Sentiment Analysis on Online Transportation Services Using Convolutional Neural Network Method. In 2021 8th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI) (pp. 335-340). IEEE.

[11] Ibrahim, A. F., Ristiawanto, S. P., Setianingsih, C., & Irawan, B. (2021, September). Micro-Expression Recognition Using VGG19 Convolutional Neural Network Architecture And Random Forest. In 2021 4th International Symposium on Agents, Multi-Agent Systems and Robotics (ISAMSR) (pp. 150-156). IEEE.

[12] Vieira, J. P. A., & Moura, R. S. (2017, September). An analysis of convolutional neural networks for sentence classification. In 2017 XLIII Latin American Computer Conference (CLEI) (pp. 1-5). IEEE.

[13] Banowati, C., Novianty, A., & Setianingsih, C. (2019, November). Cholesterol level detection based on iris recognition using convolutional neural network method. In 2019 IEEE Conference on Sustainable Utilization and Development in Engineering and Technologies (CSUDET) (pp. 116-121). IEEE.

[14] Pratama, R. M., Novianty, A., & Setianingsih, C. (2020). Cholesterol Detection Based on Eyelid Recognition Using Convolutional Neural Network Method. Proceeding of the Electrical Engineering Computer Science and Informatics, 7(2), 93-97.

[15] Sadewa, R. P., Irawan, B., & Setianingsih, C. (2019, December). Fire detection using image processing techniques with convolutional neural networks. In 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) (pp. 290-295). IEEE.

Authorized licensed use limited to: University of Plymouth. Downloaded on February 19,2024 at 14:41:13 UTC from IEEE Xplore. Restrictions apply.