BAREKENG: Journal of Mathematics and Its Applications
December 2023 Volume 17 Issue 4 Page 1901–1914
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss4pp1901-1914

A PRELIMINARY STUDY OF SENTIMENT ANALYSIS ON COVID-19 NEWS: LESSON LEARNED FROM DATA ACQUISITION, PRE-PROCESSING, AND DESCRIPTIVE ANALYTICS

Rahmatin Nur Amalia[1], Kusman Sadik[2*], Khairil Anwar Notodiputro[3]

[1,2,3]*Statistics and Data Science Department, Mathematics and Natural Sciences Faculty, IPB University*
St. Raya Dramaga, Bogor, 16680, West Java, Indonesia

*Corresponding author's e-mail: * **kusmans@apps.ipb.ac.id***

Article History:

Received: 17th May 2023
Revised:21st August 2023
Accepted: 13th September 2023

Keywords:

Sentiment Analysis;
Stratified Sampling;
Systematic Sampling;
Topic Modeling;
Web Scarping;
Word Representation.

ABSTRACT

Sentiment analysis is a method used to analyze opinions and feelings. The goal of sentiment analysis is to determine whether a document contains a positive or negative emotion. Along with the spread of COVID-19 cases, news related to COVID-19 has often become a trending topic in the mass media. Conducting sentiment analysis using all news becomes more challenging because it might take time and cost. Therefore, the sampling method is needed to obtain representative news for the analysis. Web scraping was employed to obtain the news article about COVID-19 in Indonesia. In order to select representative news, two-step sampling was employed using stratified and systematic random sampling. According to the topic modelling results using lambda 0.6, news articles are grouped into three topics: updating COVID-19 cases, vaccination, and government policy. In addition, based on the number of positive and negative words, news articles are grouped into news dominated by positive words, news dominated by negative words, and news with the same number of positive and negative words. Methods for representing text in numerical form have been developed. Some of them use tf-idf weighting and word embedding. It does not pay attention to word order or meaning, only based on the frequency of words both locally and globally. Furthermore, this method will form a vector size as large as the number of unique words in the document, so it is less effective when many documents are used. Meanwhile, the vector size generated from the word2vec method is not as much as the number of unique words in the corpus. In addition, word2vec considers the context of the words in the corpus.

How to cite this article:

R. N. Amalia, K. Sadik and K. A. Notodiputro., "A PRELIMINARY STUDY OF SENTIMENT ANALYSIS ON COVID-19 NEWS: LESSON LEARNED FROM DATA ACQUISITION, PRE-PROCESSING, AND DESCRIPTIVE ANALYTICS," *BAREKENG: J. Math. & App.*, vol. 17, iss. 4, pp. 1901-1914, December, 2023.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Back in March 2020, the President and the Minister of Health announced the first case of COVID-19 in Indonesia. Since then, the number of cases has been increasing every day. Along with the spread of COVID-19 cases, news related to COVID-19 has often become a trending topic in the mass media. As a result, information about COVID-19 circulated quickly on social media. In this digital age, it is a completely natural phenomenon. Undoubtedly, the news data dispersed across social media makes for an interesting data source for analysis. The news released by the mass media can influence public opinion, which, in turn, might influence their behavior.

Sentiment analysis, introduced in the early 2000s, is used to analyze opinions and feelings [1]. The goal of sentiment analysis is to determine whether a document contains a positive or negative emotion. Generally, sentiment analyses are grouped into three categories: statistical, lexicon-based approaches (knowledge-based methods), and hybrid approaches [2]. Pramana and Rachman applied sentiment analysis to Twitter data (*tweets*) regarding the COVID-19 vaccine using a lexicon-based approach. At the end of the research, it was stated that it was limited to using a positive-negative dictionary, so improvements are still needed in the analysis [3]. An approach that can be used to overcome this problem is employing a statistical approach for sentiment analysis [4]–[6].

Data acquisition is the first step in the data analysis process. It is the step of obtaining data from private sources, such as financial reports or public sources [7]. Of course, the data-collecting process cannot be done manually, given the vast amount of news data. Web scraping is therefore used in this research to get news information on COVID-19 in Indonesia. It is a technique for converting unstructured web data into structured data that can be stored and analyzed in a database or spreadsheet [8]. Because of these circumstances, collecting and mining such vast content has become vital and challenging [9]. Information from websites can be automatically retrieved using this method. This method can, therefore, perform the extraction procedure more rapidly and accurately than humans [10].

It can be known that due to the COVID-19 pandemic, news related to COVID-19 is released rapidly by the mass media. Consequently, in one year, there has been lots of news about COVID-19. Of course, conducting sentiment analysis using all news becomes more challenging because it might take time and cost. Therefore, the sampling method is needed to obtain representative news for the analysis. Therefore, the researcher would need to devise a scenario to select the sample from the population to get the representative data, called sample survey design [11]. The survey design should be as rigorous as possible to ensure minimum error and bias and to enhance maximum representativeness [12].

Although the majority of digital data is available as text, it is typically unstructured or semi-structed. [13]. As a result, organizing and cleaning textual data became necessary in order to make it relevant for analysis. However, given the volume of data, manual data pre-processing is essentially impossible [14]. This is one of the difficulties that people face in today's digital era. Data pre-processing is critical and must be completed to ensure data quality. If this process is not followed, data will be very inconsistent, resulting in undesirable analysis results [15].

On top of choosing the right modelling methods and representative news data, data availability is another challenge in performing sentiment analysis using a statistical machine learning approach. In general, available data tend to come without annotations [4]. Of course, this is a problem because sentiment modelling requires annotated training data, with its annotation as the response variable. Another problem arises in the labelling process. It might take a lot of time and costs, especially when large-sized data is used. As a result, pre-analysis is needed to find representative news for the labelling process. In this research, descriptive analysis and topic modelling are carried out as the basis for selecting news to be labelled.

Considering the importance of data acquisition and pre-processing in this era, therefore, this paper will discuss the preliminary steps of sentiment analysis, including data acquisition, data pre-processing, and data exploration. The result of this stage will be used for the next analysis process, namely news sentiment modelling.

2. RESEARCH METHODS

2.1 Data

This research used news articles data from news related to COVID-19 in Indonesia. The list of news portals is obtained from https://dewanpers.or.id/. Only articles that were published within the period of March 2020 – March 2021 are extracted.

2.2 Research Procedure

The research procedure is conducted in four stages: data acquisition, data cleansing, data preparation, descriptive analytics, and sentiment labeling. **Figure 1** shows the flow chart of the research procedure. Details of each stage are described below.

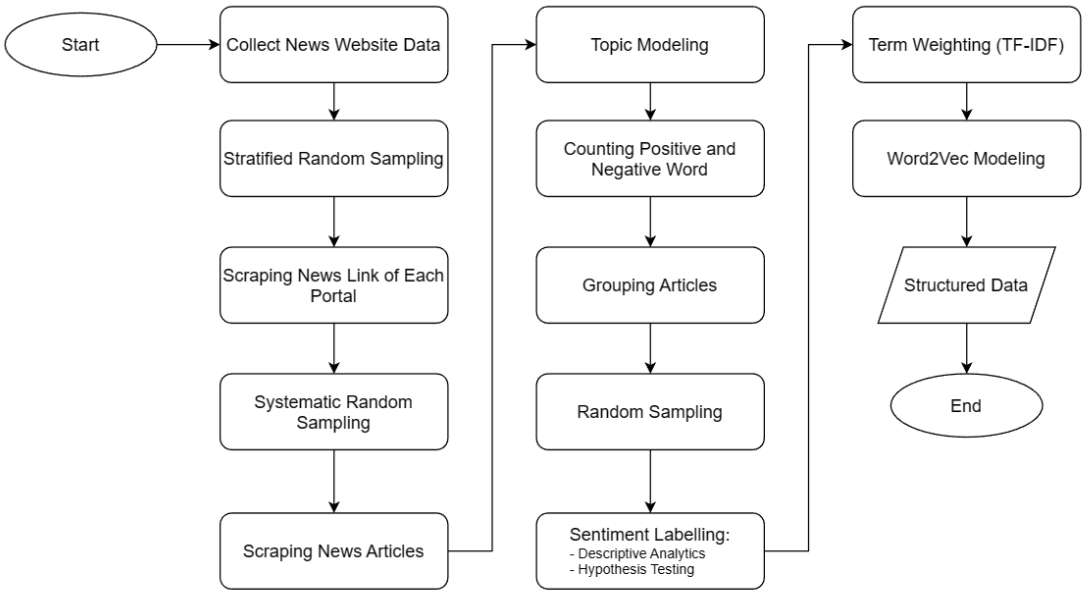


Figure 1. Research Procedure Flowchart

Data Acquisition

In the data acquisition stage, the web scrapping method is employed to get news data. The data obtained from this process are news articles, publication dates, and news titles. This stage is carried out in several steps, which are described below.

1. Determining news portals

   According to the https://dewanpers.or.id/ page, 557 official news portals were spread across 34 provinces in Indonesia (accessed on May 24, 2021). The news portals used in this research are determined using stratified random sampling according to the follower numbers on Twitter, as follows: > 500000 followers, 100000 – 500000 followers, 10000 – 100000 followers, 1000 – 10000 followers, and < 1000 followers. The sample size is determined using the following equations.

$$n = \frac{\sum_{k=1}^{l} N_k^2 p_k(1-p_k)/a_k}{N^2(B^2/4) + \sum_{k=1}^{L} N_k p_k(1-p_k)} \quad (1)$$

$$n_i = n\left(\frac{N_i\sqrt{p_i(1-p_i)}}{\sum_{k=1}^{L} N_k\sqrt{p_i(1-p_i)}}\right) \quad (2)$$

N : Population size
N_i : Population size of stratum i
p_i : Population proportion of stratum i
a_i : Fraction of observations allocated to stratum i
B : Bound of error

2. Extracting the news links of each news portal

Once the list of news portals is determined, web scraping is performed to obtain the news links. The web scraping process uses Python by utilizing BeautifulSoup and selenium packages. BeautifulSoup package is used for a static page of the website, while the selenium package is used for a dynamic page of the website. The keywords used are COVID-19 and Corona.

3. Sampling the news link data

After the link data is obtained, the next step is to apply systematic random sampling to the entire news link data. Systematic random sampling is a type of probability sampling method in which sample members from a larger population are selected according to a random starting point but with a fixed, periodic interval. This process is carried out at each stratum. The sample size is determined by the following formula.

$$n = \frac{Npq}{(N-1)\,(B^2/4) + pq} \tag{3}$$

N : Population total
p : Proportion of population
B : Bound of error

Data Cleansing

Data cleansing is the first stage of data preparation. This process removes the uninformative characters or common words inside the text. Online texts usually contain noise and uninformative parts such as URLs and advertisements. In addition, on the level of the word, many words in the text have little impact on the general orientation of it [16].

Data Preparation

At this stage, a set of operations are performed to prepare the data for sentiment modeling. The data preparation process is divided into several stages: n-gram analysis, term weighting (TF-IDF), and word embedding. The detail of each stage is as follows:

1. N-gram tokenization

A n-gram is a consecutive sequence of n words in a text [17]. This research used a combination of unigram and bigram tokenization.

2. Term weighting by using TF-IDF

TF-IDF weighting is a combination of local weighting (term frequency) and global weighting (inverse distance frequency) [1]. As a result, words frequently appearing in each document will have lower weights, while informative words will have greater weights. The formula to compute TF-IDF weights is as follows:

$$TF - IDF_{i,j} = TF_{ij} \times IDF_i \tag{4}$$

$$IDF_i = log_2\left(\frac{n}{df_i}\right) + 1 \tag{5}$$

$TF - IDF_{i,j}$: TF-IDF weight of i word in the jth document
TF_{ij} : The frequency of the occurrence i word in the jth document
IDF_i : Inverse document frequency of i word
df_i : The frequency of the occurrence i word in all documents
n : Total number of documents

3. Word embedding

In the word-embedding process, cleansed data is modelled using word2vec to create a vector that represents the meaning and context of the word [18]. This research used two different word2vec algorithms, namely CBOW (Continuous Bag of Words) and Skip-gram algorithms. Table 1 describes the scenario used in word2vec. These different dimensions are used in order to obtain the dimension size that maximizes the accuracy [19]. In the training process, hierarchical soft-max and negative sampling are applied in the optimization process.

Table 1. Scenario of Word2vec

Parameter	Value
Minimum word frequency	5.10
Learning rate	0.025
Vector size	100.300.500

Descriptive Analytics

Descriptive analytics will be conducted for news grouping to choose the news articles to label. Descriptive analytics will be carried out by modeling topics and positive-negative word counting. When the news articles are grouped, the next stage is sentiment labeling. In this research, three linguistic experts will do the labeling process.

1. Topic modelling (Latent Dirichlet Allocation)

Topic modeling is used for grouping news based on its word distribution. This process is carried out by using Latent Dirichlet Allocation (LDA) methods. The LDA model is a generative probabilistic model of the document collection or corpus [1]. The graphical illustration of the LDA model is shown in Figure 2. The number of topics will be defined by using the coherence value. The coherence score evaluates a unique topic by measuring the semantic similarity between important words of each topic.

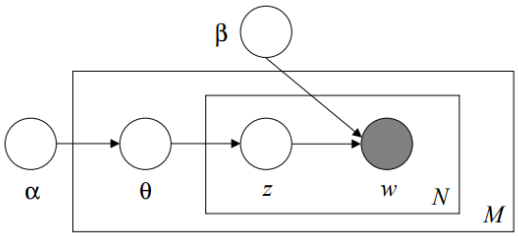


Figure 2. Graphical Illustration of LDA Model

2. Counting positive and negative word

This stage is conducted by recapping each piece of news' positive and negative words. A list of positive and negative words is obtained from a dictionary by Liu, which has been modified and translated into Indonesian.

Hypothesis Testing

Hypothesis testing will be carried out to determine if the proportions of sentiment categories are nearly the same for all populations. This stage is done by implementing the Chi-square (χ^2) test. The null hypothesis used in this test is:

$$H_0: p_{positive} = p_{negative} = p_{neutral} = 1/3$$

In addition, the test statistic will be carried out by using this formula.

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}} = \sum_{cell} \frac{(O-E)^2}{E} \tag{6}$$

where O and E symbolize an observed frequency and the corresponding expected frequency.

3. RESULTS AND DISCUSSION

3.1 Data Acquisition

Data collection is the starting point in conducting sentiment analysis in this research. According to https://dewanpers.or.id/, as of early March 2022, there were 290 factually and administratively verified news portals with accessible URLs. In this research, these news portals were stratified based on the number of Twitter followers. It was done under the assumption that the trending topic on social media may influence news sentiment.

As stated in the methodology section, news portals are grouped into five different strata based on the number of followers each has on Twitter. Using **Equation (1)**, the sample size of each stratum can be seen in **Table 2**.

Table 2. Sample Size of Each Stratum

Stratum	*N*	Sample size (n)
1st Stratum	18	4
2nd Stratum	19	5
3rd Stratum	42	11
4th Stratum	58	15
5th Stratum	162	40

A bound of error (BOE) of 0.1 was used to calculate the sample size, with the prior estimate of p being 0.5. As stated by **[11]**, calculating sample size using **Equation (1)** will yield its maximum value. Once the sample size was obtained, the web scraping process was then performed using Python. Only articles that were published within the period of March 2020 – March 2021 were extracted.

Web scraping was done on 75 news portals by using the packages BeautifulSoup and selenium. BeautifulSoup was used on websites with static pages or if the list of news links on a specific date can be accessed using a unique URL. For example, a list of news links from detiknews published on March 1, 2020, can be accessed by using the URL https://news.detik.com/indeks?date=03/01/2020. Accessing this URL will return all the news links published on March 1, 2020, in detiknews. Selenium was used to scrape websites with dynamic content, such as infinity scrolling and dynamic pages. The news portal Riau24.com, for instance, uses one single URL with a date filter feature to display links of news on a specific date as queried. Therefore, it was necessary to build a date filter automation process according to the required period.

Based on the scraping results from 75 news portals, 275 251 news URLs related to COVID-19 were obtained. Systematic random sampling was then run on the dataset and a sample of 5434 COVID-19 news URLs was obtained. **Table 3.** Sample Size of Each Stratumshows the number of samples in each stratum. Once the link data were obtained, a web scraping procedure was then performed once again to get articles and news titles of each instance. Since most of the news portals within the sample have static web pages, the BeautifulSoup package was applied to retrieve the news articles and dates.

Table 3. Sample Size of Each Stratum

Stratum	N	n
1st Stratum	87824	1099
2nd Stratum	38164	1080
3rd Stratum	67507	1094
4th Stratum	32060	1074
5th Stratum	49696	1087

3.2 Descriptive Analytics

Due to limited time and funds, labelling every article within the research was impossible. Therefore, it was necessary to determine the number of news articles to be labelled. Another challenge was choosing representative news articles to be labelled by annotators and used as training data for the modelling process. This process was necessary because it would affect the modelling result. Using representative training data, we hoped the resulting model could predict sentiment more precisely.

In order to obtain representative training data, topic modelling was carried out to group news articles based on their topic. This method assumes that each topic has a different word distribution and that the word distribution of news sharing the same topic will be similar. Once the topic groups were obtained, news articles within each group were selected to yield representative training data.

The number of topics was determined based on the coherence value. This value shows the pattern of interrelationships between sentences. A higher coherence value indicates the appropriate number of topics or better LDA performance in determining each document's topic.

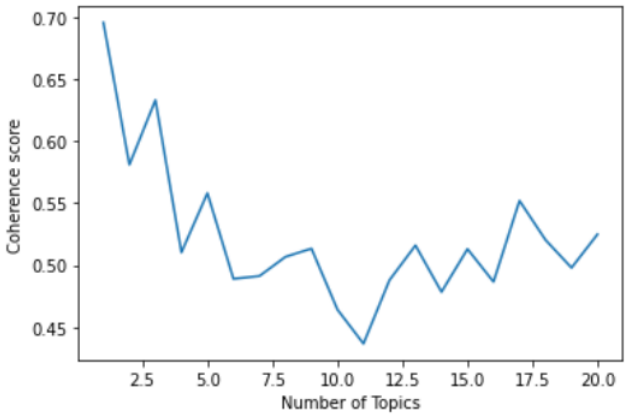
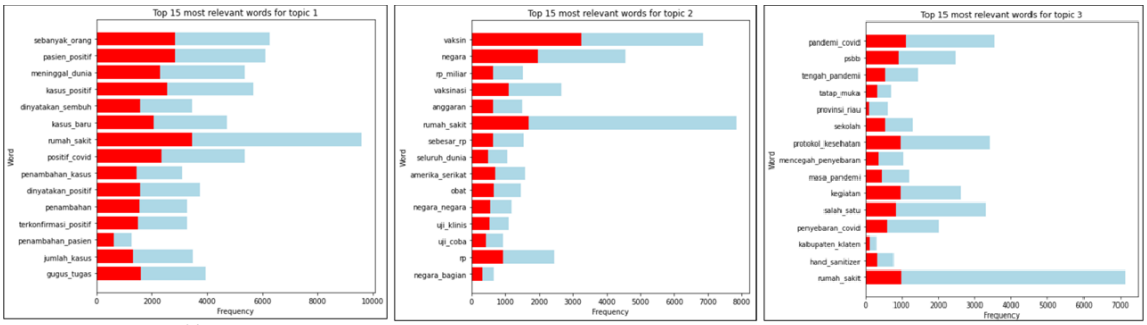


Figure 3. Coherence Plot

Coherence value is calculated by making pairwise comparisons between words on a particular topic, which results in a measure of the quality standard of a topic. **Figure 3** shows the number of topics and their respective coherence values. The highest coherence value is obtained from 3 topics. Therefore, this research grouped the news articles into three groups based on their word distribution.

LDA generates topics based on their word distribution. After the topics were defined, a descriptive analysis was then carried out to determine the word distribution within each topic. The descriptive analysis was carried out by using word relevance scores for each topic. Based on **[20]**, the optimal lambda for calculating the relevance score is 0.6.



(a) (b) (c)

Figure 4. Top 15 Most Relevant Words for Topic 1,2,3

Figure 4 (a) shows the top 15 most relevant words to describe Topic 1. The red bar chart represents the term frequency within topic 1, while the blue light bar chart represents the overall term frequency. Based on Figure 15, by using lambda of 0.6, the top three most relevant words to describe Topic 1 are sebanyak_orang, pasien_positif, and meninggal_dunia. These words are related to updates on COVID-19 cases, either updates regarding the total number of cases or deaths due to the COVID-19 virus. Thus, it could be inferred that Topic 1 is a group of news articles that shows updates regarding the total number of COVID-19 cases, deaths, and recoveries.

Figure 4 (b) shows that the words vaksin and vaksinasi are both included in the top 5 most relevant words of Topic 2. In addition, there are also the words anggaran, rp_miliyar, and rp, which may indicate the amount of budget spent by the government in tackling the COVID-19 case, especially budgets related to vaccination. Therefore, it could be inferred that news articles included in Topic 2 were related to the COVID-19 vaccination.

Figure 4 (c) shows that the words PSBB and protokol_kesehatan are within the top five most relevant words to describe Topic 3. Apart from these two words, the words tatap_muka and sekolah also contribute as the top 15 relevant words. These words are connected, as both tatap_muka and sekolah describe the government's policy in tackling COVID-19, especially regarding PSBB and health protocols.

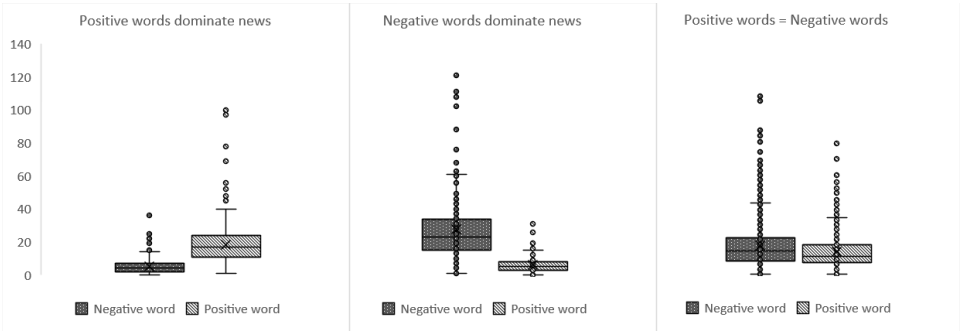


Figure 5. Boxplot of Positive and Negative Word Number of Each Group

In addition to topic modelling, this research also calculated each news article's positive and negative words. It was done with the assumption that the distribution of positive and negative words within each news article affects its perceived sentiment. Based on the number of positive and negative words, news articles were categorized into three groups; news dominated by positive words (Group 1), news dominated by negative words (Group 2), and news with the same number of positive and negative words (Group 3). **Figure 5.** Boxplot of Positive and Negative Word Number of Each Group shows the distribution of negative and positive word numbers in each group. The news dominated by negative words represents articles with more than 70% negative words, and the news dominated by positive words represents articles with more than 70% positive words.

3.3 Sentiment Labelling

Based on the results of the analyses above, news articles related to COVID-19 can be grouped into three separate groups based on the news topic. The first group are news articles discussing updates on the COVID-19 case, including updates on its total number of cases, deaths, and recoveries. The second group contains news articles that talk a lot about the COVID-19 vaccination, related to the budgeting plans or vaccine implementation in Indonesia. The third group contains news articles that discuss government policies mostly related to PSBB.

Furthermore, based on the distribution of positive and negative words, these news articles were also grouped into three groups: news dominated by positive words, news dominated by negative words, and news with the same number of positive and negative words. The results of the news grouping will be used as the basis for selecting news to be labelled.

Sentiment Labelling Scenario

In this research, labelling will be carried out on 1500 news articles. These articles were selected randomly for each topic and sentiment group. The distribution of the number of labelled news articles of each group can be seen in **Table 4.**

Table 4. News Sample Distribution Over Group

Topic	Group	N	Sample Size
1	1st Group	431	200
1	2nd Group	147	147
1	3rd Group	1200	153
2	1st Group	580	140
2	2nd Group	60	60
2	3rd Group	824	300
3	1st Group	367	150
3	2nd Group	114	114
3	3rd Group	1640	236

Sentiments are categorized into three classes which are positive, negative, and neutral. Positive statements, such as support for public policy, indicate positive sentiment. Negative sentiment is indicated by statements containing critics and complaints regarding COVID-19 was handled. Neutral sentiment describes informative news. A linguistic expert did the labelling process. Three different people repeated This process three times, and the sentiment was then determined based on the voting results of the three annotators. If all

three annotators labelled an article differently, a different annotator would re-label it. **Figure 6** shows the illustration of the sentiment labelling process.

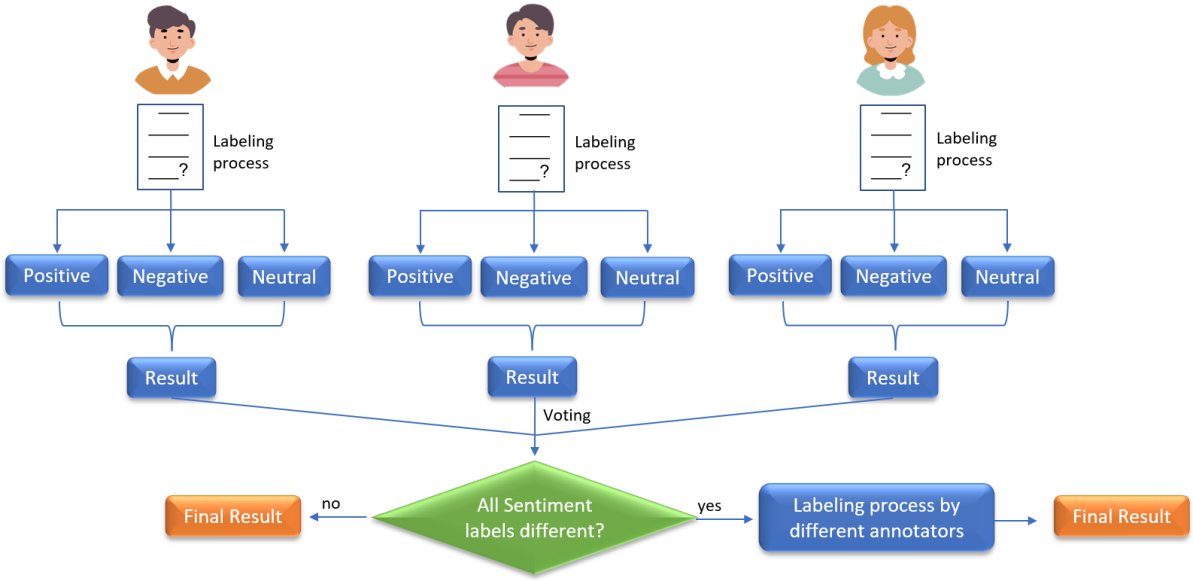


Figure 6. Illustration of Sentiment Labelling Process

Sentiment Labelling Results

The frequency of each news category from 1500 news used for modeling is shown in Table 5 below. According to **Table 5**, news sentiment tends to be neutral, with a frequency of 673 out of 1500 news articles. Meanwhile, there are 480 news articles with positive sentiment, which means the frequency between positive and neutral sentiment is not significantly different.

Table 5. News Sentiment Distribution

Sentiment	Frequency
Positive	480
Negative	347
Neutral	673

Furthermore, it is assumed that news sentiment about COVID-19 in Indonesia varies, meaning that the proportions between positive, negative, and neutral are the same. In that case, the proportion value for each category is $1/3$. Therefore, hypothesis testing was carried out using the χ^2 test to determine the dominance of news sentiment, and the results were obtained in **Table 6**.

Table 6. Result of Chi-square χ^2 Test

χ^2	P-value
53.561	0.000

Based on the hypothesis testing results above, the obtained p-value is $0.000 < \alpha = 0.05$, leading to the decision to reject the null hypothesis (H_0). It can be concluded that the proportions of each sentiment category (positive, negative, and neutral) in COVID-19 news are not equal. Looking at their frequency distribution, news sentiment about COVID-19 in the March 2020- March 2021 period tends to be neutral. It shows that most news portals only provide information, which can be in the form of updates on COVID-19 cases or information related to government policies.

3.4 Data Preparation

The difficulty in the modern digital era is not how to collect data but how to clean and prepare data to be valid for analysis. Pre-processing data is thus a crucial step before undertaking analysis, especially in text analytics. Online texts tend to contain lots of noise and uninformative parts such as HTML tags, scripts, and advertisements. In addition, on the word level, many words in the text do not have an impact on the general orientation of it **[21]**. Keeping those words makes the dimensionality of the problem high and, hence, the

classification more difficult since each word in the text is treated as one dimension. Here is the hypothesis of having the data properly pre-processed: reducing the noise in the text should help improve the performance of the classifier and speed up the classification process, thus aiding in real-time sentiment analysis [21].

Converting text into lowercase

In this step, the text was converted into the same case: the lowercase. This process is necessary because the program may give different meanings for the same word otherwise. As an example, the computer might consider the strings "Mendukung" and "mendukung" as two different words, when in fact they are referring to the same word. This pre-processing step is helpful when the dataset has mixed-case word collection. Table 7 shows the results of case conversion.

Table 7. An Illustration of Case Conversion

Before	After
2014 -- Peneliti utama vaksin Anhui dari Universitas Padjadjaran (Unpad) Rodman Tarigan mengungkap keunggulan vaksin rekombinan dibandingkan jenis vaksin Corona berbasis virus yang dimatikan	2014 -- peneliti utama vaksin anhui dari universitas padjadjaran (unpad) rodman tarigan mengungkap keunggulan vaksin rekombinan dibandingkan jenis vaksin corona berbasis virus yang dimatikan

Removing punctuation from the text

In this step, all punctuation marks within the text were removed. The results of removing punctuation from the text can be seen in Table 8.

Table 8. An Illustration of Punctuation Marks Removal

Before	After
2014 -- peneliti utama vaksin anhui dari universitas padjadjaran (unpad) rodman tarigan mengungkap keunggulan vaksin rekombinan dibandingkan jenis vaksin corona berbasis virus yang dimatikan	2014 peneliti utama vaksin anhui dari universitas padjadjaran unpad rodman tarigan mengungkap keunggulan vaksin rekombinan dibandingkan jenis vaksin corona berbasis virus yang dimatikan

Removing Link or Number from the Text

Link and number are characters considered not to provide important information related to news sentiment. Therefore, all links or words in the text were removed in this research. The results of this process can be seen in Table 9.

Table 9. An Illustration of Link or Number Removal

Before	After
2014 peneliti utama vaksin anhui dari universitas padjadjaran unpad rodman tarigan mengungkap keunggulan vaksin rekombinan dibandingkan jenis vaksin corona berbasis virus yang dimatikan	peneliti utama vaksin anhui dari universitas padjadjaran unpad rodman tarigan mengungkap keunggulan vaksin rekombinan dibandingkan jenis vaksin corona berbasis virus yang dimatikan

Stop word removal

Table 10. An Illustration of Stop Words Removal

Before	After
peneliti utama vaksin anhui dari universitas padjadjaran unpad rodman tarigan mengungkap keunggulan vaksin rekombinan dibandingkan jenis vaksin corona berbasis virus yang dimatikan	peneliti utama vaksin anhui universitas padjadjaran mengungkap keunggulan vaksin rekombinan dibandingkan jenis vaksin corona berbasis virus dimatikan

Table 10 shows the illustration of stop word removal from the text. Stop words are words that are commonly used and carry less or no meaning. They were removed from the text as they do not add any value to the analysis.

Tokenization

Table 11. An Illustration of The Tokenization Process

Before	After
peneliti utama vaksin anhui universitas padjadjaran mengungkap keunggulan vaksin rekombinan dibandingkan jenis vaksin corona berbasis virus dimatikan inactivated virus virus dilemahkan attenuated virus	[peneliti, utama, vaksin, anhui, universitas, padjajaran, mengungkap, keunggulan, rekombinan, dibandingkan, jenis, …]

The tokenization process was carried out to build a list of words for each news article. In this research, tokenization is done using a combination of unigram and bigram: uni-bigram. The results can be seen in **Table 11**.

Term Weighting

After data cleansing, the next step is to transform the text documents into a compatible format for text analysis. As previously discussed, explanatory and response variables (sentiment categories) are needed to model news sentiment. Machine learning and deep learning can work well and maximally if the data being processed is in the form of numeric data.

The document term matrix (DTM) is used to represent the frequencies of terms in documents. Meanwhile, the term frequency calculation is carried out using TF-IDF term weighting. TF-IDF weighting is a combination of local weighting (term frequency) and global weighting (inverse distance frequency) **[1]**. As a result, words frequently appearing in each document will have lower weights, while informative words will have greater weights. The results of the term frequency calculations are represented as a DTM, which will later be used as input variables for sentiment modelling. The illustration of the TF-IDF results can be seen in **Table 12**.

Table 12. An Illustration of the TF-IDF Results

Doc	Adanya	Arus	Badan	Bagus	Bank	...	Betul	Bijak
1	0.020	0.052	0.025	0.043	0.335	...	0.086	0.057
2	0.000	0.000	0.000	0.000	0.000	...	0.000	0.000
3	0.000	0.000	0.000	0.000	0.000	...	0.000	0.000
4	0.000	0.000	0.000	0.000	0.000	...	0.000	0.000
5	0.020	0.000	0.000	0.000	0.000	...	0.000	0.000
6	0.000	0.000	0.000	0.000	0.000	...	0.000	0.000
...	...	...	...	...	...	...	...	...
1500	0.000	0.000	0.000	0.000	0.000	...	0.000	0.000

Sentiment modelling using the TF-IDF weighting results as an input variable has advantages and disadvantages. The advantage of using this method is that it is easy to implement and understand in theory and meaning. This method produces the vector size as much as the number of unique words in the corpus, so it becomes inefficient for very large documents.

Word2Vec Modeling

Besides TF-IDF weighting, this research also used word2vec to represent text as a vector of numbers. This research used two-word embedding algorithms, namely CBOW and Skip-Gram. The layer sizes used are 100, 300, and 500. This layer size describes the vector length formed from each word. The layer size of 100 implies that 100 explanatory variables will be used in the model. Meanwhile, the vector representation of each article will be determined based on the average of each word vector in the news article. **Table 13** shows the results of word embedding with a layer size of 100.

Table 13. An Illustration of The Word2vec Results

doc	X1	X2	X3	X4	X5	X6	…	X100
1	0.24	-0.26	0.29	0.29	0.06	-0.80	…	0.27
2	-0.18	0.13	0.04	-0.02	-0.03	-0.05	…	0.15
3	-0.11	-0.08	0.18	0.11	-0.01	-0.12	…	0.11
4	0.09	0.14	0.07	-0.07	0.12	-0.30	…	0.08
5	-0.03	-0.12	-0.03	0.12	0.16	-0.12	…	-0.08
6	0.08	-0.20	0.16	-0.04	0.03	-0.28	…	0.03
7	-0.01	-0.15	-0.10	0.13	0.11	-0.18	…	0.03
8	0.01	0.01	-0.14	-0.02	0.09	-0.20	…	0.10
…	…	…	…	…	…	…	…	…
1500	-0.09	-0.16	0.08	0.03	-0.02	0.04	…	0.09

Word representation using word2vec has several advantages compared to TF-IDF; the vector size generated from this method is not as much as the number of unique words in the corpus. The vector size can be selected according to the corpus size and the type of project. It is particularly beneficial for very large data. In addition, word2vec considers the context of the words in the corpus, so using fewer input variables can provide better modelling results. However, this method requires a longer computational time than TF-IDF weighting.

4. CONCLUSIONS

The abundant availability of data provides a challenge for researchers to be wiser in selecting data. Using all data can increase the cost of funds and data processing time, making the analysis process ineffective. Applying the sampling method to the data acquisition can streamline the analysis process with a measurable and tolerable error rate.

In the previous era, the biggest challenge was obtaining data; now, the challenge has shifted to preparing data to obtain valid data. Of course, the steps that must be taken in preparing the data depend on the type of data used. The steps that must be carried out in the data preparation stage are very complex in the case of text data. The first step for text data must be carried out is to clean the data. After cleaning the data, researchers need to equalize the data format and form word vectors. The final stage is representing a text in a numerical representation. Methods for representing text in numerical form have been developed. Some of them use TF-IDF weighting and word embedding (word2vec). In the process, TF-IDF does not pay attention to word order or meaning, only based on the frequency of words both locally and globally.

According to the results of the topic modelling, news articles related to COVID-19 can be grouped into three groups based on the news topic. The first group is news articles discussing updates on the COVID-19 case, including updates on its total number of cases, deaths, and recoveries. The second group contains news articles that talk a lot about the COVID-19 vaccination, related to the budgeting plans or vaccine implementation in Indonesia. The third group contains news articles that discuss government policies mostly related to PSBB. Based on the hypothesis testing, news sentiment about COVID-19 in the March 2020- March 2021 period tends to be neutral. It indicates that most news portals only provide information that can be updated on COVID-19 cases or information related to government policies.

REFERENCES

[1] M. Anandarajan, C. Hill, and T. Nolan, *Practical Text Analytics*, vol. 2. in Advances in Analytics and Data Science, vol. 2. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-319-95663-3.

[2] E. Cambria, "Affective Computing and Sentiment Analysis," *IEEE Intell Syst*, vol. 31, no. 2, pp. 102–107, Mar. 2016, doi: 10.1109/MIS.2016.31.

[3] F. F. Rachman and S. Pramana, "Analisis Sentimen Pro and Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter," 2020.

[4] M. I. Abidin, K. A. Notodiputro, and B. Sartono, "Improving Classification Model Performances using an Active Learning Method to Detect Hate Speech in Twitter," *Indonesian Journal of Statistics and Its Applications*, vol. 5, no. 1, pp. 26–38, Mar. 2021, doi: 10.29244/ijsa.v5i1p26-38.

[5] H. Raza, M. Faizan, A. Hamza, A. Mushtaq, and N. Akhtar, "Scientific Text Sentiment Analysis using Machine Learning Techniques," 2019. [Online]. Available: www.ijacsa.thesai.org
[6] O. Somantri and D. Apriliani, "Support Vector Machine Berbasis Feature Selection Untuk Sentiment Analysis Kepuasan Pelanggan Terhadap Pelayanan Warung and Restoran Kuliner Kota Tegal," *Jurnal Teknologi Informasi and Ilmu Komputer*, vol. 5, no. 5, p. 537, Oct. 2018, doi: 10.25126/jtiik.201855867.
[7] R. S. Chaulagain, S. Pandey, S. R. Basnet, and S. Shakya, "Cloud Based Web Scraping for Big Data Applications," in *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, IEEE, Nov. 2017, pp. 138–143. doi: 10.1109/SmartCloud.2017.28.
[8] S. de S. Sirisuriya, "A Comparative Study on Web Scraping," 2015.
[9] M. Bahrami, M. Singhal, and Z. Zhuang, "A cloud-based web crawler architecture," in *2015 18th International Conference on Intelligence in Next Generation Networks*, IEEE, 2015, pp. 216–223. doi: 10.1109/ICIN.2015.7073834.
[10] R. Lawson, *Web scraping with Python*. UK: Packt Publishing Ltd, 2015.
[11] R. L. Scheaffer, W. Mendenhall, R. L. Ott, and K. G. Gerow, *Elementary survey sampling*, Seventh edition. USA: Brooks/Cole, 2012.
[12] S. Tyrer and B. Heyman, "Sampling in epidemiological research: issues, hazards and pitfalls," *BJPsych Bull*, vol. 40, no. 2, pp. 57–60, Apr. 2016, doi: 10.1192/pb.bp.114.050203.
[13] G. Yang, M. A. Jan, A. U. Rehman, M. Babar, M. M. Aimal, and S. Verma, "Interoperability and Data Storage in Internet of Multimedia Things: Investigating Current Trends, Research Challenges and Future Directions," *IEEE Access*, vol. 8, pp. 124382–124401, 2020, doi: 10.1109/ACCESS.2020.3006036.
[14] V. Dogra *et al.*, "A Complete Process of Text Classification System Using State-of-the-Art NLP Models," *Comput Intell Neurosci*, vol. 2022, pp. 1–26, Jun. 2022, doi: 10.1155/2022/1883698.
[15] V. Kalra and R. Aggarwal, "Importance of Text Data Preprocessing & Implementation in RapidMiner," Jan. 2018, pp. 71–75. doi: 10.15439/2017KM46.
[16] E. Haddi, X. Liu, and Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis," *Procedia Comput Sci*, vol. 17, pp. 26–32, 2013, doi: 10.1016/j.procs.2013.05.005.
[17] M. Schonlau, N. Guenther, and I. Sucholutsky, "Text Mining with n-gram Variables," *The Stata Journal: Promoting communications on statistics and Stata*, vol. 17, no. 4, pp. 866–881, Dec. 2017, doi: 10.1177/1536867X1801700406.
[18] B. Jang, I. Kim, and J. W. Kim, "Word2vec convolutional neural networks for classification of news articles and tweets," *PLoS One*, vol. 14, no. 8, p. e0220976, Aug. 2019, doi: 10.1371/journal.pone.0220976.
[19] H. Juwiantho, I. Setiawan, J. Santoso, and H. Purnomo, "Sentiment Analysis Twitter Bahasa Indonesia Berbasis WORD2VEC Menggunakan Deep Convolutional Neural Network," *Jurnal Teknologi Informasi and Ilmu Komputer (JTIIK)*, vol. 7, no. 1, pp. 181–188, Feb. 2020.
[20] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 63–70. doi: 10.3115/v1/W14-3110.
[21] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," in *Procedia Computer Science*, Elsevier B.V., 2013, pp. 26–32. doi: 10.1016/j.procs.2013.05.005.