Eduvest –Journal of Universal Studies
Volume 2 Number 8, August, 2022
p-ISSN 2775-3735-e-ISSN 2775-3727

COVID-19 SENTIMENT ANALYSIS USING CONVOLUTIONAL NEURAL NETWORK / RECCURENT NEURAL NETWORK METHOD

Ravensca Matatula [1] , Danny Manongga [2] , Hendry [3]
Master in Systems Information , Satya Wacana Christian University , Indonesia
Email: 972019017@student.uksw.edu, danny.manongga@uksw.edu, hendry@uksw.edu

A BSTRACT

Social media is a very important tool in this modern era , one of which is namely twitter. Twitter allows user for give opinion / opinion to various issues and topics hot / viral trending . Trending on twitter is so so fast in the process of spreading so that Twitter becomes a medium of information that often become a media issue conspiracy . Covid-19 is a moderate epidemic / disease _ experienced the whole world when this . Issues circulating in the population , they believe that Covid-19 is a real pandemic and a conspiracy , issue this make population confused differentiate Among second issue that . Because of that required a fast and accurate analysis _ for produce valid results , that Covid-19 a real thing _ or conspiracy seen from opinion population and corner views written on Twitter. CNN/RNN or combined from RNN(LSTM) and CNN methods are method used _ for classify opinion population about Covid-19 issues . Study this also done with compare is correct RNN/CNN accuracy same like deep RNN even more fast for in the process . Research results state that accuracy from combined RNN/CNN no different remote , even RNN/CNN in the process more fast than deep RNNs. Research results about opinion / opinion residents on twitter who believe about Covid-19 is conspiracy more low than residents who have confidence about Covid-19 is something the real thing . Percentage classification opinion / opinion from sentiment positive by 63.15% and opinion / opinion sentiment negative by 28.60%, this is results calculation use RNN/CNN method , with accuracy reached 58%. Accuracy from method used _ make Covid-19 issues that exist in the population no Becomes hoax news so population more alert against the ongoing Covid-19 pandemic happen .

KEYWORDS Twitter, Covid-19, CNN/RNN, RNN

How to cite:	Ravensca Matatula, Danny Manongga, Hendry. (2022). Covid-19 Sentiment Analysis Using Convolutional Neural Network / Reccurent Neural Network Method. Journal of Eduvest . *Vol 2*(8): 1.563-1.572
E-ISSN:	2775-3727
Published by:	https://greenpublisher.id/

INTRODUCTION

COVID-19 pandemic is an event of the spread of the *coronavirus disease* 2019, abbreviated as *COVID-19* around the world. This disease is caused by a new type of *coronavirus* named *SARS-CoV-2* (Gorbalenya et al., 2020). *SARS* and *MERS are disease variants of the* corona virus . As is known *, Covid-19* has been identified by researchers as a variant of the disease from the corona virus as well. The SARS-CoV-2 which causes this is a new type of corona virus . In 2003 the SARS-CoV corona virus caused the SARS Pandemic, and MERS- CoV caused an epidemic in 2012. Meanwhile, other epidemics and pandemics originating from the corona virus also occurred such as in 2009 *Swine Influenza* , 2004 and 2013 *Avian Influenza* , 2014 Ebola and 2020 *Covid-19 pandemic* (Pranita, 2020). There is something in this time of pandemic that spreads just as fast as a virus. It's a conspiracy theory. The human tendency is to believe in conspiracies. Conspiracy theories exist in every age. Spokespeople for each population and group have their own theories. Conspiracy theories regarding *Covid-19* have spread online which has had a considerable influence on the public's view of this. Poor information and conspiracy theories about *Covid-19* have led to incidents of *xenophobia* and racism against Chinese and other East Asian or Southeast Asian people (Wijaya, 2020). The spread of information about *Covid-19* has become a source of anxiety and panic for people around the world , especially in Indonesia. A lot of false information is spread through social media, this makes a lot of opinions appear about the accuracy or truth behind the spread of *Covid-19 information* . Accuracy of data and information is very much needed in monitoring and evaluating the results of the implementation. Decision making will also be better, if the data and information obtained are better. Through this research, the author makes a comparison between positive, negative, or neutral sentiments regarding the *Covid-19 pandemic* this is happening . Sentiment analysis is a computerized technology to analyze an opinion sentence in the form of a text whose way of working is to study and extract it like *text mining ,* so it can generate sentiment information (Boudad, Faizi, Oulad Haj Thami, & Chiheb, 2018). *Text mining* is one of the fields related to *Natural Processing language* . Sentiment analysis is carried out to see the tendency of a person to have a negative or positive opinion, even a neutral opinion that can be used as a decision supporter.

Ivanedra & Mustikasari's research (2019) with the title "Implementation of *Recurrent Neural Network Methods* in *Text* " *Summarization* With Abstract Techniques". In this study, a text summarizing system with an abstract method can be generated by calculating the weights repeatedly from the *Recurrent Neural Network* (*RNN*) systematics. Types *of Recurrent Neural Network* (*RNN)* which is used to cover the shortcomings in *RNN* which after sorting the memory but cannot save it and each word can focus more on context by adding an *Attention mechanism* is *LSTM* (*Long Short Term Memory*). This study compares the results of the summary generated by the system and the summary generated by humans by testing the performance of the system using *Precision , Recall* , and *F - Measure* . News article data with a total data of 4,515 articles is used as *a dataset* . The test is divided based on the data using *Stemming and Non - Stemming* (Ivanedra & Mustikasari, 2019)techniques .

Hikmawan, Pardamean and Khasanah's research (2020) on "Analysis of Public Sentiment Against Joko Widodo Against the Covid-19 Pandemic Using the Machine Method Learning "in this study the classification method used is *SVM* , *Naïve Bayes* and KNN. The results of this study, from the three methods used, the SVM method was chosen as the best in analyzing sentiment with an accuracy rate of 84.58%, precision 82.14% and recall 85.82% (Hikmawan, Pardamean, & Khasanah, 2020)

Research (Taufik, 2018) with the title "Comparison of *Text Mining Algorithms for Hotel* Review Classification ". This study aims to determine a stronger predictive method for classifying hotel reviews , using four different methods. The methods used in this study are the SVM method, Naïve Bayes , SVM-PSO and C45. By processing the results of the comparison test method *Support Algorithm Vector Machine* , *Naïve Bayes* , *Naive Optimization Bayes with Particle* feature selection *Swarm Optimization* and the C45 Algorithm can be concluded that the C45 algorithm is superior in predicting the classification of hotel reviews . By applying the C45 Algorithm to the classification of hotel *reviews* , it can help *online site visitors* who are looking for lodging or hotels in making more accurate and faster decisions without reviewing *comments* from previous visitors and can directly compare the facilities and prices that visitors want.

Sherstinsky's (2020) research entitled " *Fundamentals of Recurrent Neural Network (RNN) and Long Short -Term Memory (LSTM) Network* ". In this study, we present the basics of *RNN* and *LSTM networks* using a principled approach. Starting with the differential equations encountered in many branches of science and engineering, it is shown that the canonical formulation The *RNN* can be obtained by the sample delay differential equation which is used to model processes in physics, life sciences, and neural networks The main contribution in this research is the uniqueness of the pedagogical approach used to analyze the *LSTM system RNN* and Vanilla from a signal processing perspective, formal derivation of the *RNN* unrolling procedure , and thorough treatment using descriptive and meaningful notation, this is aimed at demystifying the underlying concept. Additionally, as an unexpected benefit of this analysis, two new extensions for the vanilla network were identified *LSTM :* convolutional non-causal input context window and external input gate . Then the equations for the *LSTM cells* with this extension are added together with the iterative projection layer (Sherstinsky, 2020).

Salim & Mayary's research (2020) with the title " Twitter User Sentiment Analysis Against Electronic Wallets Using the *Lexicon Method*" *Based* and *KNN* ". In this study, two methods were adopted in sentiment analysis, namely *Lexicon Based* and *KNN* . Specifically in this study the *Lexicon . method Based* using data of 949 *tweets* from three electronic wallets, namely OVO, GoPay , LinkAja . The final result of the calculation with *KNN* obtained *confusion matrix* for OVO the accuracy value is 86.91%, GoPay has 94.05% accuracy, and LinkAja has 76.31% accuracy (Salim & Mayary, 2020).

Previous studies that have been carried out have proven that sentiment analysis has been successfully applied to determine the sentiment of the population towards an issue that is *trending* in the discussion. Because of this, a study was conducted on the analysis of sentiments of the Indonesian population, regarding the problem of the *Covid-19 pandemic. Convolutional Neural Network (CNN) / Reccurent Neural Network (RNN) is the* sentiment classification method in this study.

This study uses the method of *Recurrent Neural Network (RNN)* for its sentiment classification. *Recurrent Neural Network (RNN)* based on *Long Short -Term Memory (LSTM)* is ideal for use in the aspect of text or sentiment classification in the case of a pandemic *Covid-19* . In addition , *LSTM* is better at analyzing the emotions of a long sentence and as a language model *LSTM* can be used for multiple sentiment classifications of emotional attributes of text (Li & Qian, 2016). The data for conducting sentiment analysis is taken from *twitter* using *tweet* in Indonesian obtained from *Twitter APIs* . *Recurrent Neural Network* (*RNN*) is the basis for feature extraction from a more detailed data *machine learning* . This *RNN* is a development of a *Neural Network* and its architecture has almost the same resemblance to *Multi Layer Perception* (*MLP*). *Deep* work process *Learning* has similarities with the human brain, because it can send information to each neuron (Hughes, Li, Kotoulas, & Suzumura, 2017). Based on this background, this research was conducted to find out how the population's sentiment towards the pandemic phenomenon is *Covid-19* using the *Convolutional Neural Network* (*CNN*) method combined or *hybrid* with the *Recurrent Neural Network* (*RNN*). The research objective of using the combination of the 2 methods above is to compare these methods with the *deep method RNN* on the level of accuracy and speed. The data used is the opinion of the population on social media *Twitter* . With these two combined *neural network* methods , it can produce fast and accurate data, as well as become useful information for the population .

RESEARCH METHOD

This sentiment analysis research was carried out in several stages which can be seen in Figure 3 .

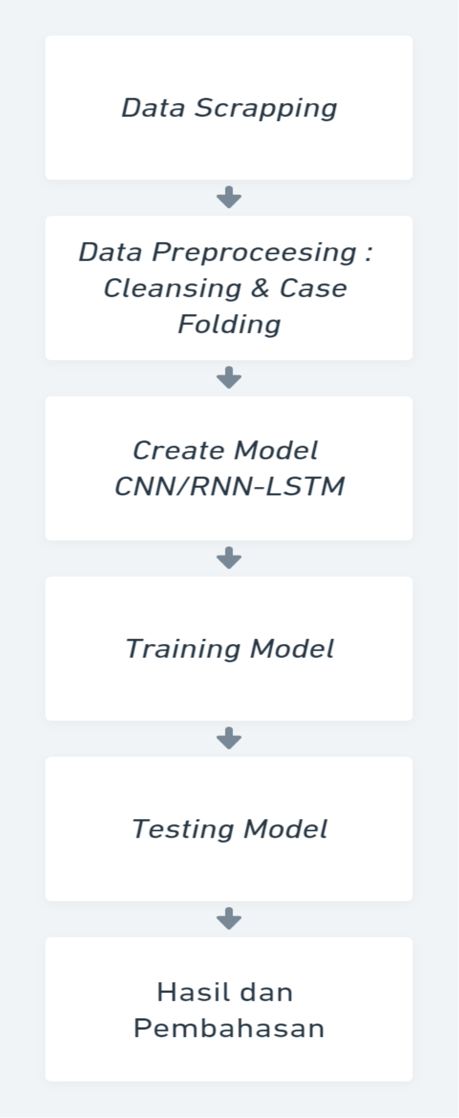
Data Scrapping
Data Preproceesing : Cleansing & Case Folding
Create Model CNN/RNN-LSTM
Training Model
Testing Model
Hasil dan Pembahasan

Figure 3. Research Stages

1) *Data Scrapping*

Process *data scrapping* conducted with retrieve *tweet* data use search with the keyword "#covid19, #coronavirus, #pandemicovid19, #konspirasicovid, #covidisreal". Data obtained use method *scrapping* with a total of 10,000 *tweets* , as for the data *range* is January 2020 to August 2020. Then the data is separated Becomes two big data in accordance with approach learning engine , to 7,000 *tweets* and 3,000 *tweets of test* data .

2) *Pre-processing Data*

Pre-processing Data applied to *training data* for make algorithm learning on the data . This process applied for more data processing easy with method get structured text _ from *data scraping* text that doesn't structured earlier . *Tweets* using _ language abbreviation or language area will interpreted to Indonesian so that the resulting data could synchronized . *Text pre-processing* conducted use a number of stages of research this , namely :

a) *Cleansing* : is for reduce *noise* with delete characters who don't including alphabet . Removed characters _ in the form of signs such as "@" and "#" mark the user and hashtag, *website url* and *emoticon , respectively* .

b) *Case Folding* : is the process of making all sentence or the word be letter small (*lower case*) in the sentence nor word.

c) *Tokenizing* : *Tokenizing* is the process of making *token* or *term* with separate words from sentence its constituents .

3) *Creating Models*

Create a model using method research used _ that is *CNN* / *RNN - LSTM* . *LSTM* is type module processing for *RNN* . *LSTM* created by *Hochreiter & Schmidhuber* , and later developed and popularized by many researchers . This is where the data is shared into 2 parts namely train data and test data.

In study this built a model with combine second method *CNN* and *RNN - LSTM* , here is the process of the model being built researcher :

***Input* :** in this process *input is done datasets* and data *pre-processing* . The process that helps organize the dataset by performing basic operations on the dataset before passing it on to the model such as removing spaces and meaningless words, converting various forms of words into their root words, and removing duplicate words, etc. is data *pre-processing* . It converts raw datasets into useful and organized datasets for further use.

***Word Embedding* : The preprocessed** *dataset* provides a unique and meaningful word order and each word has a unique *id* . This *Word Embedding* initializes words to assign random weights and learns embedding to embed all words in the training data set. This layer is used in various ways and is mostly used to learn the embedding of words that can be saved for use in other models.

***CNN* :** *Word Embedding* passes words in sentence form to a convolutional layer . *Convolution layers* convolute input using the *pooling layer* , *pooling layer helps reduce the representation of input* sentences , *input* parameters , computing in the network and control *overfitting* in the network. We apply *global max-pooling* at the end of the network layer, it gives the best *global result* of the whole network after applying different convolution layers (Yuliska, Qudsi, Lubis, Syaliman, & Najwa, 2021).

***LSTM* :** After *max-pooling is* forwarded to the *LSTM layer* **,** *the LSTM* uses three types of gates and cells to handle the flow of information across the network (Hermanto, Setyanto, & Luthfi, 2021).

- ***Dropout*** , as this prevents our model from *overfitting* . This drops irrelevant information from the network that does not contribute to further processing to improve the performance of our model.

- ***Dense Layer*** , the dense layer in the proposed model. It relates each input to each output using a weight.
- ***Sigmoid*** , a function that is widely used in the final layer of neural networks. It takes the average of the random results to be 0.1 shape.

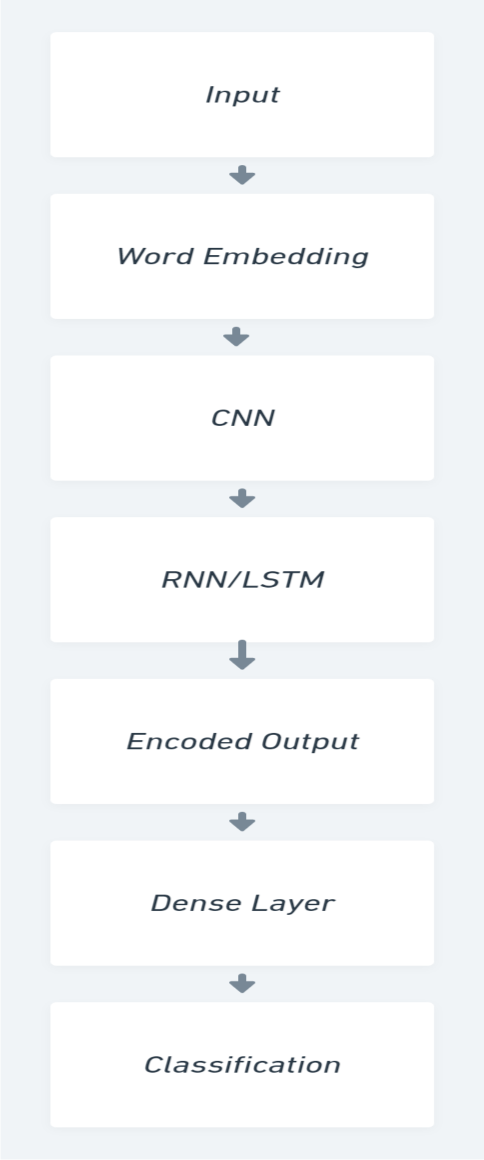

Figure 4. *CNN / RNN - LSTM model* .

RESULTS AND DISCUSSION

CNN / LSTM Testing

In *CNN / LSTM model training* is done with *epoch*. The *epoch* was determined by the researcher to get good accuracy. Before conducting *training and testing the model* , the data used must be *split* into training data and testing data. Split data *is* done with a comparison of 70% *training data* and 30% *test data* .

Results of *training* data using *epoch* 5 and *batch size* 32 resulted in *loss* : 0.9316, *accuracy* : 0.5845, *validation_loss* : 0.9287 and *validation_accuracy* : 0.5983, we can see that *loss* has decreased while *accuracy* has increased. *Loss* which is *training loss* is the calculation of the *loss function* of *training datasets* and predictions from the models

created. *Accuracy* is *training accuracy* is the calculation of the *accuracy* of the *training* data set and its prediction model. *Validation loss is the loss* calculation value *function* of *validation dataset* and prediction from model with *input* data from *validation the datasets*. *Validation accuracy* is the value of calculating accuracy from *validation dataset* and prediction from model with *input* data from *validation datasets* . While *Validation* itself is data or *datasets* that have never been seen or *trained* from the built model.
After seeing that the results of the *training* data showed good *accuracy* , then data *testing was carried out*. The results of *testing* the data using the model that has been implemented found a significant change, namely from the *real - positive number of* 3,835 the data predicted by the model to change to 6,315 data. On the other hand, *real – negative* automatically changes according to the predicted changes that occur in *real - positive* , from 5,340 data to 2,860 data.

Classification results sentiment using the *CNN/LSTM model* on against *Covid-19* in Indonesia uses opinion / opinion residents on *Twitter* . A total of *10,000* data that has been scrapped from opinion / opinion *twitter* show that opinion / opinion population for sentiment negative have number percentage by 28.60% who have confidence that *Covid-19* is a conspiracy and opinion / opinion population for sentiment positive have number percentage by 63.15% who have confidence that *Covid-19* is something the real thing .

Then the words that often appear in the results of positive and negative sentiments in this analysis are visualized in the form of a *wordcloud* . *Wordcloud* Negative sentiment was generated due to people's distrust of *Covid-19* on *tweets* with the word 'conspiracy' which was a word that had a high frequency in *tweets* during that period. While the words "*pandemic*", "*positive cases*", *and* "*real*" are words which often appears in some positive sentiment *tweets* . The *wordcloud results are* shown in Figure 5 and Figure 6.

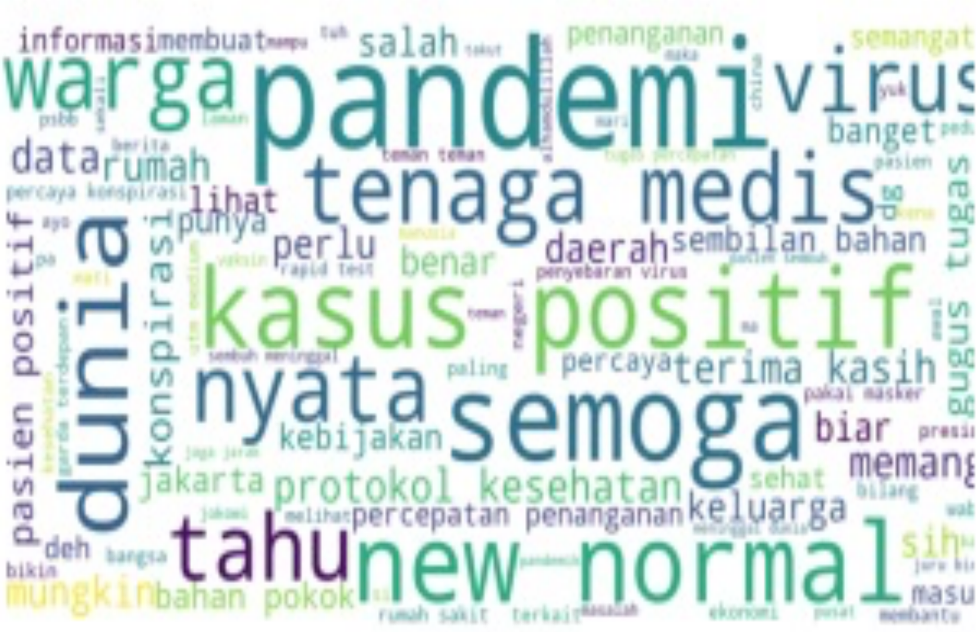

Fig 5. *Wordcloud* positive sentiment

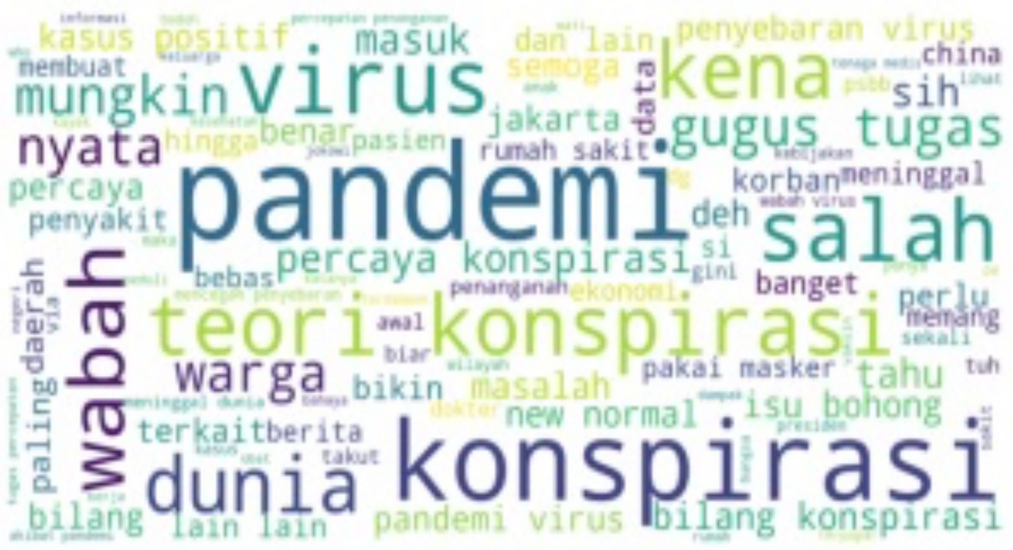

Fig 6. *Wordcloud* negative sentiment

***RNN* Testing (*LSTM*)**

The author tested the *RNN* (*LSTM*) as a comparison to the *CNN* / *LSTM model* . Tests are carried out using the same *data* , *epoch* and *split_data* .

results of the *training* data resulted in *loss* : 0.9054, *accuracy* : 0.5777, *validation_loss* : 0.8751 and *validation_accuracy* : 0.6325, we can see that *loss* has decreased while *accuracy* has increased.

In table I can be seen the comparison of the 2 tests carried out.

Model	*CNN/RNN-LSTM*	*RNN-LSTM*
loss	0.9316	0.9054
Accuracy	0.5845	0.5777
Val_Loss	0.9287	0.8751
Val_Accuracy	0.5983	0.6325
Average time	37.2s	158.8s

Table I. Table of model test results

CONCLUSION

Based on the test results of the two models above , which can be seen in the table. 1. In the table it can be seen that the *accuracy* of the *CNN* / *LSTM model is* not much different from that of the *RNN* - *LSTM* and also the time it takes to perform 1x *epoch* in the *CNN* / *LSTM model* is far superior to the *RNN* - *LSTM model* . It proves that the proposed model can do a good job in this regard.

From the sentiment analysis process that has been carried out with the *CNN* / *LSTM model* , the results of the study found that positive sentiment of 6,315 tweets with a percentage of 63.15% of the *twitter population* believed *Covid-19* was real, then the results of negative sentiment with a total of 2860 *tweets* believed that *Covid -19 19* is a conspiracy.

Negative sentiment is generated due to the population's distrust of *Covid-19* . Several *tweets* showed distrust with the word 'conspiracy' being the word that had high frequency in *tweets* during that period. While the words "*pandemic*", "*positive cases*", *and* "*real*" are words which often appears in some positive sentiment *tweets* that show the belief in *Covid-19* is real. Based on these results, it is concluded that the opinions of *Twitter residents* who believe in *Covid-19* are real, higher than those who believe that *Covid-19* is a conspiracy. It is hoped that further research can apply different methods to get more accurate results in sentiment analysis .

REFERENCES

Boudad , Naaima , Faizi, Rdouan , Oulad Haj Thami, Rachid , & Chiheb , Raddouane . (2018). Sentiment analysis in Arabic : A review of the literature . *Ain Shams Engineering Journal* , *9* (4), 2479–2490. https://doi.org/10.1016/j.asej.2017.04.007

Gorbalenya , Alexander E., Baker , Susan C., Baric , Ralph S., Groot , Raoul J. De , Gulyaeva , Anastasia A., Haagmans , Bart L., Lauber , Chris, & Leontovich , Andrey M. (2020) . The species and its viruses – a statement of the oronavirus study group . *Biorxiv (Cold Spring Harbor Laboratory)* , 1–15.

Hermanto, Dedi Tri, Setyanto, Arief, & Luthfi, Emha Taufiq. (2021). LSTM-CNN Algorithm for Binary Classification with Word2vec on Online Media. *Creative Information Technology Journal* , *8* (1). https://doi.org/10.24076/citec.2021v8i1.264

Hikmawan, Sisferi , Pardamean, Proverbs, & Khasanah, Siti Nur. (2020). Public Analysis Sentiment Against Joko Widodo on the Covid-19 outbreak using the Machine Method Learning . *Journal of Scientific Studies* , *20* (2). https://doi.org/10.31599/jki.v20i2.117

Hughes , Mark, Li, Irene, Kotoulas , Spyros , & Suzumura , Toyotaro . (2017). Medical Text Classification Using Convolutional Neural Networks . *Studies in Health Technology and Informatics* , *235* , 246–250. https://doi.org/10.3233/978-1-61499-753-5-246

Ivanedra , Kasyfi, & Mustikasari, Metty. (2019). Implementation of the Recurrent Neural Network Method on Text Summarization with Abstractive Techniques. *Journal of Information Technology And Computer Science* , *6* (4). https://doi.org/10.25126/jtiik.2019641067

Li, Dan, & Qian , Jiang . (2016). Text sentiment analysis based on long short -term memory . *2016 1st IEEE International Conference on Computer Communication and the Internet, ICCCI 2016* , 471–475. https://doi.org/10.1109/CCI.2016.7778967

Pranita, Ellyvon . (2020). Is it true that the Corona Virus Pandemic is a Conspiracy? This is Expert Explanation. Retrieved August 11, 2020, from Kompas website : https://www.kompas.com/sains/read/2020/08/11/124000823/bendakah-pandemi-virus-corona-dalam-konspirasi-ini-pencepatan-ahli? page=all

Salim, Siti Saidah, & Mayary , Joanna. (2020). ANALYSIS OF SENTIMENT OF TWITTER USERS ON ELECTRONIC WALLETS USING LEXICON BASED AND K – NEAREST NEIGHBOR METHODS. *Scientific Journal of Computer Informatics* , *25* (1). https://doi.org/10.35760/ik.2020.v25i1.2411

Sherstinsky , Alex. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short -Term Memory (LSTM) network . *Physics D: Nonlinear Phenomenon* , *404* . https://doi.org/10.1016/j.physd.2019.132306

Taufik, Andi. (2018). Text Mining Algorithm Comparison For Hotel Review Classification . *Journal of Computer Engineering* , *IV* (2). https://doi.org/10.31294/jtk.v4i2.3461

Wijaya, Diana. (2020). Victims of the COVID-19 conspiracy theory are not just the intelligence of their adherents. Retrieved June 13, 2020, from Tempo website : https://www.tempo.co/dw/2739/korban-teori-konspirasi-covid-19-not-only-kerdasan-penganutnya

Yuliska, Yuliska, Qudsi, Dini Hidayatul, Lubis, Juanda Hakim, Syaliman , Khairul General, & Najwa, Nina Fadilah. (2021). Sentiment Analysis on Student Suggestions Data on Departmental Performance in Higher Education Using Convolutional Neural Networks. *Journal of Information Technology And Computer Science* , *8* (5). https://doi.org/10.25126/jtiik.2021854842