

## CHAPTER 3

### RESEARCH METHODOLOGY

#### 3.1 Data Collection

The data used in this study uses twitter data taken by the scrapping method. By using the python scrapping method, twitter data that has a socialdistancing hashtag is taken sufficiently in accordance with the data to be used, namely 150 training data and 150 test data which will later be tested with different schemes.

Scraping itself is a process for retrieving data from a web. Because the data is taken from twitter media, an API is needed to be able to access the data to be scraped. By registering with the Twitter web developer to get the required API. If you are using this API for study purposes then Twitter might speed up our process to get the required API.

This scraping method uses python code to get data. By including the hashtags needed in the python program and how much data will be retrieved. This process does not take a long time depending on how much data you want to retrieve. With this method it is easier to get the required data. Here are some examples of the data used in this project.

- While we're all supposed to be #socialdistancing, I also encourage you to distance yourself from negativity

- @Roblox: Stay cozy. Stay healthy. Stay home. #SocialDistancingRaise awareness stylishly with these comfy, classy accessories.

- 'Only BJP MLA can organise HEALTH workers meeting and not follow any #SocialDistancing atleast least practice

- ' @Hyundai\_Global: #becauseofyou and because of all of us we must continue to play our part in maintaining #socialdistancing.

- 'Spring time walk on campus while practicing #SocialDistancing A Fellow graduate student riding his bike in the back

```

time text \
0 2020-04-25 07:01:57 b"Dear people's of #Uttarakhand\n As We are al...
1 2020-04-25 07:01:50 b'For real for real #SocialDistancing https://...
2 2020-04-25 07:01:49 b'Thank you to @POPSUGAR for naming We Are Ani...
3 2020-04-25 07:01:47 b'RT @emmieluw0: Make her go viral We want an...
4 2020-04-25 07:01:46 b'RT @DrNighatArif: Iftari is SUCH an importan...
.. ...
145 2020-04-25 06:49:28 b'Goodnight from Mekka & Downtown Los Angel...
146 2020-04-25 06:49:09 b'So... is the nhs on Westminster bridge? Why ...
147 2020-04-25 06:49:06 b'One of the special moments after this lockdo...
148 2020-04-25 06:48:47 b'#fbf to some #sunworship & #socialdistanc...
149 2020-04-25 06:48:47 b"RT @imvaishalisingh: Here's a video where #S...

sentiment
0 Positive
1 Negative
2 Positive
3 Negative
4 Positive
.. ...
145 Positive
146 Negative
147 Positive
148 Negative
149 Negative
[150 rows x 3 columns]

```

Illustration 3.1.1: Collected Data Train

```

text
0 '@flwrgardengifts: Smiles will be popping fro...
1 'Free collaboration tool for Education. Learni...
2 '@TheOfficialSBI: Banking is essential and so i...
3 '@anandmahindra @rajesh664 You are great enoug...
4 @enews Staying at home is like opening a parac...
.. ...
143 'Place Thiruvanniyur market\n\nTamilnadu gov...
144 'RT @Hyundai_Global: #becauseofyou and because...
145 'RT @Hyundai_Global: Join Hyundai and BTS @BTS...
146 'E-commerce companies aren't allowe...
147 'RT @brucetempleton: Social distancing at Coog...

[148 rows x 1 columns]

```

Illustration 3.1.2: Collected Data Test

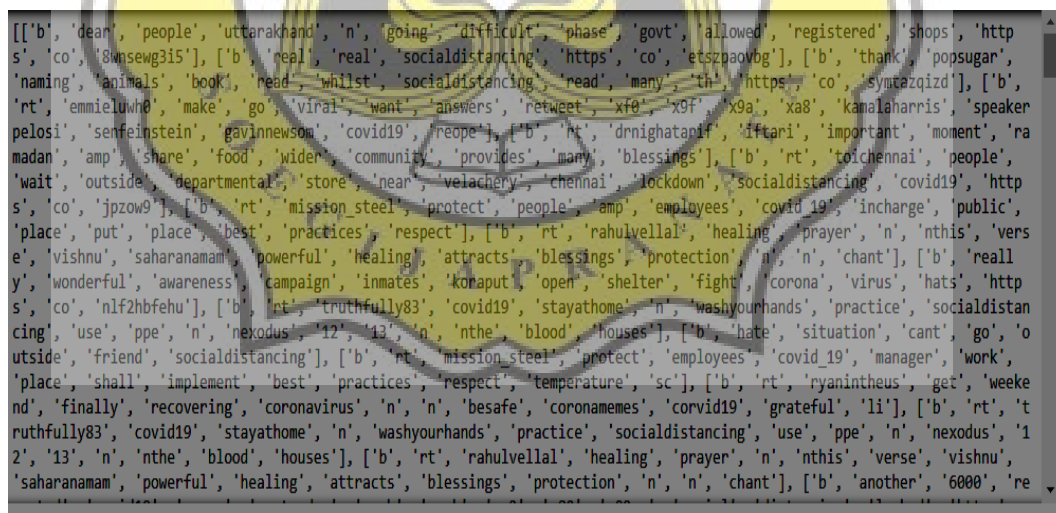
## 3.2 Preprocessing Data

After the data is obtained, this data cannot be used because it is difficult to process. therefore preprocessing data is needed to simplify the processing process. In this research, several methods are used, namely tokenization, removing stop words, case folding and stemming. This project is made in a jupyter notebook and uses the nltk library for the stopwords and tokenization processes.

The preprocessing method itself has several steps to be able to preprocess a data. Starting with case folding, is a part of converting a text data into regular letters (Lowering). Usually data is inconsistent, that is, it starts with a capital letter or there is a capital letter in the middle of the word. Therefore the case folding process is needed to equalize the data so that it is easier to process later.

Tokenization itself is the process of changing a long text data into words to make it easier to process data later. Broadly speaking, tokenization is the process of solving a set of characters in a text into word units.

After doing the above process, the process of removing stop words is carried out. This process is to eliminate words that are often-used such as the word "I", "you", "being", "from", "who" and so on. Removing this stopword reduces the index size and processing time.



```
[[['b', 'dear', 'people', 'uttarakhand', 'n', 'going', 'difficult', 'phase', 'govt', 'allowed', 'registered', 'shops', 'http', 's', 'co', '8wmswg3i5'], ['b', 'real', 'real', 'socialdistancing', 'https', 'co', 'etszpaovbg'], ['b', 'thank', 'popsugar', 'naming', 'animals', 'book', 'read', 'whilst', 'socialdistancing', 'read', 'many', 'th', 'https', 'co', 'symtazqzdz'], ['b', 'rt', 'emmiellwtd', 'make', 'go', 'viral', 'want', 'answers', 'retweet', 'xf0', 'x9f', 'x9a', 'xa8', 'kamalaharris', 'speaker', 'pelosi', 'senfeinstein', 'gavinnewson', 'covid19', 'reope'], ['b', 'rt', 'drnighatanif', 'iftari', 'important', 'moment', 'ra', 'madan', 'amp', 'share', 'food', 'wider', 'community', 'provides', 'many', 'blessings'], ['b', 'rt', 'toichennai', 'people', 'wait', 'outside', 'departmental', 'store', 'near', 'velachery', 'chennai', 'lockdown', 'socialdistancing', 'covid19', 'http', 's', 'co', 'jpoz09'], ['b', 'rt', 'mission steel', 'protect', 'people', 'amp', 'employees', 'covid 19', 'incharge', 'public', 'place', 'put', 'place', 'best', 'practices', 'respect'], ['b', 'rt', 'raahulvella', 'healing', 'prayer', 'n', 'nthis', 'verse', 'vishnu', 'saharanamam', 'powerful', 'healing', 'attracts', 'blessings', 'protection', 'n', 'n', 'chant'], ['b', 'reall', 'y', 'wonderful', 'awareness', 'campaign', 'inmates', 'konaput', 'open', 'shelter', 'fight', 'corona', 'virus', 'hats', 'http', 's', 'co', 'nlf2hbfehu'], ['b', 'rt', 'truthfully83', 'covid19', 'stayathome', 'n', 'washyourhands', 'practice', 'socialdistan', 'cing', 'use', 'ppe', 'n', 'nexodus', '12', '13', 'n', 'nthe', 'blood', 'houses'], ['b', 'hate', 'situation', 'cant', 'go', 'o', 'outside', 'friend', 'socialdistancing'], ['b', 'rt', 'mission steel', 'protect', 'employees', 'covid 19', 'managen', 'work', 'place', 'shall', 'implement', 'best', 'practices', 'respect', 'temperature', 'sc'], ['b', 'rt', 'ryanintheus', 'get', 'weeke', 'nd', 'finally', 'recovering', 'coronavirus', 'n', 'n', 'besafe', 'coronamemes', 'corvid19', 'grateful', 'li'], ['b', 'rt', 't', 'ruthfully83', 'covid19', 'stayathome', 'n', 'washyourhands', 'practice', 'socialdistancing', 'use', 'ppe', 'n', 'nexodus', '1', '2', '13', 'n', 'nthe', 'blood', 'houses'], ['b', 'rt', 'raahulvella', 'healing', 'prayer', 'n', 'nthis', 'verse', 'vishnu', 'saharanamam', 'powerful', 'healing', 'attracts', 'blessings', 'protection', 'n', 'n', 'chant'], ['b', 'another', '6000', 're
```

Illustration 3.2.1: Case Folding, Removing Stopwords, Filtering

End with a stemming process. It is a process for grouping other words which have the same basic word and meaning but have different forms or forms because they have different affixes. Broadly speaking, this process converts data from the tokenization process into basic

words in order to make it easier to group data and reduce the number of different indexes of a document..

```
[['b', 'dear', 'peopl', 'uttarakhand', 'n', 'go', 'difficult', 'phase', 'govt', 'allow', 'regist', 'shop', 'http', 'co', '8wn
sewg3i5'], ['b', 'real', 'real', 'socialdistanc', 'http', 'co', 'etszpaovbg'], ['b', 'thank', 'popsugar', 'name', 'anim', 'bo
ok', 'read', 'whilst', 'socialdistanc', 'read', 'mani', 'th', 'http', 'co', 'symtazqizd'], ['b', 'rt', 'emmieluw0', 'make',
'go', 'viral', 'want', 'answer', 'retweet', 'xf0', 'x9f', 'x9a', 'xa8', 'kamalaharri', 'speakerpelosi', 'senfeinstein', 'gavi
nnewsom', 'covid19', 'reop'], ['b', 'rt', 'drnighatarif', 'iftari', 'import', 'moment', 'ramadan', 'amp', 'share', 'food', 'w
ider', 'commun', 'provid', 'mani', 'bless'], ['b', 'rt', 'toichennai', 'peopl', 'wait', 'outsid', 'department', 'store', 'nea
r', 'velacheri', 'chennai', 'lockdown', 'socialdistanc', 'covid19', 'http', 'co', 'jpoz09'], ['b', 'rt', 'mission_steel', 'pr
otect', 'peopl', 'amp', 'employe', 'covid_19', 'incharg', 'public', 'place', 'put', 'place', 'best', 'practic', 'respect'],
['b', 'rt', 'rahulvel', 'heal', 'prayer', 'n', 'nthi', 'vers', 'vishnu', 'saharanamam', 'power', 'heal', 'attract', 'bless',
'protect', 'n', 'n', 'chant'], ['b', 'realli', 'wonder', 'awar', 'campaign', 'inmat', 'koraput', 'open', 'shelter', 'fight',
'corona', 'viru', 'hat', 'http', 'co', 'nlf2hbfehu'], ['b', 'rt', 'truthfully83', 'covid19', 'stayathom', 'n', 'washyourhan
d', 'practic', 'socialdistanc', 'use', 'ppe', 'n', 'nexodu', '12', '13', 'n', 'nthi', 'blood', 'hous'], ['b', 'hate', 'situa
t', 'cant', 'go', 'outsid', 'friend', 'socialdistanc'], ['b', 'rt', 'mission_steel', 'protect', 'employe', 'covid_19', 'mana
g', 'work', 'place', 'shall', 'implement', 'best', 'practic', 'respect', 'temperatur', 'sc'], ['b', 'rt', 'ryanintheu', 'ge
t', 'weekend', 'final', 'recov', 'coronaviru', 'n', 'n', 'besaf', 'coronamem', 'corvid19', 'grate', 'li'], ['b', 'rt', 'truth
fully83', 'covid19', 'stayathom', 'n', 'washyourhand', 'practic', 'socialdistanc', 'use', 'ppe', 'n', 'nexodu', '12', '13',
'n', 'nthi', 'blood', 'hous'], ['b', 'rt', 'rahulvel', 'heal', 'prayer', 'n', 'nthi', 'vers', 'vishnu', 'saharanamam', 'powe
r', 'heal', 'attract', 'bless', 'protect', 'n', 'n', 'chant'], ['b', 'anoth', '6000', 'report', 'covid19', 'case', 'yesterda
y', 'much', 'week', 'xe2', 'x80', 'x99ve', 'social', 'distanc', 'lock', 'http', 'co', 'ysqezbzd1'], ['b', 'work', 'home', 's
```

Illustration 3.2.2: Stemming

### 3.3 TF-IDF

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

$tf_{x,y}$  = frequency of  $x$  in  $y$   
 $df_x$  = number of documents containing  $x$   
 $N$  = total number of documents

Illustration 3.3 : TF-IDF

After preprocessing the data, it is continued by calculating the weight of each word using the TF-IDF method. Term frequency (TF) itself is the frequency of the appearance of a term in the document concerned. The greater the TF value, in a document, the greater the weight or the greater the conformity value will be.

Meanwhile, the IDF is a calculation of how the terms are widely distributed in the collection of documents concerned. The fewer the number of documents containing the term in question, the greater the idf value. Here is an example of a simple tf-idf calculation.

Query 1 = "hello, i hate socialdistancing because i wan't to play with my friends"

Query 2= "be happy on corona virus stay socialdistancing"



data which will later be tested with 3 schemes. Using 150 train data and 50 test data, 100 training data and 98 testing data, 50 training data and 148 testing data.

The data scheme is made like that because later to test how the maximum performance is against the data scheme above. If only tested in one scheme, it is likely that the analysis obtained is less than optimal, therefore a scheme like the one above is made. Training data has been manually labeled beforehand. A total of 150 training data have been labeled positive or negative to be predicted using testing data.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	time,text,sentiment																			
2	2020-04-25 07:01:57,"b"	"Dear people's of #Uttarakhand\n	As we are all going through a very difficult phase, Govt. allowed registered Shops https://t.co/8wNsewg3i5***,	Positive																
3	2020-04-25 07:01:50,"b"	For real for real #SocialDistancing https://t.co/ETs7dQyB8,	Negative																	
4	2020-04-25 07:01:49,"b"	Thank you to @PDPUSGARN for naming We Are Animals as a book to read whilst #SocialDistancing I read so many of th https://t.co/SyMTAZQId',	Positive																	
5	2020-04-25 07:01:47,"b"	RT @emmieluh70: Make her go viral We want answers RETWEET \xf0\x9f\x9a\x8a @Kamalaharris @SpeakerPelosi @SenFeinstein @GavinNewsom #COVID19 #Reope',	Negative																	
6	2020-04-25 07:01:46,"b"	RT @DilipalAut: There is SUCH an important moment in #Ramadan &amp; to share your food with the wider community provides many blessings. How,	Positive																	
7	2020-04-25 07:01:41,"b"	RT @TOChemai: People wait outside a departmental store near Velachery in Chennai #lockdown #SocialDistancing #COVID19 https://t.co/PZOW9',	Negative																	
8	2020-04-25 employees from #Covid_19, Incharge of a Public Place to put in place #Best Practices in respect",	Positive																		
9	2020-04-25 07:01:35,"b"	RT @RahulVella: Healing Prayer \n\nThis verse from Vishnu Saharanam is powerful in healing and attracts blessings and protection. \n\n#chant',	Positive																	
10	2020-04-25 07:01:29,"b"	Really wonderful awareness campaign by the inmates of Koraput Open Shelter to fight against Corona virus. Hats off https://t.co/NIF2hbFehU',	Positive																	
11	2020-04-25 07:01:27,"b"	RT @Truthfully63: On covid19 #stayAtHome\n\nWash your hands practice #social distancing, use PPE. \n\nExodus 12:13\n\nThe blood on the houses where",	Positive																	
12	2020-04-25 07:01:17,"b"	I hate this situation, I cant go outside with my friend #socialdistancing",	Negative																	
13	2020-04-25 07:01:06,"b"	RT @mission_steel: To protect employees from #Covid_19, Manager of a work place shall implement best practices in respect of temperature sc",	Positive																	
14	2020-04-25 07:01:03,"b"	RT @Ryaniotheus: When you get to the weekend after finally recovering from the #coronavirus! \n\n#BeSafe #coronamemes #Corvid19 #grateful #U',	Positive																	
15	2020-04-25 07:01:03,"b"	RT @Truthfully63: On covid19 #stayAtHome\n\nWash your hands practice #social distancing, use PPE. \n\nExodus 12:13\n\nThe blood on the houses where",	Positive																	
16	2020-04-25 07:01:00,"b"	RT @RahulVella: Healing Prayer \n\nThis verse from Vishnu Saharanam is powerful in healing and attracts blessings and protection. \n\n#chant',	Positive																	
17	2020-04-25 07:01:00,"b"	Another 600+ reported Covid19 cases yesterday - and much the same all week. We've been social distancing and locked https://t.co/y5sqEzBZdU1',	Negative																	
18	2020-04-25 07:00:47,"b"	Working from home is a solution for small businesses amid COVID-19 Social Distancing. Read more: https://t.co/OJfM6vLUl',	Positive																	
19	2020-04-25 07:00:44,"b"	We hope you've all having a relaxing weekend, and enjoying the sunshine safely and responsibly! Dont forget, we all https://t.co/a3jg8BRdZo",	Positive																	
20	2020-04-25 07:00:43,"b"	India Fights Corona\n\nGuidelines for journalists: during #COVID19\n\n#stayAtHome #CoronaVirus #socialDistancing https://t.co/y2s80qk8E',	Negative																	
21	2020-04-25 07:00:41,"b"	Hope this #shutdown ends soon \n\nAll this #social distancing means I've had to be more social \n\nGo to a store... line https://t.co/aMyrDonBLJ***,	Negative																	
22	2020-04-25 07:00:35,"b"	RT @Geeta_Mohan: Easing restrictions in India for the first time since #lockdown.\n\n#MHA order allows shops registered under Shops &amp; Established',	Positive																	

Illustration 3.4.1 : data train

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	@fwwrga:dengits: Smiles be popping from a distance with this Ghoom, Door Wreath \xf0\x9f\x9b\x89#wreath #wreathofinstagram #name #happiness #C',																			
2	Free collaboration tool for Education. Learning from any where with free video conferencing and collaboration software https://t.co/NbrNGHFBK',																			
3	@TheOfficialSBI: Banking is essential and so is your safety. Our team members from Gujarat, Telangana, Jharkhand and Uttar Pradesh encou',																			
4	@anandmahindra @rajes3664 You are great enough to ask sr. management to make this auto driver (innovator) as a visor',																			
5	@news Staying at home is like opening a parachute. The parachute will slow your fall but you don't take off the',																			
6	People wait outside a departmental store near Velachery in Chennai #lockdown #social distancing #COVID19 https://t.co/PZOW9',																			
7	@AisAmbBKK: Let's practice #social distancing by keeping 1.5 metres distance or a kangaroo, two koalas or an emu apart. https://t.co/V7PP',																			
8	@IOIBengaluru: 15 new Covid-19 positive cases reported in Karnataka from 5pm, April 24 to 12pm, April 25. Total number of cases in the s',																			
9	Only BJP MUA can organise HEALTH workers meeting and not follow any #social distancing atleast practice what',																			
10	@Lofis: Herbs Answer: RamayanHerbs Herbs Skin Care #lawlessSkin #COVID19 #CoronaVirus #quarantine',																			
11	@TOChemai: Tamil Nadu: Locals throng Kotdawal Chavadi market in Chennai to purchase ration ahead of lockdown announced by the state go',																			
12	@Roblox: Stay cozy. Stay healthy. Stay home. #social distancing raise awareness stylishly with these comfy, classy accessories. Available',																			
13	While we're all supposed to be #social distancing, I also encourage you to distance yourself from negativity, toxic',																			
14	@KTPNoorHossain: Day 30 #Rohingya camp #lockdown Start of Ramadan. Last year we held a big iftar party for Tola Tola people in the cam',																			
15	@Kat_McNamara: Say cheese...and wine? \xf0\x9f\x9a\x7d\x9f\x9b\x8d\x9f\x9b\x8d\x9f\x9b\x8d #social distancing #stayhome https://t.co/YCduCb2AYC',																			
16	Spring time walk on campus while practicing #social distancing A Fellow graduate student riding his bike in the back',																			
17	@SamiHeugan: You're a great man. It's almost holy. lol a light shines through you. And even your looks: So easy on',																			
18	@xycheesa: So sorry that exactly when you are set free you need to practice #social distancing! I wish you all the best!',																			
19	@RRPSpeaks: If companies are ready to implement strict #social distancing &amp; comply with health advisories, Central &amp; State Govt. should',																			
20	@DaviesBooks: Day 27 of Social Distancing \xe2\x9d\xa4\xe1\xb8\x8fSound up! https://t.co/1TYe1Rn4S6 via @HolyCow_Inc #gats #humor #defstar5 #smile #FunniestT',																			
21	'FYI in case if',																			
22	@ajaydevgnec: All registered shops under Shops &amp; Establishment Act of respective States/ UTs, including shops in residential complexes',																			

Illustration 3.4.2 : data test

### 3.5 Processing Data

This study uses two different algorithms, namely the vector space model using cosine similarity and naive bayes for comparison. Which will be evaluated for comparison which is more effective than three different training data trials and testing data.

The vector space model itself is an algebraic model that depicts a text document into a vector which is usually used to rank relevance. Documents in the vector space model are in the form of a matrix containing the weights of all words in each document. The weight states the importance or contribution of words to a document and a collection of documents. To get the distance or document similarity value, you can use cosine similarity as in this project. Cosine similarity is a method of measuring the closest similarity, by calculating the angle between the document vector and the query vector. If the vector is a unit of length, the cosine of the angle between them is simply the dot product of the vector. Which will produce values 0 to 1. Where 0 indicates that the documents are not similar at all and 1 indicates that the documents are completely identical.

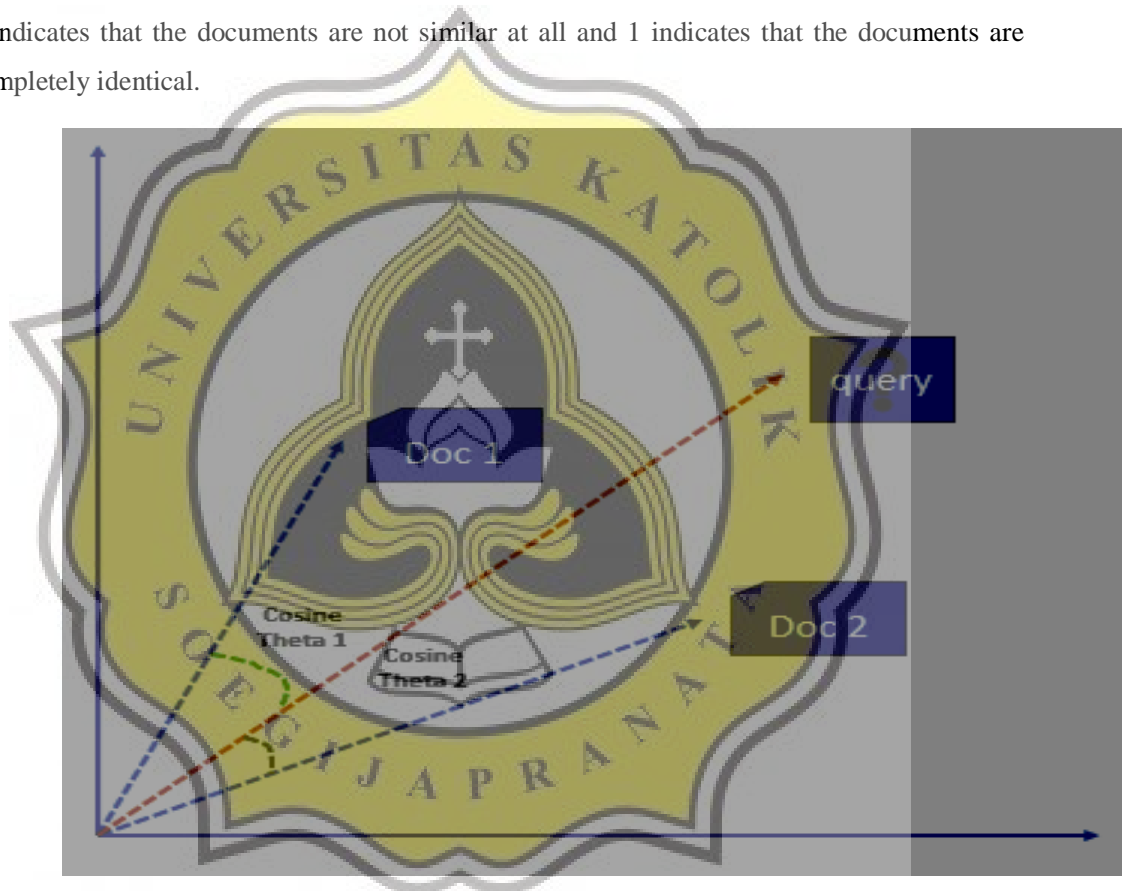


Illustration 3.4 : Vector Space Model

### 3.6 Performance Analysis

The last method used is to calculate the performance of each algorithm used. The way to evaluate the algorithm that has been used is by calculating accuracy, precision, recall, and

f1-score. how to get accuracy, precision, recall, and f1-score by manually calculating TP, TN, FP, FN.

TP is positive true where the prediction result is labeled positive from the program as desired (TRUE). Meanwhile, TN is true negative, it is the result of negative prediction and really negative (TRUE). FP is false positive, which is where the program predicts that the document is positive but if checked manually it is labeled negative (FALSE). Same is the case with FN, where the prediction result is negative but if checked manually it should be labeled positive (FALSE).

After calculating TP, TN, FP, FN, the performance of each algorithm can be calculated by calculating accuracy, precision, recall and f1-score. accuracy is the ratio of positive and negative true predictions (TRUE) to the overall data. The higher the accuracy, the greater the success ratio for predicting positive and negative labels. As for precision, it is the ratio of positive predictions (TRUE) to the overall positive predicted results. Precision answers the question "what percentage of positive tweets are predicted to be positive". Then recall is the ratio of true positive predictions compared to the overall true positive data. Recall answers the question "what percentage of tweets are predicted to be labeled positive compared to all tweets that are actually labeled positive".

