Implementation of TF-IDF Algorithm and K-mean Clustering Method to Predict Words or Topics on Twitter

Muhammad Darwis [1*)], Gatot Tri Pranoto [2], Yusuf Eka Wicaksana [3], Yaddarabullah [4]
[1]Program Studi Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur
[2]Program Studi Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur
[3]Program Studi Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur
[4]Program Studi Teknik Informatika, Fakultas Industri Kreatif dan Telematika, Universitas Trilogi
email: [1]darwis.mawardin@gmail.com, [2]gatot.tp@gmail.com, [3]ekayusuf.wicaksana@gmail.com,
[4]yaddarabullah@trilogi.ac.id

Abstract - The social media time line, especially Twitter, is still interesting to follow. Various tweets delivered by the public are very informative and varied. This information should be able to be used further by utilizing the topic of conversation trends at one time. In this paper, the authors cluster the tweet data with the TF-IDF algorithm and the K-Mean method using the python programming language. The results of the tweet data clustering show predictions or possible topics of conversation that are being widely discussed by netizens. Finally, the data can be used to make decisions that utilize community sentiment towards an event through social media like Twitter.

Keywords - data mining, clustering, K-mean Method, TF-IDF algorithm, twitter, prediction

I. INTRODUCTION

Social media, like Twitter, has been widely used and has become a new habit for people in this era. The public feels comfortable if they express their opinions or thoughts through social media. Even in critical conditions, people spend a lot of their time to send and search for information through social media, such as Twitter, Instagram and others [1] [2].

It is not surprising that many researchers analyze data and sentiment obtained through various social media [3] [4] [5] [6] [7] [8]. They use this data to predict various things based on information circulating on social media. In fact, these researchers often build algorithms like their own expert systems to utilize existing data on social media, such as Twitter and Instagram. They realize that there are many things that can be extracted and obtained from these data.

In the current situation of the Covid-19 pandemic, for example, various researchers have used the data and information circulating on Twitter to classify, cluster and study the existing conditions [9] [10]. In order to restore the situation and to learn about the novel Covid-19 virus, the researchers process the information on Twitter for various purposes, for example predicting the possible ways of spreading the Covid-19 virus, the prevention, and so on. This is certainly interesting, especially since the pandemic which began to spread at the end of 2019 is still ongoing and is predicted to continue in various parts of the world.

In this situation, social media users, especially in Indonesia, will certainly not be silent. They will continue to access social media and find out the current conditions. Furthermore, they will convey the current conditions regarding the spread of the virus. At least, Indonesia's social media timelines, especially Twitter, will continue to be flooded with topics of conversation about the corona virus for some time to come. This is because considering the number of positive patients with Covid-19 in Indonesia is still increasing, even though various efforts have been made by the government through the Ministry of Health and BNPB.

In mid of April 2020, the talk about Covid-19 in Indonesia increased, which coincided with one month since the announcement of the first covid-19 patient. In addition, many Indonesians were also worried about the pandemic, especially because Ramadan would begin. Various other conditions have increasingly encouraged Indonesians to seek and convey information about covid-19 through social media like Twitter.

To see and analyze these conditions, the authors use data mining to cluster the topics and tweets that were mostly conveyed by the public at that time. Just as other researchers have done, the authors will use the TF-IDF algorithm and the K-mean clustering method to do this. With this method, a pattern will be generated that can be used to see and predict trends and topics of conversation among Twitter users regarding the covid-19 in Indonesia.

II. RESEARCH METHOD

The research methodology in the process of clustering through TF-IDF algorithm and K-mean method to predict trends and topics of conversation via twitter community includes several stages: from literature review, the determination of the data sets and data pre-processing, which consists of tokenization, stemming, removal of stop words and rejoin words as well as data exploration using the Python programming language. The description of the flow and stages of the research that the authors did for this study is as shown in Figure 1.

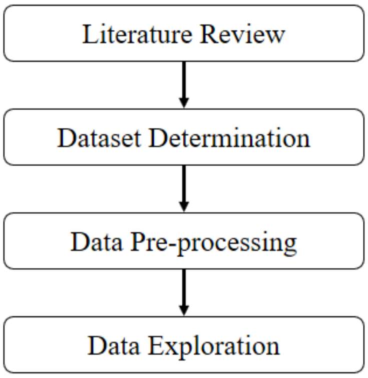


Figure 1. Flow and research stages

A. Literature Review

As literature to support this research, the authors use several trusted scientific journals, proceedings, e-books and websites. These writings include clustering data twitter, such as Cluster Analysis of Twitter Data: A Review of Algorithms [7], Clustering a Customer Base Using Twitter Data [5], Analysis and Visualization of Twitter Data using k-means Clustering [11] and Twitter data clustering and visualization [12]. In addition, the authors also use journals regarding data mining, especially those that discuss the problem of the covid-19, such as Characterizing the Propagation of Situational Information in Social Media During the COVID-19 Epidemic: A Case Study on Weibo [10], the Weakly-supervised Framework for COVID-19 Classification and Lesion Localization from Chest CT [9] and Twitter Sentiment Analysis of Covid-19 Using Term Weighting TF-IDF and Logistic Regression [13].

Based on the literature review, it is known that previous researchers have used Twitter data to perform data mining, especially clustering and sentiment analysis. The results and accuracy are very good and can be applied in solving problems. Therefore, some researchers also use data twitter for data mining to find out the Covid-19 pandemic and find solutions, such as applying the TF-IDF algorithm and regression. This study focuses on the TF-IDF algorithm and the K-Means method.

B. Dataset Determination

To solve the problems in this study, the authors started by collecting data sets. The writers obtained the dataset from Twitter with the keywords Corona Indonesia, Covid-19 Indonesia, the Ministry of Health, BNPB or patients on April 7 and 8, 2020. Furthermore, from the data tweet or document, the authors began to process it properly to the next stage: data pre-processing.

The number of datasets or tweets or documents that the authors sampled were 16 tweets or records data, which then were imported into python, which contained detailed information about the conversations of the Indonesian people about covid-19. To get this data, the authors make use of the twitter scraper module in python.

C. Data Pre-processing

The next stage is the writers take several steps in data pre-processing. This stage includes 4 main parts: tokenization, stemming, and removal of stop words and rejoin words. For information, to facilitate the implementation of these stages, the author uses `nltk` in python.

D. Data Exploration

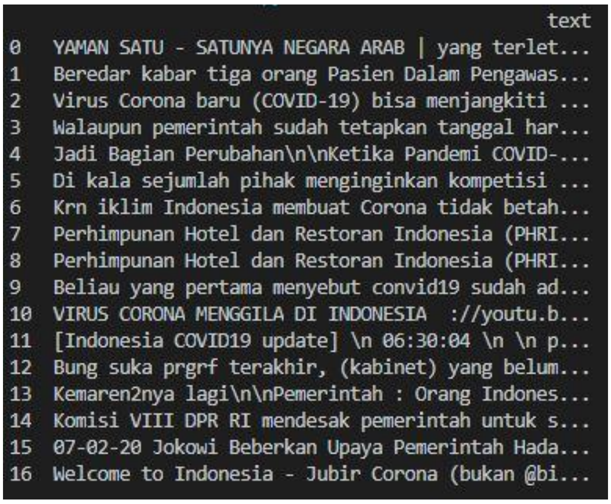
This stage is one of the important stages in the clustering process, because this is where the previously prepared data will actually be processed, its value is calculated and the results are analyzed. In this study, the authors performed data processing and calculation methods using several modules in python, such as NTLK, Literature, Pandas, Numpy, Sklearn and Matplotlib.

III. RESULTS AND DISCUSSION

The authors make a tweet data clustering design, starting from the collection of data training sets, the design stages and the application of data mining. Later, a certain pattern will be generated that can be used to see and predict trends and conversation topics about covid-19 on the date the data is taken. The research stages include identification of training sets, data pre-processing, data exploration using python and the result analysis.

A. Data Set

The first stage that the authors carry out is collecting and defining the set data as shown in Figure 2. In this data, it contains data tweets about Corona Indonesia, Covid-19 Indonesia, the Ministry of Health, BNPB or patients on April 7 and 8, 2020. Furthermore, to narrow data collection, the authors applied a limit of `limit = 10` and language `Lang= 'Indonesian'`. As a result, the authors got a document of 16 data tweets, which later in python would be converted into 16 list rows. We then exported the data into format `.xlsx`, which later would be imported into python for clustering. As seen in Figure 2, the data has a format that has not been proper or not the standard, so the data pre-processing would be carried out, to get the data ready to be processed.



```
                                                  text
0   YAMAN SATU - SATUNYA NEGARA ARAB | yang terlet...
1   Beredar kabar tiga orang Pasien Dalam Pengawas...
2   Virus Corona baru (COVID-19) bisa menjangkiti ...
3   Walaupun pemerintah sudah tetapkan tanggal har...
4   Jadi Bagian Perubahan\n\nKetika Pandemi COVID-...
5   Di kala sejumlah pihak menginginkan kompetisi ...
6   Krn iklim Indonesia membuat Corona tidak betah...
7   Perhimpunan Hotel dan Restoran Indonesia (PHRI...
8   Perhimpunan Hotel dan Restoran Indonesia (PHRI...
9   Beliau yang pertama menyebut convid19 sudah ad...
10  VIRUS CORONA MENGGILA DI INDONESIA  ://youtu.b...
11  [Indonesia COVID19 update] \n 06:30:04 \n \n p...
12  Bung suka prgrf terakhir, (kabinet) yang belum...
13  Kemaren2nya lagi\n\nPemerintah : Orang Indones...
14  Komisi VIII DPR RI mendesak pemerintah untuk s...
15  07-02-20 Jokowi Beberkan Upaya Pemerintah Hada...
16  Welcome to Indonesia - Jubir Corona (bukan @bi...
```

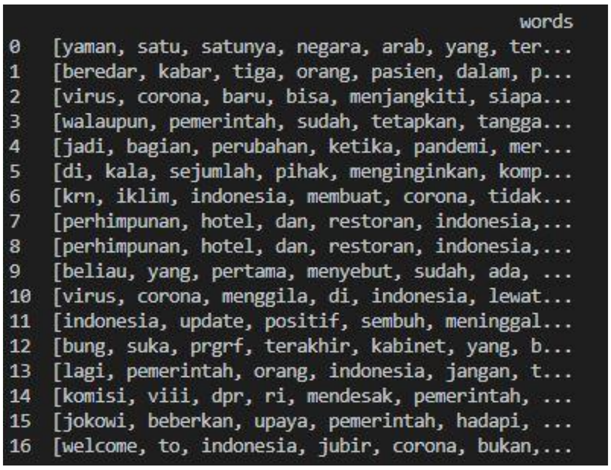
Figure 2. Dataset or document tweet

B. Data Pre-processing

This section contains the data preparation stage, which consists of tokenization, stemming and removal of and stop words rejoin words. The authors use the python programming language in all stages of this preparation to make the data processing easy and effective.

a. Tokenization

The first step that the writers took in data pre-processing process was tokenization, with the aim of selecting and identifying each of the words contained in the list tweet that the authors have collected. To do this, the authors used `word_tokenize()` method in the nltk module provided by python. As a result, we got word identification from each tweet or document as shown in Figure 3. In this picture, the data tweet in the form of a sentence is separated by a word with a comma (,), so that it can be identified easily. Not only that, in this process the authors also applied a rule so that python only fetched data in the form of those composed by alphabetic letters and not numbers using the `isalpha()` method. Thus, tweets that are numbers or dates or anything other than the alphabet will be ignored.

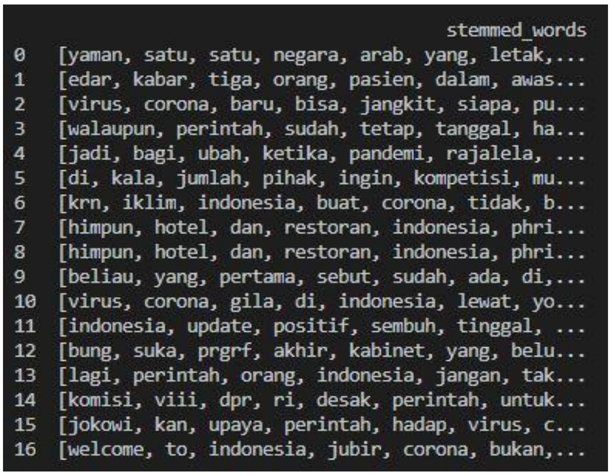


```
                                              words
0   [yaman, satu, satunya, negara, arab, yang, ter...
1   [beredar, kabar, tiga, orang, pasien, dalam, p...
2   [virus, corona, baru, bisa, menjangkiti, siapa...
3   [walaupun, pemerintah, sudah, tetapkan, tangga...
4   [jadi, bagian, perubahan, ketika, pandemi, mer...
5   [di, kala, sejumlah, pihak, menginginkan, komp...
6   [krn, iklim, indonesia, membuat, corona, tidak...
7   [perhimpunan, hotel, dan, restoran, indonesia,...
8   [perhimpunan, hotel, dan, restoran, indonesia,...
9   [beliau, yang, pertama, menyebut, sudah, ada, ...
10  [virus, corona, menggila, di, indonesia, lewat...
11  [indonesia, update, positif, sembuh, meninggal...
12  [bung, suka, prgrf, terakhir, kabinet, yang, b...
13  [lagi, pemerintah, orang, indonesia, jangan, t...
14  [komisi, viii, dpr, ri, mendesak, pemerintah, ...
15  [jokowi, beberkan, upaya, pemerintah, hadapi, ...
16  [welcome, to, indonesia, jubir, corona, bukan,...
```

Figure 3. Results of tweet data tokenization

b. Stemming

In the tokenization process, words are still not fully standard yet because there are additional words, such as affixes, prefixes or suffixes to certain words. Meanwhile, in order for the clustering process to be optimal, it would be good if what is processed is basic words. To overcome this, next the writers carried out a stemming process, where all additions, such as prefix or word endings were ignored. For this process, the authors used `PorterStemmer()` method, which was also included in the nltk python module. As a result, all words that contained affixes would be displayed in their basic word form as shown in Figure 4.



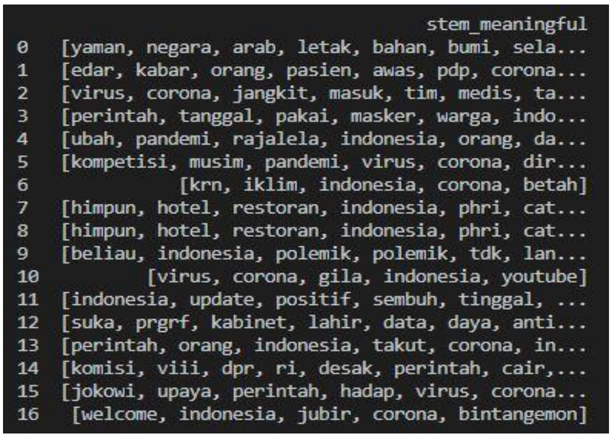
```
                                      stemmed_words
0   [yaman, satu, satu, negara, arab, yang, letak,...
1   [edar, kabar, tiga, orang, pasien, dalam, awas...
2   [virus, corona, baru, bisa, jangkit, siapa, pu...
3   [walaupun, perintah, sudah, tetap, tanggal, ha...
4   [jadi, bagi, ubah, ketika, pandemi, rajalela, ...
5   [di, kala, jumlah, pihak, ingin, kompetisi, mu...
6   [krn, iklim, indonesia, buat, corona, tidak, b...
7   [himpun, hotel, dan, restoran, indonesia, phri...
8   [himpun, hotel, dan, restoran, indonesia, phri...
9   [beliau, yang, pertama, sebut, sudah, ada, di,...
10  [virus, corona, gila, di, indonesia, lewat, yo...
11  [indonesia, update, positif, sembuh, tinggal, ...
12  [bung, suka, prgrf, akhir, kabinet, yang, belu...
13  [lagi, perintah, orang, indonesia, jangan, tak...
14  [komisi, viii, dpr, ri, desak, perintah, untuk...
15  [jokowi, kan, upaya, perintah, hadap, virus, c...
16  [welcome, to, indonesia, jubir, corona, bukan,...
```

Figure 4. Results of tweet data stemming

c. Removal of stop words

The next step that the authors do is the removal of stop words, namely the process of eliminating words that are commonly used and have no use in Natural Language Processing (NLP), such as ‘yang’, ‘di’ and so on. The output produced in the stemming process still left the ‘yang’, ‘di’ and so on words, so this stage needed to be done so that the clustering process could run well. In this stage, the authors utilized two methods to carry out the stopwords function, namely `stopword.words('english')` for the elimination of unimportant words in English and `stopword.words('indonesian')` in the NTLK module to eliminate the unnecessary Indonesian words.

The result is as shown in Figure 5, all the output words from the method are basic words which are all important, there are no conjunctions and so on. Until this stage, the available words are ready to be processed and explored further. However, the word list is still not yet in the form of a sentence.



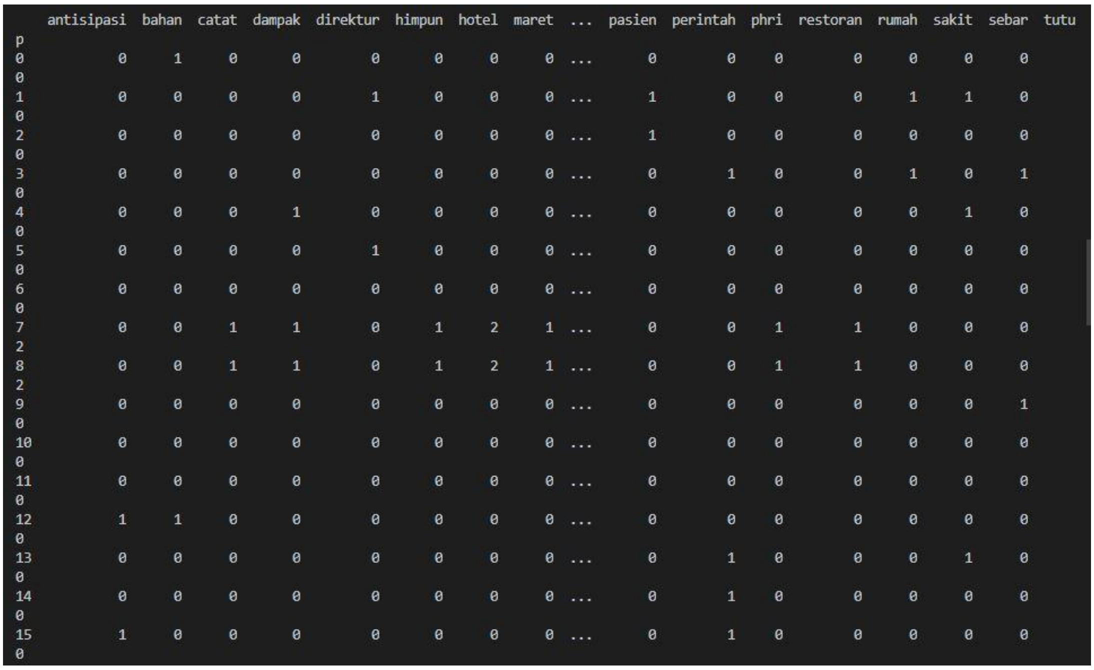
```
                                    stem_meaningful
0   [yaman, negara, arab, letak, bahan, bumi, sela...
1   [edar, kabar, orang, pasien, awas, pdp, corona...
2   [virus, corona, jangkit, masuk, tim, medis, ta...
3   [perintah, tanggal, pakai, masker, warga, indo...
4   [ubah, pandemi, rajalela, indonesia, orang, da...
5   [kompetisi, musim, pandemi, virus, corona, dir...
6             [krn, iklim, indonesia, corona, betah]
7   [himpun, hotel, restoran, indonesia, phri, cat...
8   [himpun, hotel, restoran, indonesia, phri, cat...
9   [beliau, indonesia, polemik, polemik, tdk, lan...
10         [virus, corona, gila, indonesia, youtube]
11  [indonesia, update, positif, sembuh, tinggal, ...
12  [suka, prgrf, kabinet, lahir, data, daya, anti...
13  [perintah, orang, indonesia, takut, corona, in...
14  [komisi, viii, dpr, ri, desak, perintah, cair,...
15  [jokowi, upaya, perintah, hadap, virus, corona...
16   [welcome, indonesia, jubir, corona, bintangemon]
```

Figure 5. Result of data tweet removal of stop words

d. Rejoin words

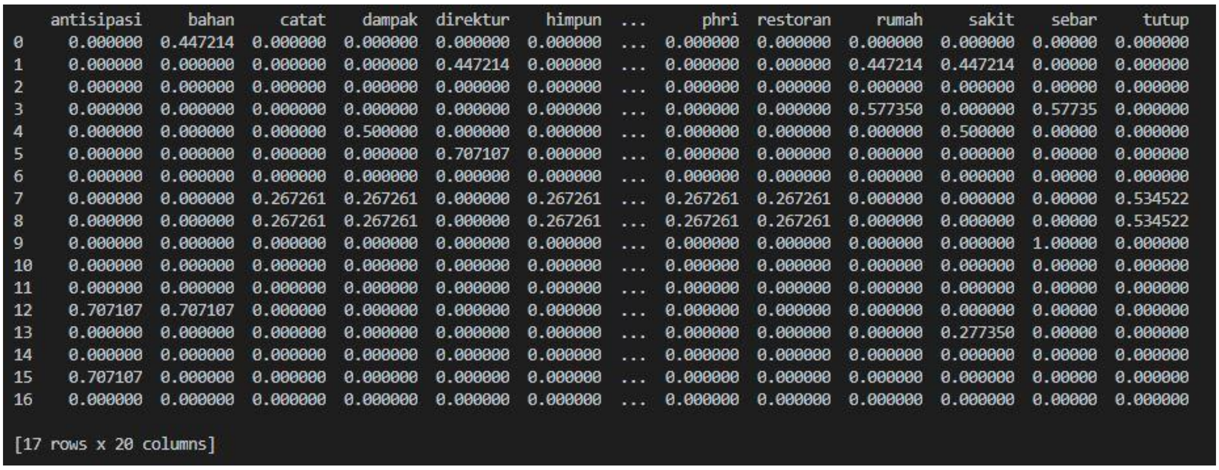
After previously the sentence in a tweet or document was tokenized or separated into words, it was then recombined into a sentence before being processed further. This is because later, the k-mean algorithm will read per sentence. To do this, authors used the `join()` method, which was also available in python. The output of this stage is to obtain a list of tweets in the form of sentences where all the constituent words are in the basic form, there are no affixes or general words that are not important as seen in Figure 6.

C. Analysis of Classification Results



```
    antisipasi  bahan  catat  dampak  direktur  himpun  hotel  maret  ...  pasien  perintah  phri  restoran  rumah  sakit  sebar  tutu
p
0            0      1      0       0         0       0      0      0  ...       0         0     0         0      0      0      0
0
1            0      0      0       0         1       0      0      0  ...       1         0     0         0      1      1      0
0
2            0      0      0       0         0       0      0      0  ...       1         0     0         0      0      0      0
0
3            0      0      0       0         0       0      0      0  ...       0         1     0         0      1      0      1
0
4            0      0      0       1         0       0      0      0  ...       0         0     0         0      0      1      0
0
5            0      0      0       0         1       0      0      0  ...       0         0     0         0      0      0      0
0
6            0      0      0       0         0       0      0      0  ...       0         0     0         0      0      0      0
0
7            0      0      1       1         0       1      2      1  ...       0         0     1         1      0      0      0
2
8            0      0      1       1         0       1      2      1  ...       0         0     1         1      0      0      0
2
9            0      0      0       0         0       0      0      0  ...       0         0     0         0      0      0      1
0
10           0      0      0       0         0       0      0      0  ...       0         0     0         0      0      0      0
0
11           0      0      0       0         0       0      0      0  ...       0         0     0         0      0      0      0
0
12           1      1      0       0         0       0      0      0  ...       0         0     0         0      0      0      0
0
13           0      0      0       0         0       0      0      0  ...       0         1     0         0      0      1      0
0
14           0      0      0       0         0       0      0      0  ...       0         1     0         0      0      0      0
0
15           1      0      0       0         0       0      0      0  ...       0         1     0         0      0      0      0
0
```

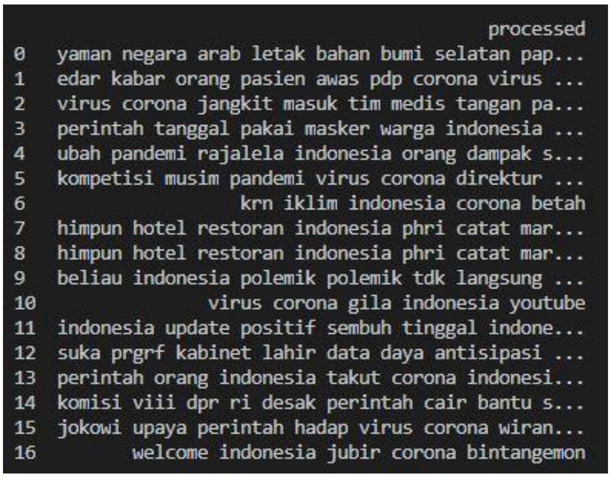
Figure 1. Data tweet word-count



```
    antisipasi     bahan     catat    dampak  direktur    himpun  ...      phri  restoran     rumah     sakit    sebar     tutup
0     0.000000  0.447214  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000  0.00000  0.000000
1     0.000000  0.000000  0.000000  0.000000  0.447214  0.000000  ...  0.000000  0.000000  0.447214  0.447214  0.00000  0.000000
2     0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000  0.00000  0.000000
3     0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.577350  0.000000  0.57735  0.000000
4     0.000000  0.000000  0.000000  0.500000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.500000  0.00000  0.000000
5     0.000000  0.000000  0.000000  0.000000  0.707107  0.000000  ...  0.000000  0.000000  0.000000  0.000000  0.00000  0.000000
6     0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000  0.00000  0.000000
7     0.000000  0.000000  0.267261  0.267261  0.000000  0.267261  ...  0.267261  0.267261  0.000000  0.000000  0.00000  0.534522
8     0.000000  0.000000  0.267261  0.267261  0.000000  0.267261  ...  0.267261  0.267261  0.000000  0.000000  0.00000  0.534522
9     0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000  1.00000  0.000000
10    0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000  0.00000  0.000000
11    0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000  0.00000  0.000000
12    0.707107  0.707107  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000  0.00000  0.000000
13    0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.277350  0.00000  0.000000
14    0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000  0.00000  0.000000
15    0.707107  0.000000  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000  0.00000  0.000000
16    0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000  0.00000  0.000000

[17 rows x 20 columns]
```

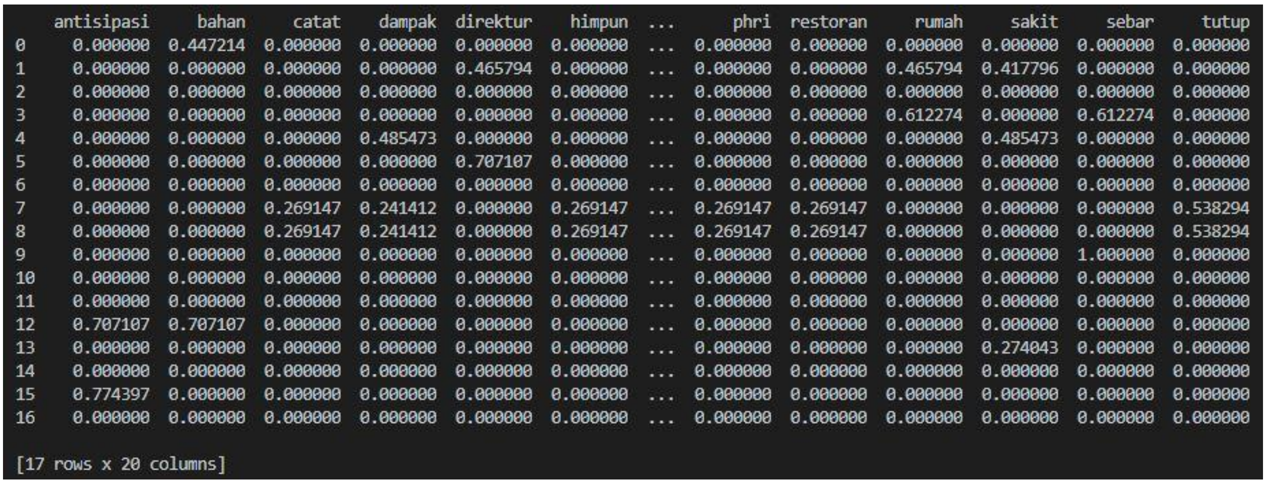
Figure 2. TF vector data tweet



```
                                           processed
0   yaman negara arab letak bahan bumi selatan pap...
1   edar kabar orang pasien awas pdp corona virus ...
2   virus corona jangkit masuk tim medis tangan pa...
3   perintah tanggal pakai masker warga indonesia ...
4   ubah pandemi rajalela indonesia orang dampak s...
5   kompetisi musim pandemi virus corona direktur ...
6                   krn iklim indonesia corona betah
7   himpun hotel restoran indonesia phri catat mar...
8   himpun hotel restoran indonesia phri catat mar...
9   beliau indonesia polemik polemik tdk langsung ...
10                 virus corona gila indonesia youtube
11  indonesia update positif sembuh tinggal indone...
12  suka prgrf kabinet lahir data daya antisipasi ...
13  perintah orang indonesia takut corona indonesi...
14  komisi viii dpr ri desak perintah cair bantu s...
15  jokowi upaya perintah hadap virus corona wiran...
16           welcome indonesia jubir corona bintangemon
```

Figure 6. Results of data tweet rejoin words

In the processing and exploration of data sets using the Python programming language, the authors emphasize more on the modules and methods and make statistical calculations. The initial stage that the writers did was to implement the TF-IDF algorithm. [14] explained that Term Frequency - Inverse Document Frequency (TF-IDF) is an algorithmic method that is useful for calculating the weight of each commonly used word. This method is also known to be efficient, easy and has accurate results. This method will calculate the values Term Frequency (TF) and Inverse Document Frequency (IDF) for each token (word) in each document in the corpus. For this reason, the first step is for the writers to look for the TF value to determine how often a word appears in a document. The more frequent a word occurs, the greater its value is. Then, the authors look for the IDF value to calculate how the terms are widely distributed in the collection of concerned documents. In contrast to TF, in the IDF, the less frequent words appear in the document, the greater the value is.



```
    antisipasi     bahan     catat    dampak  direktur    himpun  ...      phri  restoran     rumah     sakit     sebar     tutup
0     0.000000  0.447214  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
1     0.000000  0.000000  0.000000  0.000000  0.465794  0.000000  ...  0.000000  0.000000  0.465794  0.417796  0.000000  0.000000
2     0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
3     0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.612274  0.000000  0.612274  0.000000
4     0.000000  0.000000  0.000000  0.485473  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.485473  0.000000  0.000000
5     0.000000  0.000000  0.000000  0.000000  0.707107  0.000000  ...  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
6     0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
7     0.000000  0.000000  0.269147  0.241412  0.000000  0.269147  ...  0.269147  0.269147  0.000000  0.000000  0.000000  0.538294
8     0.000000  0.000000  0.269147  0.241412  0.000000  0.269147  ...  0.269147  0.269147  0.000000  0.000000  0.000000  0.538294
9     0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000  1.000000  0.000000
10    0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
11    0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
12    0.707107  0.707107  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
13    0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.274043  0.000000  0.000000
14    0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
15    0.774397  0.000000  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
16    0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000

[17 rows x 20 columns]
```

Figure 3. IDF vector data tweet

In simple terms, here is the formula for finding the IDF value:

$$tf = \begin{cases} 1 + log_{10}(f_{t,d}), & f_{t,d} > 0 \\ 0, & f_{t,d} = 0 \end{cases}$$

$$idf_j = log\left(\frac{D}{df_j}\right)$$

$$w_{ij} = tf_{ij} + idf_{ij}$$

where Wij is the weight of the term (tj) to the document (di); tfij is the number of occurrences of term (tj) in the document (di); D is the number of all documents in the database; and dfj is the number of documents that contain term (tj) (at least one word is term (tj)).

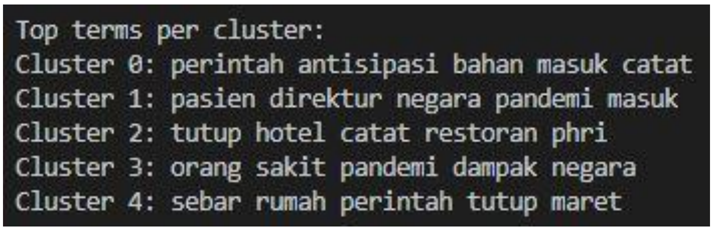
Figure 7 above shows the word count in each document. To do this, we use the `CountVectorizer()` method in the sklearn python module. The TF results in this study are shown in Figure 8. These results show the words in the list document with a minimum support of 20%. The TF results of this study consisted of 20 columns which indicated that there were 20 words, which met the minimum support of 20% after the frequency term being calculated. Not much different, the results of the IDF in this study are as shown in Figure 9. To do this, the authors use the `TfidVectorizer()` method with the attribute `use_idf= False` in the python sklearn module for the TF value and the attribute `use_idf = False` for the IDF value. Of course, these two values must be combined in order to produce a maximum TF-IDF value.

In Figure 9, it can be seen that some words have better IDF values than TF values. This indicates that the word or term is properly distributed in the document collection, so that it becomes more meaningful.

Furthermore, the authors began to apply the K-mean algorithm in the study. One of the reasons why it is important to know TF-IDF is because of document similarity. By knowing these same documents, other related documents can be found and can automatically be grouped into clusters.

[15] explained that the K-means algorithm flowchart is a flowchart that contains input sequences in the form of vector terms from each document, namely the weight of each word/term, finding the number of clusters, determining the initial centroid (center point), looking for distances, classifying documents based on the closest distance to the centroid, as well as the process of finding a new centroid. The Flowchart is completely shown in Figure 10.

In order to cluster the K-mean method, the authors used the K-mean method in the sklearn module. The method is quite simple, just determine the number of clusters, and the program will run it. In this study, the authors chose the number of clusters of 5, considering that the document was also limited to only 16 tweets. The cluster results are as shown in Figure 11. It is also known that the center of the



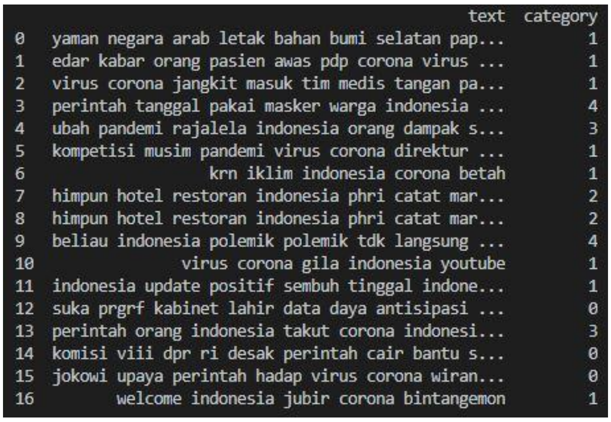
```
Top terms per cluster:
Cluster 0: perintah antisipasi bahan masuk catat
Cluster 1: pasien direktur negara pandemi masuk
Cluster 2: tutup hotel catat restoran phri
Cluster 3: orang sakit pandemi dampak negara
Cluster 4: sebar rumah perintah tutup maret
```

Figure 11. Five clusters on data tweet

cluster is in cluster 3: ‘orang’, ‘sakit’, ‘pandemi’, ‘dampak’, and ‘negara’.

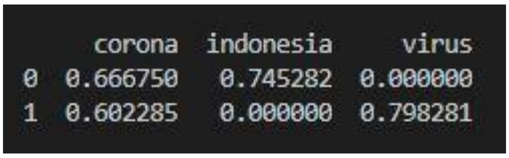
To see the document per cluster category more clearly, consider Figure 12. From this figure, it appears that the words of the cluster do have a pattern that is almost the same in every document or tweet. From here, we can predict what words might become a trend at a certain time. The authors can find out what clusters of words might become trending topics on twitter, for example, if people talk about a pandemic, they might talk about people who are affected, or people who are sick.

There is also another way to create clusters of data tweet by specifying the results of the K-mean method. This method directly utilizes the `max_features=` in method the Sklearn module. The way it works is by looking up a list of the most important words, which will be measured from the K-mean result. This feature can also run automatically in python, without having to make more effort to search for it. Consider Figure 13.



```
                                                  text  category
0    yaman negara arab letak bahan bumi selatan pap...         1
1    edar kabar orang pasien awas pdp corona virus ...         1
2    virus corona jangkit masuk tim medis tangan pa...         1
3    perintah tanggal pakai masker warga indonesia ...         4
4    ubah pandemi rajalela indonesia orang dampak s...         3
5    kompetisi musim pandemi virus corona direktur ...         1
6                  krn iklim indonesia corona betah         1
7    himpun hotel restoran indonesia phri catat mar...         2
8    himpun hotel restoran indonesia phri catat mar...         2
9    beliau indonesia polemik polemik tdk langsung ...         4
10               virus corona gila indonesia youtube         1
11   indonesia update positif sembuh tinggal indone...         1
12   suka prgrf kabinet lahir data daya antisipasi ...         0
13   perintah orang indonesia takut corona indonesi...         3
14   komisi viii dpr ri desak perintah cair bantu s...         0
15   jokowi upaya perintah hadap virus corona wiran...         0
16          welcome indonesia jubir corona bintangemon         1
```

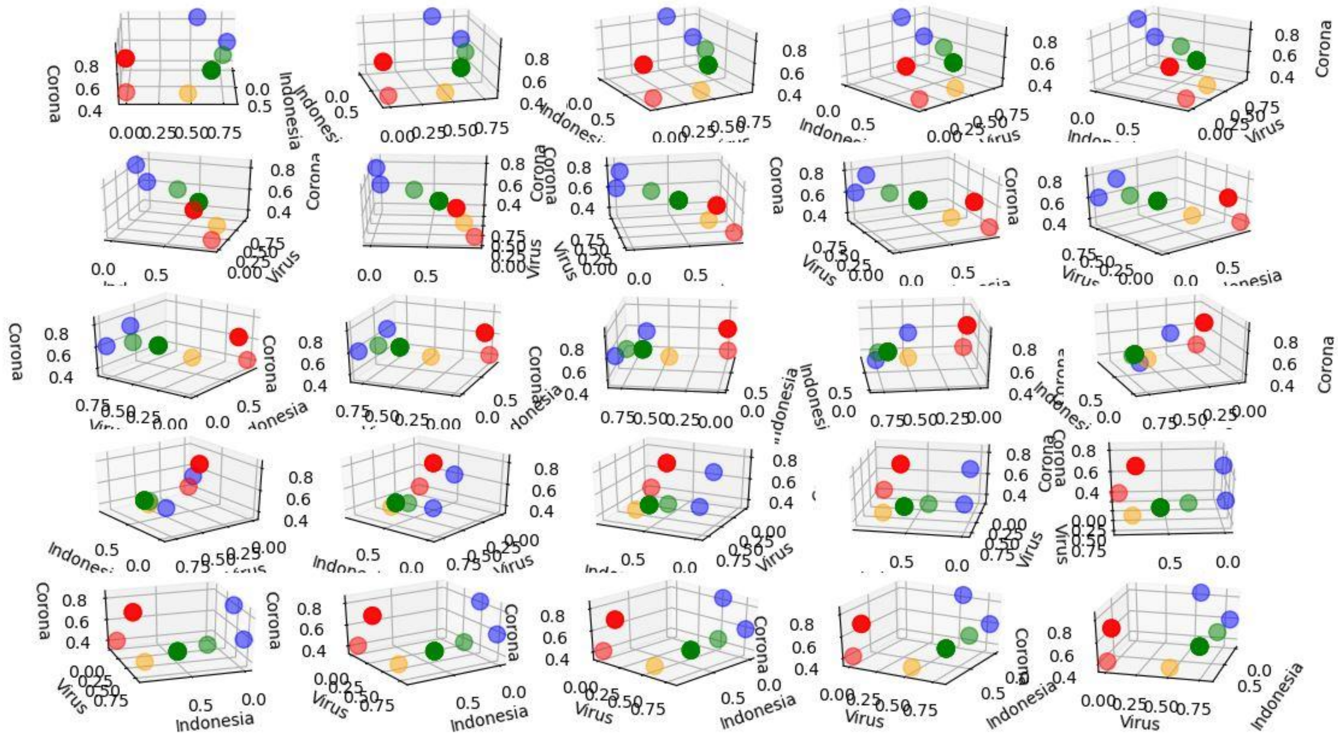
Figure 12. Cluster category



```
     corona  indonesia     virus
0  0.666750   0.745282  0.000000
1  0.602285   0.000000  0.798281
```

Figure 13. 3 best words or tokens

Start



End

Figure 14. 3D graphic of clusterisation

In this study, the authors immediately looked for the 3 most important words from the K-mean method. When the K-mean algorithm is run, the program will request to issue only the most important words that match the K-mean results. Incidentally in this case, the appropriate words are corona, Indonesia and virus. This can also be taken into consideration and predictions to find out the possibility of conversation topics that will be a trend on Twitter.

If displayed in 3D graphics, the cluster shape is as shown in Figure 14.

IV. CONCLUSION

The TF-IDF algorithm and the K-mean method can be used to cluster the topics and tweets that many people convey at one time. The clustering results are used as a prediction to know the possible topics or words that might trend in the public conversation on Twitter. The K-mean method is carried out by utilizing methods and modules regarding data mining contained in the python programming language, so it is quite fast and accurate.

The results of this clustering, for example, can be used as material for creating sentiment to something on Twitter. In addition, the results of clustering like the K-mean model can be used as material for generating opinions through social media.

REFERENCES

[1] A. Mukkamala and R. Beck, “The Role Of Social Media For Collective Behavior Development,” *ECIS 2018 Proc.*, 2018.

[2] K. Rudra, S. Ghosh, N. Ganguly, P. Goyal, and S. Ghosh, “Extracting situational information from microblogs during disaster events: A classification-summarization approach,” *Int. Conf. Inf. Knowl. Manag. Proc.*, vol. 19-23-Oct-, pp. 583–592, 2015.

[3] W. E. Nurjanah, R. S. Perdana, and M. A. Fauzi, “Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 1, no. 12, pp. 1750–1757, 2017.

[4] Tiara, M. K. Sabariah, and V. Effendy, “Analisis Sentimen pada Twitter untuk Menilai Performansi Program Televisi dengan Kombinasi Metode Lexicon-Based dan Support Vector Machine,” *e-Proceeding Eng.*, vol. 2, no. 1, pp. 1237–1247, 2015.

[5] V. Friedemann, “Clustering a Customer Base Using Twitter Data,” *Cs*, vol. 229, no. 1, pp. 1–5, 2015.

[6] R. Soni and K. J. Mathai, “Improved Twitter Sentiment Prediction through Cluster-then-Predict Model,” vol. 4, no. 4, pp. 559–563, 2015.

[7] N. Alnajran, K. Crockett, D. McLean, and A. Latham, “Cluster analysis of twitter data: A review of algorithms,” *ICAART 2017 - Proc. 9th Int. Conf. Agents Artif. Intell.*, vol. 2, no. Icaart, pp. 239–249, 2017.

[8] M. Vicente, F. Batista, and J. P. Carvalho, “Twitter gender classification using user unstructured information,” *IEEE Int. Conf. Fuzzy Syst.*, vol. 2015-Novem, 2015.

[9] X. Wang *et al.*, “A Weakly-supervised Framework for COVID-19 Classification and Lesion Localization from Chest CT,” vol. XX, no. XX, pp. 1–11, 2020.

[10] L. Li *et al.*, “Characterizing the Propagation of Situational Information in Social Media during COVID-19 Epidemic: A Case Study on Weibo,” *IEEE Trans. Comput. Soc. Syst.*, vol. 7, no. 2, pp. 556–562, 2020.

[11] N. Garg and R. Rani, “Analysis and Visualization of Twitter Data using k-means Clustering,” *Int. Conf. Intell. Comput. Control Syst. 2017*, pp. 670–675, 2017.

[12] A. Sechelea, T. Do Huu, E. Zimos, and N. Deligiannis, “Twitter data clustering and visualization,” *2016 23rd Int. Conf. Telecommun. ICT 2016*, vol. 8, pp. 1–5, 2016.

[13] Imamah and F. H. Rachman, “Twitter Sentiment Analysis of Covid-19 Using Term Weighting TF-IDF And Logistic Regresion,” *2020 6th Inf. Technol. Int. Semin.*, p. pp 238-242, 2020.

[14] Delta Sierra, “Algoritma TF — IDF,” *Medium The Startup*, 2019. [Online]. Available: https://medium.com/@dltsierra/algoritma-tf-idf-633e17d10a80.

[15] A. Y. Putri *et al.*, “Ekstraksi Fitur Situs Berita Online Untuk Kaleidoskop,” 2018.