Convolutional Neural Network Method in Determining Pfizer Vaccination Sentiment Analysis

Riza Adrianti Supono¹*, Hizkia Abiel Muljana²
Master of Information Systems Management, Universitas Gunadarma, Depok, Indonesia¹
Information Systems, Faculty of Computer Science and Industrial Technology, Universitas Gunadarma, Depok, Indonesia²
Email: riza.adrianti.s@gmail.com¹*, hizkia@gmail.com²

ABSTRACT
Coronavirus (COVID-19) is a disease caused by the SARS-CoV-2 virus by attacking the respiratory system in humans and because of the rapid spread of infection, WHO declared COVID-19 a pandemic. Over time, several types of vaccines have been discovered which are thought to minimize the possibility of infection. One of the vaccines is Pfizer. During the use of the Pfizer vaccine, there have been pros and cons caused by the side effects of using the vaccine. Therefore, sentiment analysis was carried out on public opinion with data sourced from tweets on Twitter. The method used in making the model is Convolutional Neural Network (CNN). This model has been successfully created and has been tested on 1158 training data and 773 test data. The training data obtained an accuracy level of 98.87% and the test data obtained an accuracy level of 69.46%.

Keywords: Pfizer Vaccination, Sentiment Analysis, Convolutional Neural Network

INTRODUCTION
Coronavirus (COVID-19) is a disease caused by the SARS-CoV-2 virus by attacking the human respiratory system (Zhang et al., 2020). The first case occurred in Wuhan, China at the end of 2019 and then spread to all parts of the world, so that the World Health Organization (WHO) declared COVID-19 a pandemic (Chaurasiya et al., 2020). Because of this, various countries are trying to suppress the spread by closing access to several areas (Saadat et al., 2020). Of course, this has a big impact on the lifestyle of various groups of society (Wilkinson, 2020). Therefore, researchers and medical personnel are trying to overcome this virus by finding the right vaccine (Krubiner et al., 2021). So far several vaccines have been discovered that can help minimize the possibility of contracting COVID-19 (Sultana et al., 2020).

One of the vaccines that has been successfully created is Pfizer (Bernal et al., 2021). This vaccine was produced from collaboration between the BioTech, Fosun and Pfizer companies (Feix & Feix, 2021). The Pfizer vaccine has a high effectiveness value in stimulating the production of antibodies for COVID-19, namely 90% (Inchingolo et al., 2022). Quoting from covid.go.id, this vaccine has entered Indonesia in August 2021 and is being distributed in stages for the activity of administering dose 1 to those who have not had the vaccine and dose 2.

On January 11 2022, the government decided to carry out the 3rd dose of vaccine (booster) which was implemented on January 12 2022 by prioritizing elderly and
vulnerable groups who had received the 2nd vaccine more than 2 months ago. One of
the vaccines used to carry out Booster activities is Pfizer (Yehezkie & Ramatillah, 2023).

After using the vaccine, it causes several side effects, giving rise to pros and cons
in the opinion of the public. This opinion can be found on several social media available
in Indonesia. Social media has helped various people to communicate long distances,
communication can even be done without needing to know the person beforehand. Apart
from communicating between individuals, social media can be used to disseminate
information and public opinion. One of the social media commonly used by Indonesian
people is Twitter. Quoting writing done by Angeline Puput Giovani, et al; Twitter has
become popular in Indonesian society because of its simplicity and ease of use, and
users can freely express their views or opinions. To search for information, you can type
certain keywords to find the desired information. Apart from that, Twitter has become a
forum for accommodating various public opinions regarding certain topics. Therefore, to
make it easier to respond to certain topics, sentiment analysis can be carried out and
the results can be used as a consideration for public response.

Quoting from writing by Sukma Nindi Listyarini, sentiment analysis is a
computational study of individual attitudes, opinions and emotions towards an entity.
Entities can represent individuals, events, or topics. Algorithms commonly used when
conducting sentiment analysis in Indonesian are Naive Bayes, Maximum Entropy (ME),
Support Vector Machine (SVM), and Decision Tree. Meanwhile, research on sentiment
analysis in English has applied a deep learning method, namely Convolutional Neural
Network (CNN), which produces much better output than other algorithms, namely;
Precision 7%, Recall 8%, F-1 Score 9%.

Related research that has been conducted previously provides an important
foundation in understanding Convolutional Neural Network (CNN) methods. Azhar Eka
Mulia Wiguna et al. (2021) applied CNN to detect threat speech on Twitter social media
posts, with system accuracy results reaching 80.63 % . Meanwhile, Hans Juwiantho et
al. (2020) developed the Word2Vec model for Twitter sentiment analysis in Indonesian,
achieving an average accuracy of up to 76.40 % . Apart from that, research by Sukma
Nindi Listyarini and Dimas Aryo Anggoro produced the highest accuracy of 90% in
analyzing regional election activities using CNN. Kzar (2023) also used CNN for product
sentiment analysis, with the best model achieving an accuracy of 81.4 %. Finally,
Parameswari (2022) achieved the highest accuracy in environmental opinion analysis in
Depok City with a score of 86%.

The problem formulation in this research includes two main questions: first, how
to carry out sentiment analysis of public opinion regarding the Pfizer vaccine? Second,
what are the results of sentiment analysis using the CNN algorithm? This research has
several limitations, namely the use of tweets in Indonesian with the keyword Pfizer,
tweets analyzed from April 9 2022 to June 6 2022 with a total of 5,131 data, as well as
the classification of data into three categories: positive, neutral and negative. The aim
of this research is to carry out sentiment analysis of tweets using Indonesian with the
keyword Pfizer on Twitter, with a time span from April 9 2022 to June 6 2022, using a
total of 5,131 data. The tweet data will be classified into positive, neutral and negative,
and the results can be used as a consideration for the public in using the Pfizer vaccine.

RESEARCH METHOD

The methodology that will be carried out consists of several stages such as
system requirements analysis, data collection, data pre-processing, sentiment labeling,
data vector presentation, model training and testing and displaying sentiment results.
RESULTS AND DISCUSSION

This section will explain the stages carried out in the research to produce the previously planned output. The first stage is an analysis of the system requirements required during the research. The second stage, collecting tweets with the keyword "Pfizer" in Indonesian which will be used as a dataset. The third stage carries out data preprocessing such as case folding, cleaning text, normalization, stopword removal, stemming. The fourth stage is to label the dataset into "positive", "neutral", "negative". The fifth stage converts the data that has been labeled into a vector. The sixth stage creates a CNN model which will be used for the training data model, validation model, testing model to get a value for the accuracy of the model created. The seventh stage, create data visualization from the results of the model that has been created.

A. System Requirements Analysis

System requirements analysis is a stage that aims to determine the functional and non-functional requirements required.

1. Functional Requirements Analysis

Functional requirements in this research:

a. Preparation data results.

b. Classification results: positive, neutral and negative.

c. Results of the Convolutional Neural Network (CNN) method.

2. Analysis of Non-Functional Requirements

Non-functional needs in this research:

a. Software requirements:
   1) MacOS High Sierra version 10.13.6
   2) Google Chrome
   3) Google Collab
   4) Python
   5) Google Sheets

b. Hardware requirements:
   1) 2.5GHz Intel Core i5
   2) 4 GB 1600 MHz DDR3
   3) 500GB HDD storage

B. Data Collection

This research uses secondary data obtained by collecting tweets that contain the keyword "Pfizer" in Indonesian. This data was obtained by the crawling method using the tweepy and twint libraries. The data crawling process succeeded in collecting tweet data with a total of 5,131 data published on April 9 2022 - June 6 2022. Table 1 is some examples of tweet data that were successfully collected.

<table>
<thead>
<tr>
<th>No</th>
<th>Date</th>
<th>Username</th>
<th>Tweet</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2022-06-06</td>
<td>Ndhr</td>
<td>Anyway, you can check the Malaysian version of the complete vaccine definition here</td>
</tr>
<tr>
<td></td>
<td>11:47:01</td>
<td></td>
<td><a href="https://t.co/WDFimMvRbD">https://t.co/WDFimMvRbD</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>The point is, if you are 60+ and/or have 1-2 Sinovac or AstraZeneca vaccines, you must get a booster. If you're 60 or under, just get the Pfizer</td>
</tr>
</tbody>
</table>
### C. Data Pre-processing

This stage processes unstructured data into more structured data. This is needed to assist the data processing process at the next stage.

1. **Case Folding**

   Case Folding is the initial stage carried out when processing raw data into data that is ready to be used. This stage changes each letter character in the data to lower case.

2. **Cleaning Text**

   At this stage, clean the tweet data from URLs, punctuation, special characters and repeated spaces. How this stage works.

### Table 2. Cleaning Text

<table>
<thead>
<tr>
<th>No</th>
<th>Tweet</th>
<th>Tweet Clean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Anyway, you can check the Malaysian version of the complete vaccine definition here <a href="https://t.co/WDfimMvRbD">https://t.co/WDfimMvRbD</a></td>
<td>Basically, if you are 60+, and/or have 1-2 Sinovac or AstraZeneca vaccines, you need a booster. If you are 60 or under, just get the Pfizer or Moderna vaccine twice and consider being fully vaccinated. <a href="https://t.co/wdfimmvrbd">https://t.co/wdfimmvrbd</a></td>
</tr>
<tr>
<td>2</td>
<td>FDA Accepts Pfizer Application for Covid-19 Vaccine for Children Under 5 Years Read more information at <a href="https://t.co/ujudUa62kz">https://t.co/ujudUa62kz</a></td>
<td>FDA accepts Pfizer application for Covid-19 vaccine for children under 5 years old see complete information at <a href="https://t.co/ujudua62kz">https://t.co/ujudua62kz</a></td>
</tr>
<tr>
<td></td>
<td>#styleid #fda #vaksinacovid19 #vaksinpfizer #vaksinanak #health</td>
<td>#styleid #fda #vaksinacovid19 #vaksinpfizer #vaksinanak #health</td>
</tr>
</tbody>
</table>
3. Normalization
Standard words into standard ones. This normalization refers to the rules for writing in Indonesian.

Table 3. Normalization

<table>
<thead>
<tr>
<th>No</th>
<th>Tweet</th>
<th>Tweet_Clean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Anyway, the Malaysian version of the complete vaccine definition can be checked here. Basically, if you are old, and/or the vaccine is Sinovac or AstraZeneca, you must get a booster. If the vaccine is lower than Pfizer or Moderna, I'd consider being fully vaccinated.</td>
<td>Anyway, the Malaysian version of the complete vaccine definition can be checked here. The point is, if you are old, and/or the vaccine is Sinovac or AstraZeneca, you must get a booster. If the vaccine is lower than Pfizer or Moderna, consider being fully vaccinated.</td>
</tr>
<tr>
<td>2</td>
<td>fda accepts pfizer application for covid vaccine - children under 1 year old see more information at inasicovid inanak</td>
<td>fda accepts pfizer application for covid vaccine for children under 1 year old see complete information at inasicovid inanak</td>
</tr>
<tr>
<td>3</td>
<td>_cl: has anyone tried the main whole virus vaccine (sinovac, sinopharm), mRNA booster (pfizer, moderna), bo...</td>
<td>_cl: has anyone tried the main whole virus vaccine (sinovac, sinopharm), mRNA booster (pfizer, moderna), bo...</td>
</tr>
</tbody>
</table>

4. Stopword Removal
Stopword removal is a stage for removing words that are considered unimportant and do not affect analysis activities.

Table 4. Stopword Removal

<table>
<thead>
<tr>
<th>No</th>
<th>Tweet</th>
<th>Tweet_Clean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Anyway, the Malaysian version of the complete vaccine definition can be checked here. The point is, if you are old, the Sinovac AstraZeneca vaccine requires a booster. If the vaccine is lower than Pfizer or Moderna, consider being fully vaccinated.</td>
<td>Anyway, check the Malaysian version of the complete vaccine definition here. Yes, the point is that if you are old, the Sinovac AstraZeneca vaccine requires a booster. If the vaccine is lower than Pfizer or Moderna, consider fully vaccinated.</td>
</tr>
<tr>
<td>2</td>
<td>fda accepts pfizer application for covid vaccine for children under 1 year old see complete information at inasicovid inanak</td>
<td>FDA accepts Pfizer Covid vaccine application for under-year-olds. See complete information about Covid-19 in children</td>
</tr>
<tr>
<td>3</td>
<td>_cl: has anyone tried the main whole virus vaccine (sinovac, sinopharm), mRNA booster (pfizer, moderna), bo...</td>
<td>_cl: have you ever tried the main whole virus vaccine (sinovac, sinopharm), mRNA booster (pfizer, moderna), bo...</td>
</tr>
</tbody>
</table>

5. Stemming
Stemming is the process of changing a word back to its basic form. This can happen by deleting affixes at the beginning and end of a word. The stemming process can be seen in Table 5.
Table 5. Stemming

<table>
<thead>
<tr>
<th>No</th>
<th>Tweet</th>
<th>Tweet_Clean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Anyway, check the Malaysian version of the complete vaccine definition here. Yes, the point is that if you are old, the Sinovac Astrazeneca vaccine requires a booster. If the vaccine is lower than Pfizer or Moderna, consider fully vaccinated.</td>
<td>Anyway, the Malaysian version of the complete vaccine definition, check here. Yes, the point is, if the vaccine is old, Sinovac Astrazeneca requires a booster, if it’s lower than the Pfizer and Moderna vaccines, consider being fully vaccinated.</td>
</tr>
<tr>
<td>3</td>
<td>cl: have you ever tried the main whole virus vaccine (sinovac, sinopharm), mrna booster (pfizer, moderna), bo...</td>
<td>cl has tried the principal whole virus vaccine Sinovac Sinopharm booster mrna pfizer moderna bo</td>
</tr>
</tbody>
</table>

D. Labeling
Labeling in this research uses Textblob. Textblob plays a role in determining the value of a tweet. After obtaining this value, a label will be given with the value criteria -1, 0, 1. If the resulting value from the textblob is smaller than 0, then the tweet will have a negative value. If the value is greater than 0 it will be positive and it will be neutral if the value is equal to 0.

E. Research Dataset
After going through the data preprocessing process, the data set has a total of 1932 tweets with three labels, namely 797 positive data, 835 neutral data and 300 negative data.

1. Training Dataset
The training dataset used amounts to 60% of the dataset, namely 1,159 data.

2. Test Dataset
The test dataset used amounts to 40% of the dataset, namely 773 data.

F. Word Embedding
In this section, we use the Keras library, namely Tokenizer, to convert text into word index or binary vector form.

G. CNN Model Architecture
The CNN model architecture used is as follows:

Figure 1. CNN architecture
H. CNN Models

In this section, we will explain CNN model training which will be followed by model validation and CNN model testing.

1. CNN Model Training

Model training uses 60% of the dataset. This data is used to train a CNN model to determine sentiment classification in tweets. Next, use a confusion matrix consisting of a true positive, two false positives, a true negative, two false negatives, a true neutral and two false neutrals.

<table>
<thead>
<tr>
<th>Predictions</th>
<th>Negative</th>
<th>Neural</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Actual</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>True</td>
<td>False</td>
<td>False</td>
</tr>
<tr>
<td>Neural</td>
<td>False</td>
<td>True</td>
<td>False</td>
</tr>
<tr>
<td>Positive</td>
<td>False</td>
<td>False</td>
<td>True</td>
</tr>
</tbody>
</table>

Based on table 6, true negative shows data that is predicted correctly with negative sentiment, while false negative shows data that is predicted incorrectly with negative sentiment. True positive shows data that is predicted correctly with positive sentiment, while false positive shows data that is predicted incorrectly with positive sentiment. True neutral indicates data that is predicted correctly with neutral sentiment, while false neutral indicates data that is predicted incorrectly with neutral sentiment.

2. Model Testing

Model testing uses test data of 40% of the dataset. Testing data is not included in the data used for model training. This test will use a confusion matrix which contains true positive, false positive, true negative, false negative, true neutral and false neutral values.

I. Visualization

The final stage is designing a visualization of the results of sentiment analysis for Pfizer. This visualization aims to make it easier to understand the data that has been analyzed. Visualization in the form of a pie chart and wordcloud.

IMPLEMENTATION AND TRYING

A. System Requirements Analysis

Stage is carried out on all data that has been collected. Data preprocessing consists of case folding, text cleaning, normalization, stopword removal and stemming stages. The following is an attachment to the program carried out during the data preprocessing process.
Convolutional Neural Network Method in Determining Pfizer Vaccination Sentiment Analysis

```python
#buat fungsi untuk mengecilkan text
def lower(text):
    text = text.lower()
    return text

data_raw['Tweet_Clean'] = data_raw['Tweet'].apply(lower)
data_raw.head(5)
```

Figure 2. Case folding function code

```python
def cleanText(text):
    text = re.sub(r'@[A-Za-z0-9]+', '', text) #menghapus mention
    text = re.sub(r'\#[A-Za-z0-9]+', '', text) #menghapus hashtag
    text = re.sub(r'\^[\s\S]+', '', text) #menghapus AT
    text = re.sub(r'\d+', '', text) #menghapus angka
    text = re.sub(r'https?://[A-Za-z0-9./]+', '', text) #menghapus link
    text = re.sub(r'[\s]+', ' ', text) #menghapus spasi diawal dan akhir kalimat
    text = re.sub(r'\s\s+', ' ', text) #menghapus spasi dua buah
    text = re.sub(r'\s+', ' ', text) #menghapus spasi dua buah
    text = text.strip() #text = text.strip()
    #text = ' '.join(re.sub(r'@[A-Za-z0-9]+', '', text))
    #text = ' '.join(re.sub(r'\d+', '', text))
    #text = ' '.join(re.sub(r'\s\s+', '', text))
    return text

data_raw['Tweet_Clean'] = data_raw['Tweet_Clean'].apply(cleanText)
```

Figure 3. Cleaning Text Function Code

```python
dicts = {row[0]: row[1] for _, row in pd.read_csv('/content/drive/MyDrive/Sriyani_Sangat_FIX/Dataset/kbba/kbba.txt', delimiter=':').itertuples()}

def kbbi(text):
    token = text.split()
    final_string = ' '.join([dicts.get(word, word) for word in token])
    return final_string

data_raw['Tweet_Clean'] = data_raw['Tweet_Clean'].apply(kbbi)
```

Figure 4. Normalization Function Code

```python
#ini yang benar
def stopword2(text):
    factory = StopWordRemoverFactory()
    stopwords = factory.get_stop_words()
    textwords = text.split()
    text1 = [text for text in textwords if text not in stopwords]
    text = ' '.join(text1)
    return text

data_raw['Tweet_Clean'] = data_raw['Tweet_Clean'].apply(stopword2)
data_raw.head()
```

Figure 5. Stopword Removal function code
Figure 6. Stemming function code

```python
#1 jam 9 menit 14 detik
def stemming1(text):
    factory = StemmerFactory()
    stemmer = factory.create_stemmer()
    text = stemmer.stem(text)

    return text

data_raw['Tweet_Clean'] = data_raw['Tweet_Clean'].apply(stemming1)
```

B. Labeling

Labeling uses the Textblob library where Indonesian text will be translated into English to get the polarity value. Then this value will be used to determine the label for a text with the condition that if the value is greater than 0 it will be given a positive label, if the value is equal to 0 it will be given a neutral label and if the value is smaller than zero it will be given a negative label. Labeling results can be seen in table 7.

**Table 7. Tweet Labeling**

<table>
<thead>
<tr>
<th>No</th>
<th>Tweet Preprocessing Results</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pfizer booster vaccine is effective against omicron in children</td>
<td>Positive</td>
</tr>
<tr>
<td>2</td>
<td>information on pfizer booster vaccine magelang dongg</td>
<td>Neutral</td>
</tr>
<tr>
<td>3</td>
<td>the weak pfizer booster fever</td>
<td>Negative</td>
</tr>
</tbody>
</table>

C. Word Embedding

Word Embedding aims to convert tweet data into vector. The word embedding results can be seen in table 8.

**Table 8. Word Embedding Results**

<table>
<thead>
<tr>
<th>No</th>
<th>Tweet Preprocessing Results</th>
<th>Vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pfizer booster vaccine is effective against omicron in children</td>
<td>[2, 3, 1, 55, 113, 62, 25]</td>
</tr>
<tr>
<td>2</td>
<td>information on pfizer booster vaccine magelang dongg</td>
<td>[34, 2, 3, 1, 1140, 1086]</td>
</tr>
<tr>
<td>3</td>
<td>the weak pfizer booster fever</td>
<td>[193, 626, 3, 1, 29]</td>
</tr>
</tbody>
</table>

D. CNN Model Training Results

CNN model training was carried out on 60% of the entire dataset, namely 1,159 data. The highest accuracy value on the training data is 1.0 which is at the 100th epoch and the highest accuracy on the validation data is at the 25th epoch, namely 0.6947, so this research uses a model with a total of 25 epochs.

However, if you look at table 8, the loss value in the validation data is higher than the loss value in the training data. This is called overfitting where the model works better during training.
In Figure 4.6, the model created succeeded in predicting 501 of the 505 data labeled as actually negative, but the model considered that 4 data were neutral. Then 483 data were labeled neutral, the model succeeded in predicting 474 data correctly and 9 data labeled neutral were considered positive and the model succeeded in predicting 171 as data with a positive label.

E. CNN Model Test Results

Testing on the CNN model was carried out on 40% of the entire dataset, namely 773 data. This test was carried out using a number of epochs of 25. Figure 6 is a Confusion Matrix generated using test data.

In Figure 8, the model created succeeded in predicting 214 of the 261 data with actual negative labels, but the model considered 33 data to be neutral and 14 data to have a positive label. Then, of the 411 data labeled neutral, the model succeeded in predicting 263 data as neutral, but the model considered 102 data
labeled negative and 46 data labeled positive. In 101 positive data, the model actually succeeded in predicting 60 data labeled positive, but the model considered 18 data as negative and 23 data as neutral.

F. Visualization of Results
At this stage, a visualization of the results of the sentiment analysis that has been carried out will be displayed. The visualization will use a line plot to determine the total number of tweets made on each date, a pie chart to display the percentage of negative, neutral and positive labels and a wordcloud visualization to determine the highest volume of word usage.

CONCLUSION
Sentiment analysis towards the Pfizer vaccine using Convolutional Neural Network (CNN) has been successfully carried out using data in the form of tweets in Indonesian taken from Twitter. After the pre-processing process, the dataset size was reduced from 5,131 data to 1,932 data, which was then classified into three labels: positive, neutral, and negative. The highest percentage belonged to neutral labels (43.2%), followed by positive labels (41.3%) and negative labels (15.5%). The CNN model achieved training accuracy of 98.87% on 1,159 training data and testing accuracy of 69.46% on 773 test data. Suggestions for future research are to use a larger dataset and a wider time span to increase modeling accuracy.

REFERENCES


