CHAPTER 4
ANALYSIS AND DESIGN

4.1 Analysis

4.1.1 Collecting data

Stages of collecting data from websites with web scrapping techniques. Sources taken from online news in Indonesia are liputan6.com, tribunnews.com, BBC.com, also turnbackhoax.id. The data is taken from the title and content of the news only, both parts will be processed and then analyzed using an algorithm.

Retrieval of data from the website liputan6.com, tribunnews.com and BBC.com using beautifulsoap. Beautifulsoap is a library of Python, which helps in retrieving data from web pages. Furthermore, the results of data collection are saved in CSV format. The following table 4.1 is an example of training data and table 4.2 is an example of testing data.

Table 1: Table Training Data

No.	Title	Content	Source
1	[SALAH] Pasar Tebet Barat Ditutup Total Karena Ada Pedagang Positif Corona	Beredar informasi yang menyebutkan Pasar Tebet Barat ditutup lantaran ada pedagang terpapar virus Corona atau wabah COVID-19. Dalam pesan itu juga disebutkan bahwa terdapat dua orang karyawan toko yang positif dan 12 orang lainnya saspek. Berikut kutipan narasinya: *Breaking news* Orang Tampa Gejala - OTG (semoga jadi PERHATIAN	Turnbackhoax.id

		berSAMA) *JUST INFO* _*Pasar Tebet Barat*_ ditutup total karena pemilik _*Toko Kristal*_ di lantai 2 positif dan meninggal kemarin, lalu Kepala Pasar Jaya Tebet Barat inisiatif untuk rapid test ke semua pedagang Pasar Tebet Barat, hasilnya : _*2 orang karyawan toko kristal positif, dan 12 orang suspect*_ Hebatnya ke-14 orang itu baik2 aja, gak demam, gak sakit tenggorokan atau sesak nafas. Terinfeksi tanpa gejala, akhirnya dibawa langsung ke Wisma Atlet dan semua pedagang Pasar Tebet Barat wajib karantina mandiri di rumah dibawah Dinkes.	
2.	2 Hal Ini Sebabkan Kasus Positif Corona di Indonesia Terus Bertambah	Sampai dengan hari ini, total kumulatif pasien terinfeksi virus Corona Covid-19 sebanyak 4.839 orang.	Liputan 6

Table 2: Table Testing Data

1.	Pasien Positif Corona di Jakarta Kini Didampingi Psikolog	Hingga hari ini, jumlah pasien positif Covid-19 di Jakarta mencapai 3.112 orang. Pasien yang dinyatakan sembuh sebanyak 237 orang, meninggal 297 orang.	Liputan 6	?

4.1.2 Pre-Processing

Pre-Processing is the initial stage in processing a data. The pre-processing stage consists of several steps, namely folding, tokenizing, filtering, stemming. In addition, in this study the researchers added the stages of removing the '[]' mark in each document also changing each document to ASCII code. Here are some steps when doing pre-processing.

a. Change to ASCII type

This stage returns the string to type ASCII so that it can be processed to the next process. ASCII is an international standard in letters and symbols. The following are Table 4.3 and Table 4.4 which show the results after going through the ASCII type.

Table 3: Training Data Results that have been changed to ASCII type

Before			**After**	
No.	**Title**	**Content**	**Title**	**Content**
1	[SALAH] Pasar Tebet Barat Ditutup Total Karena Ada	Beredar informasi yang menyebutkan Pasar Tebet Barat	[SALAH] Pasar Tebet Barat Ditutup Total Karena Ada	Beredar informasi yang menyebutkan

	Pedagang Positif Corona	ditutup lantaran ada pedagang terpapar virus Corona atau wabah COVID-19. Dalam pesan itu juga disebutkan bahwa terdapat dua orang karyawan toko yang positif dan 12 orang lainnya saspek. Berikut kutipan narasinya: *Breaking news* Orang Tampa Gejala - OTG (semoga jadi PERHATIAN berSAMA) *JUST INFO* _*Pasar Tebet Barat*_ ditutup total karena pemilik _*Toko Kristal*_ di lantai 2 positif dan meninggal kemarin, lalu Kepala Pasar Jaya Tebet Barat inisiatif untuk rapid test ke semua pedagang Pasar	Pedagang Positif Corona	Pasar Tebet Barat ditutup lantaran ada pedagang terpapar virus Corona atau wabah COVID-19. Dalam pesan itu juga disebutkan bahwa terdapat dua orang karyawan toko yang positif dan 12 orang lainnya saspek. Berikut kutipan narasinya:\n\n*Breaking news* Orang Tampa Gejala - OTG\n(semoga jadi PERHATIAN berSAMA)\n\n*JUST INFO* _*Pasar Tebet Barat*_

		Tebet Barat, hasilnya : _*2 orang karyawan toko kristal positif, dan 12 orang suspect*_ Hebatnya ke-14 orang itu baik2 aja, gak demam, gak sakit tenggorokan atau sesak nafas. Terinfeksi tanpa gejala, akhirnya dibawa langsung ke Wisma Atlet dan semua pedagang Pasar Tebet Barat wajib karantina mandiri di rumah dibawah Dinkes.		ditutup total karena pemilik _*Toko Kristal*_ di lantai 2 positif dan meninggal kemarin, lalu Kepala Pasar Jaya Tebet Barat inisiatif untuk rapid test ke semua pedagang Pasar Tebet Barat, hasilnya : _*2 orang karyawan toko kristal positif, dan 12 orang suspect*_ Hebatnya ke-14 orang itu baik2 aja, gak demam, gak sakit tenggorokan atau sesak nafas. Terinfeksi tanpa gejala, akhirnya dibawa

				langsung ke Wisma Atlet dan semua pedagang Pasar Tebet Barat wajib karantina mandiri di rumah dibawah Dinkes.
2.	2 Hal Ini Sebabkan Kasus Positif Corona di Indonesia Terus Bertambah	Sampai dengan hari ini, total kumulatif pasien terinfeksi virus Corona Covid-19 sebanyak 4.839 orang.	2 Hal Ini Sebabkan Kasus Positif Corona di Indonesia Terus Bertambah	Sampai dengan hari ini, total kumulatif pasien terinfeksi virus Corona Covid-19 sebanyak 4.839 orang.

Table 4: Results of Testing Data that have been changed to ASCII type

Before			**After**	
No.	**Title**	**Content**	**Title**	**Content**
1.	Pasien Positif Corona di Jakarta Kini Didampingi Psikolog	Hingga hari ini, jumlah pasien positif Covid-19 di Jakarta mencapai 3.112 orang. Pasien yang dinyatakan sembuh	Pasien Positif Corona di Jakarta Kini Didampingi Psikolog	Hingga hari ini, jumlah pasien positif Covid-19 di Jakarta mencapai

		sebanyak 237 orang, meninggal 297 orang.		3.112 orang. Pasien yang dinyatakan sembuh sebanyak 237 orang, meninggal 297 orang.

b. Stage of removing ‘[] ’

This stage removes ‘[]’ and the contents of the words contained in the character. At this stage, using the regular expression technique. The following Tables 4.5 and 4.6 contain before and after undergoing the process of removing the mark.

Table 5: Results of Training Data that have been removed ‘[]’

Before			**After**	
No.	**Title**	**Content**	**Title**	**Content**
1	[SALAH] Pasar Tebet Barat Ditutup Total Karena Ada Pedagang Positif Corona	Beredar informasi yang menyebutkan Pasar Tebet Barat ditutup lantaran ada pedagang terpapar virus Corona atau wabah COVID-19. Dalam pesan itu juga disebutkan bahwa terdapat dua orang karyawan toko yang positif dan 12 orang lainnya saspek.	Pasar Tebet Barat Ditutup Total Karena Ada Pedagang Positif Corona	Beredar informasi yang menyebutkan Pasar Tebet Barat ditutup lantaran ada pedagang terpapar virus Corona atau wabah COVID-19. Dalam pesan itu juga disebutkan

		Berikut kutipan narasinya: *Breaking news* Orang Tampa Gejala - OTG (semoga jadi PERHATIAN berSAMA) *JUST INFO* _*Pasar Tebet Barat*_ ditutup total karena pemilik _*Toko Kristal*_ di lantai 2 positif dan meninggal kemarin, lalu Kepala Pasar Jaya Tebet Barat inisiatif untuk rapid test ke semua pedagang Pasar Tebet Barat, hasilnya : _*2 orang karyawan toko kristal positif, dan 12 orang suspect*_ Hebatnya ke-14 orang itu baik2 aja, gak demam, gak sakit tenggorokan atau sesak nafas.		bahwa terdapat dua orang karyawan toko yang positif dan 12 orang lainnya saspek. Berikut kutipan narasinya:\n\n\n*Breaking news* Orang Tampa Gejala - OTG\nN berSAMA)\n\n\n*JUST INFO* _*Pasar Tebet Barat*_ ditutup total karena pemilik _*Toko Kristal*_ di lantai 2 positif dan meninggal kemarin, lalu Kepala Pasar Jaya Tebet Barat inisiatif untuk rapid test ke semua pedagang Pasar Tebet Barat, hasilnya : _*2 orang karyawan toko kristal

		Terinfeksi tanpa gejala, akhirnya dibawa langsung ke Wisma Atlet dan semua pedagang Pasar Tebet Barat wajib karantina mandiri di rumah dibawah Dinkes.		positif, dan 12 orang suspect*_ Hebatnya ke-14 orang itu baik2 aja, gak demam, gak sakit tenggorokan atau sesak nafas. Terinfeksi tanpa gejala, akhirnya dibawa langsung ke Wisma Atlet dan semua pedagang Pasar Tebet Barat wajib karantina mandiri di rumah dibawah Dinkes.\n
2.	2 Hal Ini Sebabkan Kasus Positif Corona di Indonesia Terus Bertambah	Sampai dengan hari ini, total kumulatif pasien terinfeksi virus Corona Covid-19 sebanyak 4.839 orang.	2 Hal Ini Sebabkan Kasus Positif Corona di Indonesia Terus Bertambah	Sampai dengan hari ini, total kumulatif pasien terinfeksi virus Corona Covid-19 sebanyak

				4.839 orang.

Table 6: Results of Testing Data that have been removed ‘[]’

Before			After	
No.	**Title**	**Content**	**Title**	**Content**
1	Pasien Positif Corona di Jakarta Kini Didampingi Psikolog	Hingga hari ini, jumlah pasien positif Covid-19 di Jakarta mencapai 3.112 orang. Pasien yang dinyatakan sembuh sebanyak 237 orang, meninggal 297 orang.	Pasien Positif Corona di Jakarta Kini Didampingi Psikolog\n	Hingga hari ini, jumlah pasien positif Covid-19 di Jakarta mencapai 3.112 orang. Pasien yang dinyatakan sembuh sebanyak 237 orang, meninggal 297 orang.

c. Case Folding

This stage turns all letters into letters in the document into lowercase letters. Letters received ‘a’ through ‘z’. The following table 4.7 is the result of changes before and after going through the folding process for training data and table 4.8 is the result of changes before and after going through the folding process for testing data.

Table 7: Results of Training Data that have been through the Folding process

Before			After	
No.	**Title**	**Content**	**Title**	**Content**
1	Pasar Tebet Barat	Beredar informasi	pasar tebet barat	beredar

	Ditutup Total Karena Ada Pedagang Positif Corona	yang menyebutkan Pasar Tebet Barat ditutup lantaran ada pedagang terpapar virus Corona atau wabah COVID-19. Dalam pesan itu juga disebutkan bahwa terdapat dua orang karyawan toko yang positif dan 12 orang lainnya saspek. Berikut kutipan narasinya:\n\ n*Breaking news* Orang Tampa Gejala - OTG\nN berSAMA)\n\ n*JUST INFO* _*Pasar Tebet Barat*_ ditutup total karena pemilik _*Toko Kristal*_ di lantai 2 positif dan meninggal kemarin, lalu Kepala Pasar Jaya Tebet Barat inisiatif untuk rapid test ke semua pedagang Pasar Tebet Barat, hasilnya : _*2 orang karyawan toko kristal positif, dan 12 orang suspect*_ Hebatnya ke-14 orang itu baik2 aja, gak demam, gak sakit tenggorokan atau sesak nafas. Terinfeksi tanpa gejala, akhirnya dibawa langsung ke Wisma Atlet dan semua pedagang Pasar Tebet Barat wajib karantina mandiri di rumah	ditutup total karena ada pedagang positif corona	informasi yang menyebutkan pasar tebet barat ditutup lantaran ada pedagang terpapar virus corona atau wabah covid-19. dalam pesan itu juga disebutkan bahwa terdapat dua orang karyawan toko yang positif dan 12 orang lainnya saspek. berikut kutipan narasinya:\n\ n*breaking news* orang tampa gejala - otg\nn bersama)\n\ n*just info* _*pasar tebet barat*_ ditutup total karena pemilik _*toko kristal*_ di lantai 2 positif dan meninggal kemarin, lalu kepala pasar jaya tebet barat inisiatif untuk rapid test ke semua pedagang pasar tebet barat, hasilnya : _*2 orang karyawan toko kristal positif, dan 12 orang suspect*_ hebatnya ke-14

		dibawah Dinkes.\n		orang itu baik2 aja, gak demam, gak sakit tenggorokan atau sesak nafas. terinfeksi tanpa gejala, akhirnya dibawa langsung ke wisma atlet dan semua pedagang pasar tebet barat wajib karantina mandiri di rumah dibawah dinkes.
2.	2 Hal Ini Sebabkan Kasus Positif Corona di Indonesia Terus Bertambah	Sampai dengan hari ini, total kumulatif pasien terinfeksi virus Corona Covid-19 sebanyak 4.839 orang.	2 hal ini sebabkan kasus positif corona di indonesia terus bertambah	sampai dengan hari ini, total kumulatif pasien terinfeksi virus corona covid-19 sebanyak 4.839 orang

Table 8: Testing Data Results that have been through the Folding process

Before			**After**	
No.	**Title**	**Content**	**Title**	**Content**
1	Pasien Positif Corona di Jakarta Kini Didampingi Psikolog	Hingga hari ini, jumlah pasien positif Covid-19 di Jakarta mencapai 3.112 orang. Pasien yang dinyatakan sembuh sebanyak 237 orang, meninggal 297	pasien positif corona di jakarta kini didampingi psikolog	hingga hari ini, jumlah pasien positif covid-19 di jakarta mencapai 3.112 orang. pasien yang dinyatakan

		orang.		sembuh sebanyak 237 orang, meninggal 297 orang.

d. Case Stemming

This stage is a process of changing the words in a document into basic form words. Following table 4.9 and table 4.10 which contains a comparison before and after the stemming process.

Table 9: Results of Training Data that have gone through the Stemming process

Before			**After**	
No.	**Title**	**Content**	**Title**	**Content**
1	pasar tebet barat ditutup total karena ada pedagang positif corona	beredar informasi yang menyebutkan pasar tebet barat ditutup lantaran ada pedagang terpapar virus corona atau wabah covid-19. dalam pesan itu juga disebutkan bahwa terdapat dua orang karyawan toko yang positif dan 12 orang lainnya saspek. berikut kutipan narasinya:\n\ n*breaking news* orang tampa gejala -	pasar tebet barat tutup total karena ada dagang positif corona	edar informasi yang sebut pasar tebet barat tutup lantar ada dagang papar virus corona atau wabah covid-19 dalam pesan itu juga sebut bahwa dapat dua orang karyawan toko yang positif dan 12 orang lain saspek ikut kutip narasi

		otg\nn bersama)\n\n*just info* _*pasar tebet barat*_ ditutup total karena pemilik _*toko kristal*_ di lantai 2 positif dan meninggal kemarin, lalu kepala pasar jaya tebet barat inisiatif untuk rapid test ke semua pedagang pasar tebet barat, hasilnya : _*2 orang karyawan toko kristal positif, dan 12 orang suspect*_ hebatnya ke-14 orang itu baik2 aja, gak demam, gak sakit tenggorokan atau sesak nafas. terinfeksi tanpa gejala, akhirnya dibawa langsung ke wisma atlet dan semua pedagang pasar tebet barat wajib karantina mandiri di rumah dibawah dinkes.		breaking news orang tampa gejala - otg n sama just info pasar tebet barat tutup total karena milik toko kristal di lantai 2 positif dan tinggal kemarin lalu kepala pasar jaya tebet barat inisiatif untuk rapid test ke semua dagang pasar tebet barat hasil 2 orang karyawan toko kristal positif dan 12 orang suspect hebat ke-14 orang itu baik2 aja gak demam gak sakit tenggorok atau sesak nafas infeksi tanpa gejala akhir bawa langsung ke wisma atlet

				dan semua dagang pasar tebet barat wajib karantina mandiri di rumah bawah dinkes
2.	2 hal ini sebabkan kasus positif corona di indonesia terus bertambah	sampai dengan hari ini, total kumulatif pasien terinfeksi virus corona covid-19 sebanyak 4.839 orang	2 hal ini sebab kasus positif corona di indonesia terus tambah	sampai dengan hari ini total kumulatif pasien infeksi virus corona covid-19 banyak 4 839 orang

Table 10: Results of Testing Data that have gone through the Stemming process

Before			**After**	
No.	**Title**	**Content**	**Title**	**Content**
1	pasien positif corona di jakarta kini didampingi psikolog	hingga hari ini, jumlah pasien positif covid-19 di jakarta mencapai 3.112 orang. pasien yang dinyatakan sembuh sebanyak 237 orang, meninggal 297 orang.	pasien positif corona di jakarta kini damping psikolog	hingga hari ini jumlah pasien positif covid-19 di jakarta capai 3 112 orang pasien yang nyata sembuh banyak 237 orang tinggal 297 orang

e. Case Filtering

This filtering stage is to remove words that contain conjunctions such as di, ke, dari, dan, atau, etc. The following table 4.11 contains a comparison of Training data before and after the Filtering process and 4.12 which contains a comparison of Training data before and after the Filtering process.

Table 11: Results of Training Data that have been through the filtering process

Before			After	
No.	**Title**	**Content**	**Title**	**Content**
1	pasar tebet barat tutup total karena ada dagang positif corona	edar informasi yang sebut pasar tebet barat tutup lantar ada dagang papar virus corona atau wabah covid-19 dalam pesan itu juga sebut bahwa dapat dua orang karyawan toko yang positif dan 12 orang lain saspek ikut kutip narasi breaking news orang tampa gejala - otg n sama just info pasar tebet barat tutup total karena milik toko kristal di lantai 2 positif dan tinggal kemarin lalu kepala pasar jaya tebet barat inisiatif untuk rapid test ke semua	pasar tebet barat tutup total ada dagang positif corona	edar informasi sebut pasar tebet barat tutup lantar dagang papar virus corona wabah covid-19 pesan juga sebut dapat orang karyawan toko positif 12 orang saspek ikut kutip narasi breaking news orang tampa gejala - otg n sama just info pasar tebet barat tutup total milik toko kristal lantai 2 positif tinggal

		dagang pasar tebet barat hasil 2 orang karyawan toko kristal positif dan 12 orang suspect hebat ke-14 orang itu baik2 aja gak demam gak sakit tenggorok atau sesak nafas infeksi tanpa gejala akhir bawa langsung ke wisma atlet dan semua dagang pasar tebet barat wajib karantina mandiri di rumah bawah dinkes		kemarin lalu kepala pasar jaya tebet barat inisiatif rapid test semua dagang pasar tebet barat hasil 2 orang karyawan toko kristal positif 12 orang suspect hebat ke-14 orang baik2 aja gak demam gak sakit tenggorok sesak nafas infeksi gejala akhir bawa langsung wisma atlet semua dagang pasar tebet barat wajib karantina mandiri rumah bawah dinkes
2.	2 hal ini sebab kasus positif corona di indonesia terus tambah	sampai dengan hari ini total kumulatif pasien infeksi virus corona covid-19	2 ini kasus positif corona indonesia terus tambah	dengan hari total kumulatif pasien infeksi virus corona

		banyak 4 839 orang		covid-19 banyak 4 839 orang

Table 12: Results of Testing Data that have been through the filtering process

Before			After	
No.	**Title**	**Content**	**Title**	**Content**
1	pasien positif corona di jakarta kini damping psikolog	hingga hari ini jumlah pasien positif covid-19 di jakarta capai 3 112 orang pasien yang nyata sembuh banyak 237 orang tinggal 297 orang	pasien positif corona jakarta kini damping psikolog	hingga hari jumlah pasien positif covid-19 jakarta capai 3 112 orang pasien nyata sembuh banyak 237 orang tinggal 297 orang

4.1.3 Calculate Term Frequency

This stage counts for the number of times the term or keyword appears in each document. Before doing this stage, the data needs to be broken down by word.

Training Data :

Document 1 (Title) : pasar tebet barat tutup total karena ada dagang positif corona

Document 1 (Content) : edar informasi yang sebut pasar tebet barat tutup lantar ada dagang papar virus corona atau wabah covid-19 dalam pesan itu juga sebut bahwa dapat dua orang karyawan toko yang positif dan 12 orang lain saspek ikut kutip narasi breaking news orang tampa gejala - otg n sama just info pasar tebet

barat tutup total karena milik toko kristal di lantai 2 positif dan tinggal kemarin lalu kepala pasar jaya tebet barat inisiatif untuk rapid test ke semua dagang pasar tebet barat hasil 2 orang karyawan toko kristal positif dan 12 orang suspect hebat ke-14 orang itu baik2 aja gak demam gak sakit tenggorok atau sesak nafas infeksi tanpa gejala akhir bawa langsung ke wisma atlet dan semua dagang pasar tebet barat wajib karantina mandiri di rumah bawah dinkes

Document 2 (Title) : 2 hal ini sebabkan kasus positif corona di indonesia terus bertambah.
Document 2 (Content) : sampai dengan hari ini total kumulatif pasien infeksi virus corona covid-19 banyak 4 839 orang.

Testing Data :
Document 3 (Title) : pasien positif corona di jakarta kini damping psikolog.
Document 3 (Content) : hingga hari ini jumlah pasien positif covid-19 di jakarta capai 3 112 orang pasien yang nyata sembuh banyak 237 orang tinggal 297 orang.

The following table 4.13 which contains the results of the title data split process and 4.14 which contains the results of the content part split process.

Table 13: Split Data Results (Title)

Word(Title)		
2	ada	barat
corona	dagang	damping
indonesia	ini	jakarta
kasus	kini	pasar
pasien	positif	psikolog
tambah	tebet	terus
total	tutup	

Table 14: Split Data Results (Content)

Word(Content)		
-	112	12
2	237	297
3	4	839
aja	akhir	atlet
baik2	banyak	barat
bawa	bawah	breaking
capai	corona	covid-19
dagang	dapat	demam
dengan	dinkes	edar
gak	gejala	hari
hasil	hebat	hingga
ikut	infeksi	info
informasi	inisiatif	jakarta
jaya	juga	jumlah
just	karantina	karyawan
ke-14	kemarin	kepala
kristal	kumulatif	kutip
lalu	langsung	lantai
lantar	mandiri	milik
n	nafas	narasi
news	nyata	orang
otg	papar	pasar
pasien	pesan	positif
rapid	rumah	sakit
sama	saspek	sebut
sembuh	semua	sesak
suspect	tampa	tebet
tenggorok	test	tinggal
toko	total	tutup
virus	wabah	wajib

wisma		

Result Term Frequency (Title) :

Training Data :

Document 1 (Title) : pasar tebet barat tutup total karena ada dagang positif corona.

Document 2 (Title) : 2 hal ini sebabkan kasus positif corona di indonesia terus. bertambah.

Testing Data :

Document 3 (Title) : pasien positif corona di jakarta kini damping psikolog.

The following table 4.15 contains the results of the title split process.

Table 15: Term Frequency Results (Title)

No.	Term	Document	Count
1	2	Document 1	0
		Document 2	1
		Document 3	0
2	ada	Document 1	1
		Document 2	0
		Document 3	0
3	barat	Document 1	1
		Document 2	0
		Document 3	0
4	corona	Document 1	1
		Document 2	1
		Document 3	1
5	dagang	Document 1	1
		Document 2	0
		Document 3	0
6	damping	Document 1	0
		Document 2	0

		Document 3	1
7	indonesia	Document 1	0
		Document 2	1
		Document 3	0
8	ini	Document 1	0
		Document 2	1
		Document 3	1
9	jakarta	Document 1	0
		Document 2	0
		Document 3	1
10	kasus	Document 1	0
		Document 2	1
		Document 3	0
11	kini	Document 1	0
		Document 2	0
		Document 3	1
12	pasar	Document 1	1
		Document 2	0
		Document 3	0
13	pasien	Document 1	0
		Document 2	0
		Document 3	1
14	positif	Document 1	1
		Document 2	1
		Document 3	1
15	psikolog	Document 1	0
		Document 2	0
		Document 3	1
16	tambah	Document 1	0
		Document 2	1
		Document 3	0
17	tebet	Document 1	1

		Document 2	0
		Document 3	0
18	terus	Document 1	0
		Document 2	1
		Document 3	0
19	total	Document 1	1
		Document 2	0
		Document 3	0
20	tutup	Document 1	1
		Document 2	0
		Document 3	0

Training Data :

Document 1 (Content) : edar informasi yang sebut pasar tebet barat tutup lantar ada dagang papar virus corona atau wabah covid-19 dalam pesan itu juga sebut bahwa dapat dua orang karyawan toko yang positif dan 12 orang lain saspek ikut kutip narasi breaking news orang tampa gejala - otg n sama just info pasar tebet barat tutup total karena milik toko kristal di lantai 2 positif dan tinggal kemarin lalu kepala pasar jaya tebet barat inisiatif untuk rapid test ke semua dagang pasar tebet barat hasil 2 orang karyawan toko kristal positif dan 12 orang suspect hebat ke-14 orang itu baik2 aja gak demam gak sakit tenggorok atau sesak nafas infeksi tanpa gejala akhir bawa langsung ke wisma atlet dan semua dagang pasar tebet barat wajib karantina mandiri di rumah bawah dinkes.

Document 2 (Content) : sampai dengan hari ini total kumulatif pasien infeksi virus corona covid-19 banyak 4 839 orang.

Testing Data :

Document 3 (Content) : hingga hari ini jumlah pasien positif covid-19 di jakarta capai 3 112 orang pasien yang nyata sembuh banyak 237 orang tinggal 297 orang.

The following table 4.16 contains the results of the process of split data content section.

Table 16: Term Frequency Results (Content)

No.	Term	Document	Count	No.	Term	Document	Count
1	-	Document 1	3	46	kepala	Document 1	1
		Document 2	1			Document 2	0
		Document 3	1			Document 3	0
2	112	Document 1	0	47	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	1			Document 3	0
3	12	Document 1	2	48	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	1			Document 3	0
4	2	Document 1	5	49	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	3			Document 3	0
5	237	Document 1	0	50	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	1			Document 3	0
6	297	Document 1	0	51	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	1			Document 3	0
7	3	Document 1	0	52	kepala	Document 1	1
		Document 2	1			Document 2	0
		Document 3	2			Document 3	0
8	4	Document 1	1	53	kepala	Document 1	1
		Document 2	1			Document 2	0
		Document 3	0			Document 3	0
9	839	Document 1	0	54	kepala	Document 1	1
		Document 2	1			Document 2	0
		Document 3	0			Document 3	0
10	aja	Document 1	0	55	kepala	Document 1	1

		Document 2	1			Document 2	0
		Document 3	0			Document 3	0
11	akhir	Document 1	1	56	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
12	atlet	Document 1	1	57	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
13	banyak	Document 1	0	58	kepala	Document 1	1
		Document 2	1			Document 2	0
		Document 3	1			Document 3	0
14	barat	Document 1	5	59	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
15	bawa	Document 1	2	60	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
16	bawah	Document 1	1	61	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
17	breaking	Document 1	1	62	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
18	capai	Document 1	0	63	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	1			Document 3	0
19	corona	Document 1	1	64	kepala	Document 1	1
		Document 2	1			Document 2	0
		Document 3	0			Document 3	0
20	covid-19	Document 1	1	65	kepala	Document 1	1
		Document 2	1			Document 2	0
		Document 3	1			Document 3	0

21	dagang	Document 1	3	66	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
22	dapat	Document 1	1	67	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
23	demam	Document 1	0	68	kepala	Document 1	1
		Document 2	1			Document 2	0
		Document 3	0			Document 3	0
24	dinkes	Document 1	1	69	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
25	edar	Document 1	1	70	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
26	gak	Document 1	2	71	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
27	gejala	Document 1	2	72	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
28	hari	Document 1	0	73	kepala	Document 1	1
		Document 2	1			Document 2	0
		Document 3	1			Document 3	0
29	hasil	Document 1	1	74	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
30	hebat	Document 1	1	75	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
31	hingga	Document 1	0	76	kepala	Document 1	1
		Document 2	0			Document 2	0

		Document 3	1			Document 3	0
32	ikut	Document 1	1	77	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
33	infeksi	Document 1	1	78	kepala	Document 1	1
		Document 2	1			Document 2	0
		Document 3	0			Document 3	0
34	info	Document 1	2	79	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
35	informasi	Document 1	1	80	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
36	inisiatif	Document 1	0	81	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	1			Document 3	0
37	jakarta	Document 1	0	82	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	1			Document 3	0
38	jaya	Document 1	1	83	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
39	juga	Document 1	1	84	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
40	jumlah	Document 1	0	85	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	1			Document 3	0
41	just	Document 1	1	86	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
42	karantina	Document 1	1	87	kepala	Document 1	1

		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
43	karyawan	Document 1	2	88	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
44	ke-14	Document 1	1	89	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0
45	kemarin	Document 1	1	90	kepala	Document 1	1
		Document 2	0			Document 2	0
		Document 3	0			Document 3	0

4.1.4 Calculate Document Frequency

Document Frequency is the number of documents based on the terms that appear. The following table 4.17 contains the results of the Document Frequency section title. While table 4.18 contains the results of the Document Frequency content section.

Table 17: Document Frequency Result (Title)

No.	Term	df
1	2	1
2	ada	1
3	barat	1
4	corona	3
5	dagang	1
6	damping	1
7	indonesia	1
8	ini	2
9	jakarta	1
10	kasus	1
11	kini	1
12	pasar	1

13	pasien	1
14	positif	3
15	psikolog	1
16	tambah	1
17	tebet	1
18	terus	1
19	total	1
20	tutup	1

Table 18: Document Frequency Result (Content)

No.	Term	df	No.	Term	df
1	839	1	46	pesan	1
2	just	1	47	jaya	1
3	orang	10	48	nafas	1
4	akhir	1	49	sakit	1
5	informasi	1	50	sebut	2
6	rapid	1	51	total	2
7	banyak	2	52	demam	1
8	baik2	1	53	hari	2
9	4	2	54	capai	1
10	lantar	1	55	atlet	1
11	jumlah	1	56	papar	1
12	dagang	3	57	2	8
13	wisma	1	58	juga	1
14	ke-14	1	59	otg	1
15	edar	1	60	gak	2
16	lalu	1	61	tampa	1
17	297	1	62	info	2
18	toko	3	63	12	3
19	saspek	1	64	wabah	1
20	kutip	1	65	n	49
21	kemarin	1	66	nyata	1
22	112	1	67	bawah	1

23	dapat	1	68	positif	4
24	semua	2	69	tutup	2
25	tinggal	2	70	kepala	1
26	rumah	1	71	milik	1
27	tenggorok	1	72	gejala	2
28	3	3	73	inisiatif	1
29	dengan	1	74	corona	2
30	suspect	1	75	237	1
31	aja	1	76	kumulatif	1
32	narasi	1	77	infeksi	2
33	hingga	1	78	sama	1
34	sembuh	1	79	-	5
35	kristal	2	80	langsung	1
36	tebet	5	81	pasar	5
37	barat	5	82	test	1
38	pasien	3	83	dinkes	1
39	lantai	1	84	bawa	2
40	news	1	85	hebat	1
41	mandiri	1	86	karantina	1
42	hasil	1	87	virus	2
43	karyawan	2	88	jakarta	1
44	sesak	1	89	covid-19	3
45	wajib	1	90	ikut	1

4.1.5 Inverse Document Frequency

$$idf_i = \log(N / df_i)$$

N = Number of documents in the collection

df = Number of documents containing term (document frequency)

The following table 4.19 is the results of the Inverse Document Frequency title section and the results of the Inverse Document Frequency section content.

Table 19: Inverse Document Frequency Result (Title)

No.	Term	idf
1	2	1.386294
2	ada	1.386294
3	barat	1.386294
4	corona	0.693147
5	dagang	1.386294
6	damping	1.386294
7	indonesia	1.386294
8	ini	0.693147
9	jakarta	1.386294
10	kasus	1.386294
11	kini	1.386294
12	pasar	1.386294
13	pasien	1.386294
14	positif	0.693147
15	psikolog	1.386294
16	tambah	1.386294
17	tebet	1.386294
18	terus	1.386294
19	total	1.386294
20	tutup	1.386294

Table 20: Inverse Document Frequency Result (Content)

No.	Term	idf	No.	Term	idf
1	839	1.386294	46	pesan	1.386294
2	just	1.386294	47	jaya	1.386294
3	orang	0	48	nafas	1.386294
4	akhir	1.386294	49	sakit	1.386294
5	informasi	1.386294	50	sebut	0.693147
6	rapid	1.386294	51	total	0.693147
7	banyak	0.693147	52	demam	1.386294
8	baik2	1.386294	53	hari	0.693147
9	4	0.693147	54	capai	1.386294

10	lantar	1.386294	55	atlet	1.386294
11	jumlah	1.386294	56	papar	1.386294
12	dagang	0.693147	57	2	0.693147
13	wisma	1.386294	58	juga	1.386294
14	ke-14	1.386294	59	otg	1.386294
15	edar	1.386294	60	gak	0.693147
16	lalu	1.386294	61	tampa	1.386294
17	297	1.386294	62	info	0.693147
18	toko	0.693147	63	12	0.693147
19	saspek	1.386294	64	wabah	1.386294
20	kutip	1.386294	65	n	0
21	kemarin	1.386294	66	nyata	1.386294
22	112	1.386294	67	bawah	1.386294
23	dapat	1.386294	68	positif	0
24	semua	0.693147	69	tutup	0.693147
25	tinggal	0.693147	70	kepala	1.386294
26	rumah	1.386294	71	milik	1.386294
27	tenggorok	1.386294	72	gejala	0.693147
28	3	0.693147	73	inisiatif	1.386294
29	dengan	1.386294	74	corona	0.693147
30	suspect	1.386294	75	237	1.386294
31	aja	1.386294	76	kumulatif	1.386294
32	narasi	1.386294	77	infeksi	0.693147
33	hingga	1.386294	78	sama	1.386294
34	sembuh	1.386294	79	-	0
35	kristal	0.693147	80	langsung	1.386294
36	tebet	0	81	pasar	0
37	barat	0	82	test	1.386294
38	pasien	0.693147	83	dinkes	1.386294
39	lantai	1.386294	84	bawa	0.693147
40	news	1.386294	85	hebat	1.386294
41	mandiri	1.386294	86	karantina	1.386294
42	hasil	1.386294	87	virus	0.693147

43	karyawan	0.693147	88	jakarta	1.386294
44	sesak	1.386294	89	covid-19	0.693147
45	wajib	1.386294	90	ikut	1.386294

4.1.6 Term Frequency - Inverse Document Frequency

$$W_{ij} = tf_{ij} * idf_j$$

TF-IDF = the term frequency results times the inverse document frequency results.

Training Data :

Document 1 (Title) : pasar tebet barat tutup total karena ada dagang positif corona.

Document 2 (Title) : 2 hal ini sebabkan kasus positif corona di indonesia terus bertambah.

Testing Data :

Document 3 (Title) : pasien positif corona di jakarta kini damping psikolog.

The following table 4.21 is the results of the Term Frequency - Inverse Document Frequency section title.

Table 21: Term Frequency Result - Inverse Document Frequency (Title)

No.	Document	tf-idf
1	pasar tebet barat tutup total karena ada dagang positif corona	11.090355
2	2 hal ini sebabkan kasus positif corona di indonesia terus bertambah	9.010913
3	pasien positif corona di jakarta kini damping psikolog	9.010913

Training Data :

Document 1 (Content) : edar informasi yang sebut pasar tebet barat tutup lantar ada dagang papar virus corona atau wabah covid-19 dalam pesan itu juga sebut bahwa dapat dua orang karyawan toko yang positif dan 12 orang lain saspek ikut kutip narasi breaking news orang tampa gejala - otg n sama just info pasar tebet

barat tutup total karena milik toko kristal di lantai 2 positif dan tinggal kemarin lalu kepala pasar jaya tebet barat inisiatif untuk rapid test ke semua dagang pasar tebet barat hasil 2 orang karyawan toko kristal positif dan 12 orang suspect hebat ke-14 orang itu baik2 aja gak demam gak sakit tenggorok atau sesak nafas infeksi tanpa gejala akhir bawa langsung ke wisma atlet dan semua dagang pasar tebet barat wajib karantina mandiri di rumah bawah dinkes.

Document 2 (Content) : sampai dengan hari ini total kumulatif pasien infeksi virus corona covid-19 banyak 4 839 orang.

Testing Data :

Document 3 (Content) : hingga hari ini jumlah pasien positif covid-19 di jakarta capai 3 112 orang pasien yang nyata sembuh banyak 237 orang tinggal 297 orang.

The following table 4.22 is the results of the Term Frequency - Inverse Document Frequency content section.

Table 22: Term Frequency Result - Inverse Document Frequency (Content)

No.	Document	tf-idf
1	edar informasi yang sebut pasar tebet barat tutup lantar ada dagang papar virus corona atau wabah covid-19 dalam pesan itu juga sebut bahwa dapat dua orang karyawan toko yang positif dan 12 orang lain saspek ikut kutip narasi breaking news orang tampa gejala - otg n sama just info pasar tebet barat tutup total karena milik toko kristal di lantai 2 positif dan tinggal kemarin lalu kepala pasar jaya tebet barat inisiatif untuk rapid test ke semua dagang pasar tebet barat hasil 2 orang karyawan toko kristal positif dan 12 orang suspect hebat ke-14 orang itu baik2 aja gak demam gak sakit tenggorok atau sesak nafas infeksi tanpa gejala akhir bawa langsung ke wisma atlet dan semua dagang pasar tebet barat wajib karantina mandiri di rumah bawah dinkes.	89.415986

2	sampai dengan hari ini total kumulatif pasien infeksi virus corona covid-19 banyak 4 839 orang.	11.090355
3	hingga hari ini jumlah pasien positif covid-19 di jakarta capai 3 112 orang pasien yang nyata sembuh banyak 237 orang tinggal 297 orang.	18.714974

4.1.7 Define Term Frequency - Inverse Document Frequency Category

The stages of determining this category as additional data for the analysis phase in the use of the algorithm. In determining this category there are provisions:

1. tf-idf value is less than or equal to 0 give value 5
2. tf-idf value is more than 0 and less or equal to 50, give a value of 10
3. tf-idf value more than 50 and less or equal to 100 give a value of 15
4. tf-idf value more than 100 and less or equal to 150 give a value of 20
5. tf-idf value more than 150 and l less or equal to 200 give a value of 25
6. tf-idf value more than 200 and less or equal to 250, give a value of 30
7. tf-idf value more than 250 and less or equal to 300, give a value of 35
8. tf-idf value more than 300 and less or equal to 350, give a value of 40
9. tf-idf value more than 350 and less or equal to 400, give a value of 45
10. tf-idf value more than 400 and less or equal to 450, give a value of 50
11. tf-idf value more than 450 and less or equal to 500, give a value of 55
12. tf-idf value more than 500 and less or equal to 550, give a value of 60
13. tf-idf value more than 550 give it value 65

The following table 4.23 is the results of the Term Frequency category - Inverse Document Frequency title section and table 4.24 is the results of the Term Frequency category - Inverse Document Frequency content section.

Table 23: Results for the Term Frequency - Inverse Document Frequency category (Title)

No.	Document	tf-idf	Tf-idf Category
1	pasar tebet barat tutup total karena ada dagang positif corona	11.090355	10
2	2 hal ini sebabkan kasus positif corona di indonesia terus bertambah	9.010913	10
3	pasien positif corona di jakarta kini damping psikolog	9.010913	10

Table 24: Results for the Term Frequency - Inverse Document Frequency category (Content)

No.	Document	tf-idf	Tf-idf Category
1	edar informasi yang sebut pasar tebet barat tutup lantar ada dagang papar virus corona atau wabah covid-19 dalam pesan itu juga sebut bahwa dapat dua orang karyawan toko yang positif dan 12 orang lain saspek ikut kutip narasi breaking news orang tampa gejala - otg n sama just info pasar tebet barat tutup total karena milik toko kristal di lantai 2 positif dan tinggal kemarin lalu kepala pasar jaya tebet barat inisiatif untuk rapid test ke semua dagang pasar tebet barat hasil 2 orang karyawan toko kristal positif dan 12 orang suspect hebat ke-14 orang itu baik2 aja gak demam gak sakit tenggorok atau sesak nafas infeksi tanpa gejala akhir bawa langsung ke wisma atlet dan semua dagang pasar tebet barat wajib karantina mandiri di rumah bawah dinkes.	89.415986	15
2	sampai dengan hari ini total kumulatif pasien infeksi virus corona covid-19 banyak 4 839 orang.	11.090355	10
3	hingga hari ini jumlah pasien positif covid-19 di jakarta capai 3 112 orang pasien yang nyata sembuh banyak 237 orang tinggal 297 orang.	18.714974	10

4.1.8 Processing Data using Random Forest Classifier

Random Forest is a method of classifying large amounts of data. This method is a branch of the decision tree method. The process of this algorithm starts from randomly solving sample data in a decision tree. After a tree is formed the voting is done in each class from the sample data. Then combining the votes in each class from there can be seen the most votes.

There are several steps in processing data using Random Forest :

1. Divide data into several classes randomly
2. Then go to the decision tree method by checking the label on the data, if you have 2 labels, then the value is False, it will go to stage number 4. Conversely, if you have 1 label in 1 the data will be True and can be directly classified into stage number 3.
3. The classification stage, this stage is determined by the number of votes on each label. The most votes are the classification results.
4. If the value is False, then it goes to the stage of calculate for potential splits, this stage where search for the midpoint between the initial data and the final data.

The following example in table 4.25 is the result of Training Data after going through Pre-processing and TF-IDF. And table 4.26 is the result of Testing Data after going through Pre-processing and TF-IDF.

Table 25: Training Data Results after going through Pre-processing,TF-IDF and TF-IDF Category

No.	Document Training	tf-idf	Tf-idf Category
1	pasar tebet barat tutup total karena ada dagang positif corona	11.090355	10
2	2 hal ini sebabkan kasus positif corona di indonesia terus bertambah	9.010913	10

Potensial split tf-idf = (Document 1 (tf-idf) + Document 2 (tf-idf)) / 2

= (11.090355 + 9.010913)/2

= 10.050634

Potensial split tf-idf category = (10 +10) / 2 = 10

Table 26: Testing Data Results after going through Pre-processing,TF-IDF and TF-IDF Category

No.	Document	tf-idf	Tf-idf Category
1	edar informasi yang sebut pasar tebet barat tutup lantar ada dagang papar virus corona atau wabah covid-19 dalam pesan itu juga sebut bahwa dapat dua orang karyawan toko yang positif dan 12 orang lain saspek ikut kutip narasi breaking news orang tampa gejala - otg n sama just info pasar tebet barat tutup total karena milik toko kristal di lantai 2 positif dan tinggal kemarin lalu kepala pasar jaya tebet barat inisiatif untuk rapid test ke semua dagang pasar tebet barat hasil 2 orang karyawan toko kristal positif dan 12 orang suspect hebat ke-14 orang itu baik2 aja gak demam gak sakit tenggorok atau sesak nafas infeksi tanpa gejala akhir bawa langsung ke wisma atlet dan semua dagang pasar tebet barat wajib karantina mandiri di rumah bawah dinkes.	89.415986	15
2	sampai dengan hari ini total kumulatif pasien infeksi virus corona covid-19 banyak 4 839 orang.	11.090355	10

Potensial split tf-idf = (Document 1 (tf-idf) + Document 2 (tf-idf)) / 2

= (89.415986+ 11.090355)/2

= 50,2531705

Potensial split tf-idf category = (15 +10) / 2

= 12.5

5. Search for the best split to determine the center line of all available data. In calculate for the best split, it is necessary to search for minimum and maximum data values as well as determining overall entropy from the results of the minimum and maximum data values obtained.

To find the minimum and maximum data values needed between the overall data value with the results of the potential split value obtained previously.

Example :

Potensial split (title) : 10.050634

data minimal = data value overall tf-idf title section <= Potensial split (title)

= [11.090354888959125 9.010913347279288] <= 10.050634118119206

= 9.010913347279288

data maximal = data value overall tf-idf title section > Potensial split (title)

= [11.090354888959125 9.010913347279288] > 10.050634118119206

= 11.090354888959125

Potensial split (content) = 50,2531705

data minimal = data value overall tf-idf content section <= Potensial split (content)

= [89.41598629223293 11.090354888959125] <= 50.25317059059603

= 11.090354888959125

data maximal = data value overall tf-idf content section > Potensial split (content)

= [89.41598629223293 11.090354888959125] > 50.25317059059603

= 89.41598629223293

Total length of data = Minimum length of data array + Maximum length of data array

average minimum data length = Minimum array data length / Amount of data length.

average maximal data length = Maximum data array length / Amount of data length.

Overall entropy = (average minimum data length * minimum entropy data value) + (average maximum data length * maximum entropy data value).

Before calculating for overall entropy, it is necessary to calculate the value of entropy first.

Probabilitas = $jumlah\,label / \sum jumlah\,label$

Entropy = $\sum (probabilitas * -\log probabilitas)$

Then counts for overall entropy by using the entropy results obtained previously.

Overall entropy = (average minimum data length * entropy minimum data value) + (average maximal data length * entropy maximum data value)

average minimum data length = Minimum array data length / Total data length = 1/2 = 0.5

average maximum data length = Maximum data array length / Total data length = 1/2 = 0.5

Example :

Calculate Entropy part title:

Document 1 : Hoax , tf-idf : 11.090355 (Data maximum)

Document 2 : Real, tf-idf : 9.010913 (Data minimal)

Minimum Entropy data = $\sum (probabilitas * -\log probabilitas)$

= 1 * - log (1)

= 0.0

Maximum Entropy data = $\sum (probabilitas * -\log probabilitas)$

= 1 * - log (1)

= 0.0

Overall entropy = (0.5 * minimum entropy data) + (0.5 * maximum entropy data)

= (0.5 * 0.0) + (0.5 * 0.0)

= 0.0

Calculates Entropy content section :

Document 1 : Hoax , tf-idf : 11.090355 (Maximum Data)
Document 2 : Real, tf-idf : 9.010913 (Minimum Data)

Minimum Entropy data = $\sum(probabilitas * -\log probabilitas)$
= 1 * - log (1)
= 0.0

Maximum Entropy data = $\sum(probabilitas * -\log probabilitas)$
= 1 * - log (1)
= 0.0

Overall entropy = (0.5 * minimum entropy data) + (0.5 * maximum entropy data)
= (0.5 * 0.0) + (0.5 * 0.0)
= 0.0

6. Finally to the classification stage, using the minimum data value and the maximum data value previously calculated. With n-trees that are 2 and max depth 3. n-trees are the number of branches in a tree, while max depth is the maximum number of branch depths.

Result :

{'tf_idf <= 10.0506341181': [u'REAL', u'HOAX']}

explanation: the first branch produces 10.0506341181. is the split potential value previously calculated. If tf_idf <= 10.0506341181 is true then the result is REAL, if wrong then the result is HOAX.

To get prediction results, compare with the trained tree using training data. The value of tf-idf in the testing data is worth 9.010913 **then the prediction results of testing the title data section are Real. Because the value of 9.010913 is less than 10.0506341181 it is labeled REAL.**

Whereas the content section, produces branches:

{'tf_idf <= 50.2531705906': [u'REAL', u'HOAX']}

Explanation: the first branch produces 50.2531705906. is the split potential value previously calculated. If tf_idf <= 50.2531705906 is true then the result is REAL, if wrong then the result is HOAX.

Testing data on tf-idf is worth 18.714974, showing that the value of tf-idf on testing data is less than 50.2531705906, **then the predicted result of the content section is REAL.**

4.1.9 Processing Data using Support Vector Machine Classifier

Support Vector Machine is a method used in classification. Linear support vector machine is a linear kernel in the support vector machine algorithm where hyperplane lines are usually straight in the shape of dividing two classes (Maulina Dina, Sagara Rofie, 2018). Hyperplane lines are lines that divide both two vector groups and are usually valued $w*x-b=0$ and margin value $w*x-b=1$ or $w*x-b=-1$. SVM linear equation (Maulina Dina, Sagara Rofie, 2018).

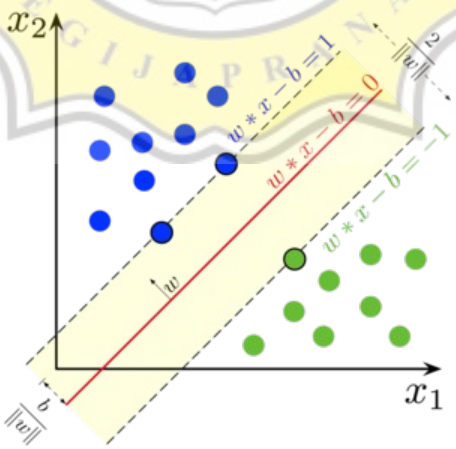

Illustration 1: SVM images with maximum hyperplane margins

Linear SVM :

$w.x-b=0$

$w.x-b\geqslant 1$ if $y_i=1$

$y_i w.x - b \geqslant 1$

$w.x_i - b \leqslant -1$ if $y_i = -1$

$-y_i w.x_i - b \leqslant -1$

Where :

w = hyperplane parameter sought

x = SVM input data

b = hyperplane parameters sought (bias)

Hinge Loss :

$max(0, 1 - y_i(w.x_i - b))$

$f(x) = w.x_i - b$

if $y * f(x) \geqslant 1$

0

else :

$1 - y * f(x)$

Regularization Technique :

R = $\lambda \|w\|^2 + \frac{1}{n} f(x) \sum_{i=1}^{n} (0, 1 - y_i(w.x_i - b))$

Where :

w = hyperplane parameter sought

x = SVM input data

b = hyperplane parameters sought (bias)

y_i = parameter y index

λ = lambda parameter

if $y * f(x) \geqslant 1$:

$\lambda \|w\|^2$

else :

$\lambda\|w\|^2+1-y_i(w.x_i-b)$

Gradient :

if $y*f(x)\geqslant 1$:

$$\frac{da}{dw}=2\lambda\|w\|$$

else :

$$\frac{da}{dw}=2\lambda\|w\|-y_i x_i$$

$$\frac{da}{db}=y_i$$

for each training sample x_i :

$w=w-\alpha*dw$

$b=b-\alpha*db$

Gradient Update :

if $y*f(x)\geqslant 1$:

$w=w-\alpha*2\lambda w$

else

$w=w-\alpha*(2\lambda w-y_i x_i)$

$b=b-\alpha*y_i$

dimana :

w = Hyperplane parameters are sought

b = Hyperplane parameters sought (biased)

a = learning rate

dw = dR/dw

db = dR/db

x_i = input data for each index

Next calculate the gradient of the training data. Gradient formula as above is $y*f(x)\geqslant 1$ then create conditions where if the result is more than one or equal to 1 the results will go into the formula $w=w-a*2\lambda w$ conversely if not less than one then enter the process $w=w-a*(2\lambda w-y_i x_i)$ and $b=b-a*y_i$

Training Data :

Document 1 (Title) : pasar tebet barat tutup total karena ada dagang positif corona

Document 2 (Title) : 2 hal ini sebabkan kasus positif corona di indonesia terus bertambah

Testing Data :

Document 3 (Title) : pasien positif corona di jakarta kini damping psikolog

The following table 4.27 contains all the documents and their labels, table 4.28 contains the results of calculations on the title of each document which constitutes traning data. Table 4.29 is the result of prediction on the title part testing data using the Support Vector Machine algorithm.

Table 27: Document tables with labels

Document	Label
Document 1 (Training Data)	Hoax
Document 2 (Training Data)	Real
Document 3 (Training Data)	Real

Table 28: Results of the calculation table for each Training data

Document (Training Data)	$y*w.x_i-b$	$y*w.x_i-b\geqslant 1$
Document 1	0.0	False
Document 2	-0.010093422689498588	False

Table 29: Table Prediction calculation results

Document (Testing Data)	$w.x-b$	**Prediction**
Document 3	-0.00177378	Hoax

Result: -0.00177378. If the value is 0, then the value is Hoax, if the value is more than 1, the value is Real. **Then the result of document 3 in the title analysis using the SVM algorithm is HOAX.**

Analysis ‘Content’

Example :

Document 1 (Content) : edar informasi yang sebut pasar tebet barat tutup lantar ada dagang papar virus corona atau wabah covid-19 dalam pesan itu juga sebut bahwa dapat dua orang karyawan toko yang positif dan 12 orang lain saspek ikut kutip narasi breaking news orang tampa gejala - otg n sama just info pasar tebet barat tutup total karena milik toko kristal di lantai 2 positif dan tinggal kemarin lalu kepala pasar jaya tebet barat inisiatif untuk rapid test ke semua dagang pasar tebet barat hasil 2 orang karyawan toko kristal positif dan 12 orang suspect hebat ke-14 orang itu baik2 aja gak demam gak sakit tenggorok atau sesak nafas infeksi tanpa gejala akhir bawa langsung ke wisma atlet dan semua dagang pasar tebet barat wajib karantina mandiri di rumah bawah dinkes.

Document 2 (Content) : sampai dengan hari ini total kumulatif pasien infeksi virus corona covid-19 banyak 4 839 orang.

Testing Data :

Document 3 (Content) : hingga hari ini jumlah pasien positif covid-19 di jakarta capai 3 112 orang pasien yang nyata sembuh banyak 237 orang tinggal 297 orang.

The following table 4.30 contains all the documents and their labels, table 4.31 contains the results of calculations on the contents of each document which constitutes traning data. Table 4.32 is the result of prediction on the testing data part of the content using the Support Vector Machine algorithm.

Table 30: Document tables with labels

Document	**Label**
Document 1 (Training Data)	Hoax
Document 2 (Training Data)	Real
Document 3 (Training Data)	Real

Table 31: The results of the calculation table for each Training data

Document (Training Data)	$y*w.x_i-b$	$y*w.x_i-b\geqslant 1$
Document 1	0.0	False
Document 2	-0.11426550207271675	False

Table 32: Table Prediction calculation results

Document (Training Data)	$w.x-b$	**Prediction**
Document 3	-0.15158628	Hoax

Result: -0.15158628.If the value of 0 is Hoax, if the value of more than 1 is Real. **Then the result of document 3 in content analysis using SVM algorithm is HOAX.**

4.1.10 Accuracy measurement Random Forest and Support Vector Machine

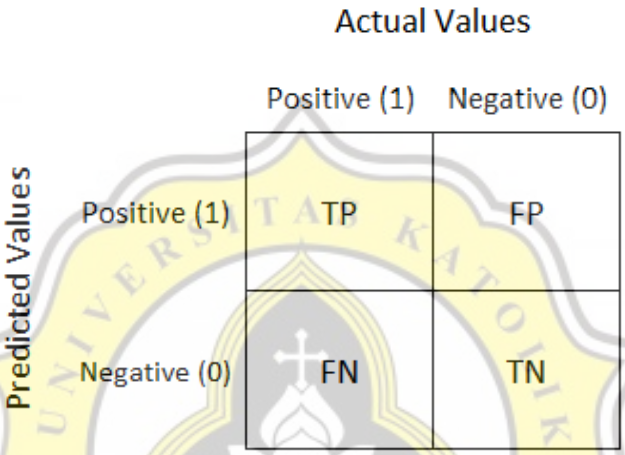

Illustration 2: Illustration Confusion Matrix

Explanation :

TP: True Positive (the actual document is True but the prediction results are True too).

FP: False Positive (documents that are actually True but predictions are False).

TN: True Negative (documents that are actually False but predictions are True).

FN: False Negative (the document is actually False but predictions are False).

a. $Accuracy = \frac{tp+tn}{tp+tn+fp+fn}$

b $Precision = \frac{tp}{tp+fp}$

c. $Recall = \frac{tp}{tp+fn}$

d. $F1-Score = \frac{2 \times recall \times precision}{recall + precision}$

Testing Data :

Document 3 (Title) : pasien positif corona di jakarta kini damping psikolog

Document 3 (Content) : hingga hari ini jumlah pasien positif covid-19 di jakarta capai 3 112 orang pasien yang nyata sembuh banyak 237 orang tinggal 297 orang.

The following table 4.33 is the result of the Confusion Matrix on the Support Vector Machine algorithm, while table 4.34 is the result of the Confusion Matrix on the Random Forest algorithm. Table 4.35 is a comparison of accuracy, precision, recall, f1-score of the two algorithms.

Table 33: Table Confusion Matrix Results on Support Vector Machine algorithm (Title and Content)

TP = 0	FP = -
FN = -	TN = -

Table 34: Table Confusion Matrix Results on Random Forest algorithm (Title and Content)

TP = 1	FP = -
FN = -	TN = -

Table 35: Result Performance (Title and Content)

Algorithm	Accuracy	Precision	Recall	F1-Score
Support Vector Machine	0 %	0 %	0 %	0 %
Random Forest	100 %	100 %	100 %	100 %

4.2 Desain

4.2.1 Flow Chart

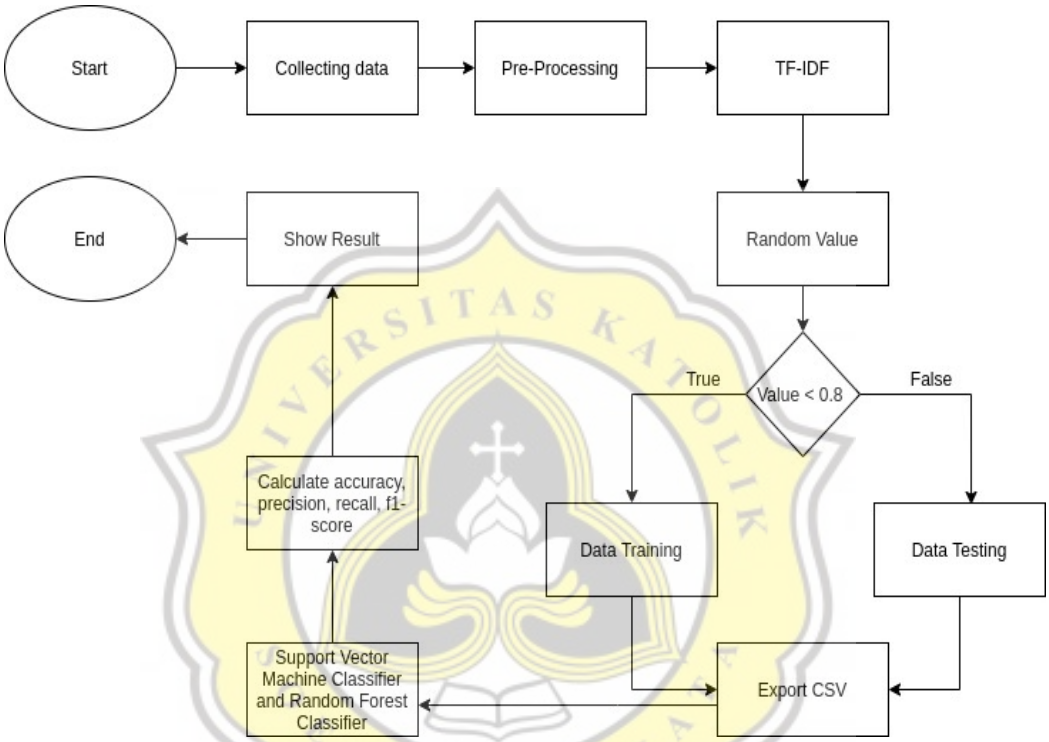

Illustration 3: Flow Chart

In image 4.3 above is a flow chart diagram of the course of the program starts from the process of collecting data. The process aims to extract data from online news websites in Indonesia. Online news such as liputan6.com, tribunnews.com, BBC.com and turnbackhoax.id. In taking data from the website turnbackhoax.id using the API, while the rest use the BeautifulSoap library from python for web scrapping techniques.

After getting all the news data. The data will experience data processing which is often called the pre-processing stage. The pre processing stages include folding, stemming, filtering, tokenizing. Also in data processing the researchers added the process of changing to ASCII type, and deleting characters in each sentence.

Then the next step is to calculate TF-IDF. Calculating TF-IDF as explained in chapter 3, TF-IDF is the result of the weighting of each document. Before calculating for TF-IDF results, there are several steps that must be carried out such as calculating the term frequency, then the document frequency, then calculate for the Inverse Document Frequency. And finally calculate TF-IDF.

The next step is determining training data and testing data. To divide the training data by 80% portion and 20% for testing data, all data are given a temporary number with a random value between 0-1. If the document or data is less than 0.8 then enter the training data, then if more than 0.8 then enter the testing data. After everything is divided into training data and testing data, all training and testing data are stored in CSV format.

Then enter the algorithm stage, the Support Vevtor Machine and Random Forest algorithm. The two have differences in the way of processing, but have the same goal of being able to predict the data hoax or facts.

After obtaining the prediction results from each use of both algorithms. As the example in table 4.35 the prediction results in the Support Vector Machine get an accuracy of 0% because they cannot predict the document correctly while Random Forest gets an accuracy of 100% because it can predict the document correctly.

4.2.2 Use Case

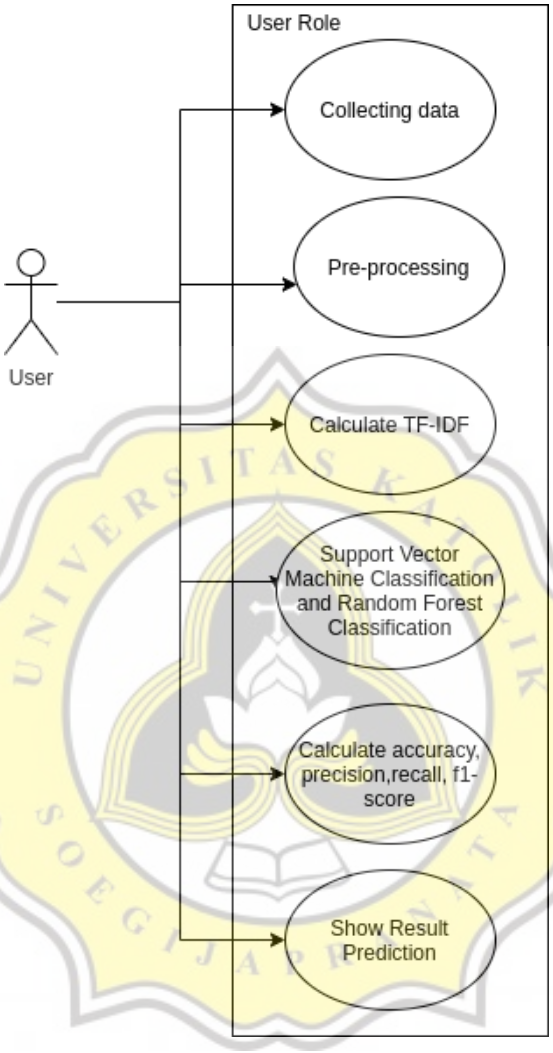

Illustration 4: Use Case Diagram

In the image 4.4 is the use case of the role user in running a program. As shown, the user has 7 accesses that can collect data, pre-processing, calculate tf-idf, calculate Support Vector Machine Classification and Random Forest Classification processes. the user can compare calculate accuracy between the two algorithms. And finally get the prediction results from both algorithm.