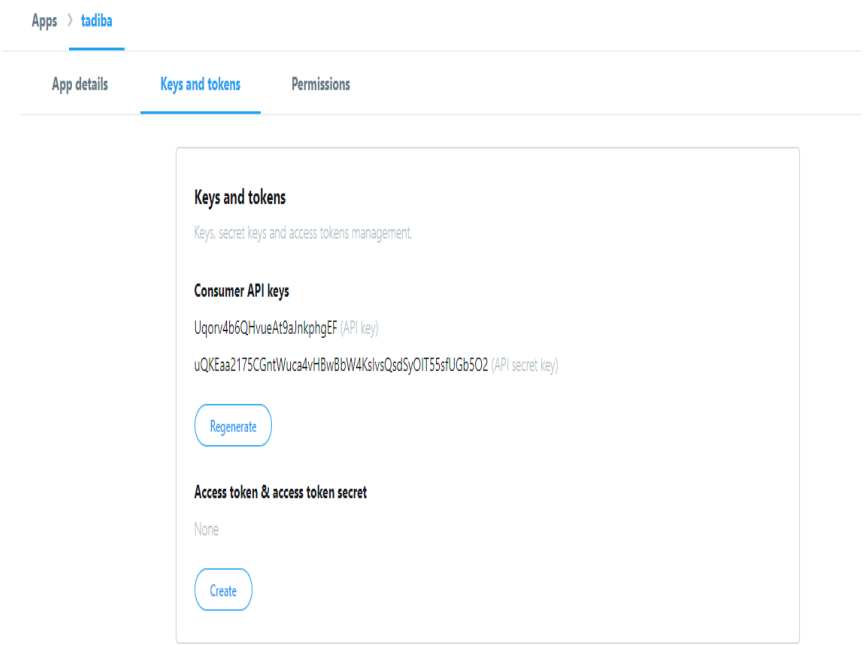
BAB V
PEMBAHASAN

Pada bab ini, akan dijelaskan hasil dari penelitian yang telah dilakukan mulai dari *crawling* data, pembersihan data, pelabelan data hingga analisis *Naive Bayes Classifier* yang digunakan untuk pengklasifikasian mengenai ulasan positif dan ulasan negatif.

5.1. *Authentication*

Untuk melakukan *crawling* data pada *twitter* dibutuhkan sebuah kode yang didapat dari *twitter API* untuk mengakses data *twitter* tersebut. *Twitter API* merupakan aplikasi yang diciptakan oleh pihak *twitter* dengan tujuan agar mempermudah pihak developer untuk mengakses informasi *web twitter*. Pendaftaran *API* digunakan untuk mengonfirmasi kepada pihak *twitter* agar memberikan izin menjelajahi lebih luas terkait dengan data yang berkaitan dengan *twitter*.



Gambar 5. 1 Konfigurasi Pendaftaran API

Setelah registrasi dan bergabung dengan *twitter API* pada tanggal 21 Agustus 2019, dari *twitter API* didapatkan beberapa kode berupa *consumer key*, *consumer secret*, *access token* dan *acces key* dari *twitter*. Kode *API* tersebut adalah sebagai jembatan antara *twitter* dengan aplikasi lainnya, dalam penelitian ini kode tersebut dapat digunakan untuk proses integrasi antara *twitter API* dengan *Python.*

5.2. Pengambilan Data dari *Twitter*

Pengambilan data dari *twitter* menggunakan kode yang didapat dari *keys and tokens* pada *twitter API*, sedangkan kata kunci yang digunakan dalam pencarian data yaitu kata kunci "Jokowi", "#HidupMahasiswa", "#SayaBersamaJokowi dan "#TrisaktiTurunLagi", maka pada *python* digunakan fungsi "`twitter_stream = Stream(auth, MyListener()) twitter_stream.filter(track=['jokowi','#HidupMahasiswa','#SayaBersamaJokowi','#trisaktiTurunLagi'])`"

Tabel 5. 1 Data *Tweet*

No	*Text*	Lokasi	*username*
1	Saya menghargai aspirasi yang disampaikan para mahasiswa dan berbagai elemen masyarakat selama beberapa hari terakhir... .	Bali	sisilianaD1
2	Politik itu rumit, apalagi dimasa transisi seperti ini. Saya yakin Pak Jokowi sedang dalam posisi yang amat pelik... .	Bandung	risma_rsmnrlstr
3	Jokowi hanya ingin berkuasa. Ia tak peduli berapa dan apapun ongkosnya. Termasuk hilangnya kemerdekaan berekspresi, pengembangan nalar kritis... .	Majalengka	Zambulazzam

Pada penelitian yang dilakukan data yang diambil mengenai komentar masyarakat terkait Rancangan Undang-Undang (RUU) yang dianggap kontroversial dengan menggunakan cara *crawling* data *twitter*. Hasil data yang didapat dari *crawling* sebesar 4903 data *tweet* dari media sosial *twitter* pada

tanggal 24-27 Sepetember 2019 yang selanjutnya akan digunakan untuk analisis. Dapat dilihat pada tabel 5.1 terdapat beberapa data yang diperoleh dari proses *crawling* data *twitter* menggunakan *python* seperti *text*, lokasi, dan *username. Text* merupakan komentar-komentar pengguna *twitter* atau yang biasa disebut dengan *tweet*, lokasi merupakan tempat atau daerah dibuatnya *tweet* tersebut dan *username* merupakan nama pengguna atau akun *twitter* yang digunakan untuk menulis komentar tersebut.

5.3. Mengubah Data Kedalam bentuk CSV

Data yang diperoleh dalam proses *scrapping* menggunakan *phyton* memiliki format ".json". agar lebih mudah untuk diolah, data tersebut di*convert* kedalam bentuk ".csv" menggunakan *package* "jsonlite". Kemudian informasi yang diambil dimasukkan pada sebuah data *frame* menggunakan perintah "`mrtgab=data.frame(text=paste(c(k1,k2,k3)),lokasi=paste(c(k4,k5,k6)),username=paste(c(k7,k8,k9)))`" kemudian data tersebut disimpan dalam bentuk ".csv" menggunakan perintah "`write.csv(mrtgab,file="D:\\SKRIPSI\\data full gabungan.csv")`". Terdapat pada lampiran 2.

5.4. *Preprocessing*

Setelah data yang didapatkan sudah dalam bentuk "csv" kemudian dilakukan tahap *preprocessing*, tahap ini bertujuan untuk membersihkan data-data dari *noise* dan pembenahan bahasa seperti menghilangkan singkatan, bahasa gaul, serta menghapus kata yang tidak diperlukan, karena data awal yang didapatkan berupa data yang tidak terstruktur maka dilakukan tahap *preprocessing* agar data tersebut dapat di analisis.

Tabel 5. 2 Contoh Data untuk *Preprocessing*

No.	Data
1.	#PercayakanPadaJokowi\nMoeldoko menyebut sikap DPR belum jelas. DPR perlu mendiskusikan pasal-pasal yang masih perlu didalami. https://t.co/EQRN0qUvMv
2.	Kalau sepatu kotor saja jadi perhatian, bagaimana dengan isikepala wakil rakyat yang kotor? ... https://t.co/lmUjhYrLT6
3.	Dasar rezim ugal-ugalan. Tanpa punya dasar hukum, pemerintah DIAM-

	DIAM membatasi akses ke Twitter dan instagram. Kalau gak mau didemo, makanya kerja yang becus dong @jokowi! #TurunkanJokowi #TolakRUUKUHP #TolakRevisiUUKPK

Dalam Tabel 5.2 menunjukkan beberapa contoh *tweet* terkait sikap atau komentar masyarakat mengenai RUU yang kontroversial baik RKUHP. Berikut merupakan tahapan-tahapan *preprocessing*:

1. *Cleaning* Data

Hasil dari *crawling* merupakan data mentah atau data yang diperoleh masih terdapat unsur simbol, *URL* dan sebagainya yang tidak mempunyai arti pada kalimat tersebut. Hal ini dapat menyulitkan para pembaca untuk menemukan topik atau pembahasan informasi terkait. Dari permasalahan tersebut maka diperlukan proses *cleaning* guna membersihkan data sehingga pembaca dapat mengetahui informasi dengan mudah. Proses *cleaning* data adalah proses untuk merapihkan dan membersihkan kalimat dari kata-kata yang tidak memiliki arti sehingga lebih mudah dan cepat dalam mendapatkan informasi dari data yang didapat. Perintah yang digunakan pada proses *cleaning* yaitu "`removeURL <- function(x) gsub("http[^[:space:]]*", "", x)`

`twitclean <- tm_map(komenc, removeURL)`" yang terdapat pada lampiran 3. Pada proses *cleaning* didapatkan hasil pada Tabel 5.3:

Tabel 5. 3 Proses *Cleaning Data*

Sebelum *Cleaning*	Sesudah *Cleaning*
#PercayakanPadaJokowi Moeldoko menyebut sikap DPR belum jelas. DPR perlu mendiskusikan pasal-pasal yang masih perlu didalami https://t.co/EQRN0qUvMy	#PercayakanPadaJokowi Moeldoko menyebut sikap DPR belum jelas DPR perlu mendiskusikan pasal-pasal yang masih perlu didalami
Kalau sepatu kotor saja jadi perhatian bagaimana dengan isi kepala wakil rakyat yang kotor? ..https://t.co/lmUjhYrLT6	Kalau sepatu kotor saja jadi perhatian bagaimana dengan isi kepala wakil rakyat yang kotor
Dasar rezim ugal-ugalan. Tanpa punya dasar hukum, pemerintah DIAM-DIAM membatasi akses ke Twitter dan instagram. Kalau gak mau di demo, makanya kerja yang becus dong @jokowi	Dasar rezim ugal-ugalan. Tanpa punya dasar hukum pemerintah DIAM-DIAM membatasi akses ke Twitter dan instagram Kalau gak mau di demo makanya kerja yang becus dong, @jokowi #TurunkanJokowi

#TurunkanJokowi #TolakRUUKUHP #TolakRevisiUUKPK	#TolakRUUKUHP #TolakRevisiUUKPK

Setelah melakukan tahapan *cleaning* yang telah disebutkan sebelumnya, ada beberapa tahapan lainnya dalam proses *cleaning* untuk data teks *tweet* pada *twitter*. Tahapan *cleaning* lainnya yang dilakukan seperti penghapusan *URL*, angka dan lain sebagainya seperti pada Tabel 5.3 merupakan contoh *cleaning URL*, kata yang bewarna kuning merupakan kata yang dihapus pada proses *cleaning*.

2. *Case Folding*

Pada tahap *case folding* merupakan tahap pengubahan huruf kapital menjadi huruf non kapital atau semuanya menjadi huruf kecil menggunakan perintah `"twitclean <- tm_map(twitclean, tolower) inspect(twitclean[2:4])"`. Pada Tabel 5.4 dibawah ini terdapat huruf kapital yang bewarna kuning, dimana huruf tersebut yang dirubah pada proses *case folding*.

Tabel 5. 4 Proses *Case folding*

Sebelum *Case Folding*	**Sesudah *Case Folding***
#PercayakanPadaJokowi Moeldoko menyebut sikap DPR belum jelas. DPR perlu mendiskusikan pasal-pasal yang masih perlu didalami	#percayakanpadajokowi moeldoko menyebut sikap dpr belum jelas. Dpr perlu mendiskusikan pasal-pasal yang masih perlu didalami.
Kalau sepatu kotor saja jadi perhatian bagaimana dengan isi kepala wakil rakyat yang kotor?	kalau sepatu kotor saja jadi perhatian bagaimana dengan isi kepala wakil rakyat yang kotor?
Dasar rezim ugal-ugalan. Tanpa punya dasar hukum, pemerintah DIAM-DIAM membatasi akses ke Twitter dan instagram. Kalau gak mau di demo, makanya kerja yg becus dong @jokowi #TurunkanJokowi #TolakRUUKUHP	dasar rezim ugal-ugalan. tanpa punya dasar hukum, pemerintah diam-diam membatasi akses ke twitter dan instagram. kalau gak mau di demo, makanya kerja yg becus dong @jokowi #turunkanjokowi #tolakruukuhp #tolakrevisiuukpk

#TolakRevisiUUKPK	

3. *Filtering*

Tahap *filtering* yaitu tahapan untuk mengambil kata-kata yang penting. Proses *filtering* dapat menggunakan algoritma *stopword* (menghapus kata tidak penting). Contoh *stopword* yaitu "yang", "dan", "ke", "dari","oleh" dan lainnya. Kata-kata tersebut merupakan kata yang berfrekuensi tinggi dan dapat ditemukan di hampir setiap kalimat. *Stopword* atau menghapus kata dapat mengurangi ukuran indeks dan waktu pemrosesan serta dapat mengurangi *noise*. Pada tahap *filtering* menggunakan perintah `"twitclean <- tm_map(twitclean , removeWords, readLines("D:/SKRIPSI/kata.txt")) inspect(twitclean[2:4])"`. Pada Tabel 5.5 terdapat kata yang bewarna kuning, dimana kata tersebut yang dihapus pada tahap filtering.

Tabel 5. 5 Proses *Filtering*

Sebelum *Filtering*	**Sesudah *Filtering***
#percayakanpadajokowi moeldoko menyebut sikap dpr belum jelas. dpr perlu mendiskusikan pasal-pasal yang masih perlu didalami.	#percayakanpadajokowi moeldoko menyebut sikap dpr. dpr mendiskusikan pasal-pasal didalami.
kalau sepatu kotor saja jadi perhatian bagaimana dengan isi kepala wakil rakyat yang kotor?	sepatu kotor perhatian, isi kepala wakil rakyat kotor?
dasar rezim ugal-ugalan. tanpa punya dasar hukum, pemerintah diam-diam membatasi akses ke twitter dan instagram. kalau gak mau di demo, makanya kerja yg becus dong @jokowi #turunkanjokowi #tolakruukuhp #tolakrevisiuukpk	dasar rezim ugal-ugalan. dasar hukum, pemerintah diam-diam membatasi akses twitter instagram. di demo, kerja yg becus @jokowi #turunkanjokowi #tolakruukuhp #tolakrevisiuukpk

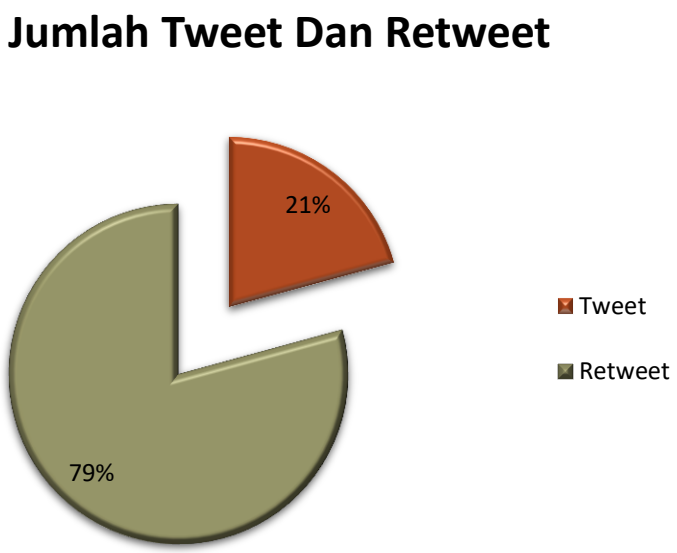
4. *Tokenizing*

Tahapan *tokenizing* adalah proses untuk memisahkan kata di dalam dokumen menjadi potongan kata yang tidak saling berpengaruh yang disebut token untuk kemudian dapat diidentifikasi. Pada tahap *tokenizing* menggunakan *package* "tokenizers". Pada Tabel 5.6 merupakan contoh dari proses *tokenizing*.

Tabel 5. 6 Proses *Tokenizing*

Sebelum *Tokenizing*	Sesudah *Tokenizing*
#percayakanpadajokowi moeldoko menyebut sikap dpr dpr mendiskusikan pasal-pasal didalami	“#percayakanpadajokowi” “moeldoko” “menyebut” “sikap” “dpr” “ dpr” “mendiskusikan” “pasal-pasal” “didalami”
sepatu kotor perhatian isi kepala wakil rakyat kotor	“sepatu” “kotor” “perhatian” “isi” “kepala” “wakil” “rakyat” “kotor”
dasar rezim ugal-ugalan dasar hukum pemerintah diam-diam membatasi akses twitter instagram di demo kerja becus @jokowi #turunkanjokowi #tolakruukuhp #tolakrevisiuukpk	“dasar” “rezim” “ugal-ugalan” “dasar” “hukum” “pemerintah” “diam-diam” “membatasi” “twitter” “instagram” “di” “demo” “kerja” “becus” “@jokowi” “#turunkanjokowi” “#tolakruukuhp” “#tolakrevisiuukpk”

Kemudian dari data yang telah melewati tahap *preprocessing*, dilakukan identifikasi objek terlebih dahulu sebelum di analisis lebih lanjut.

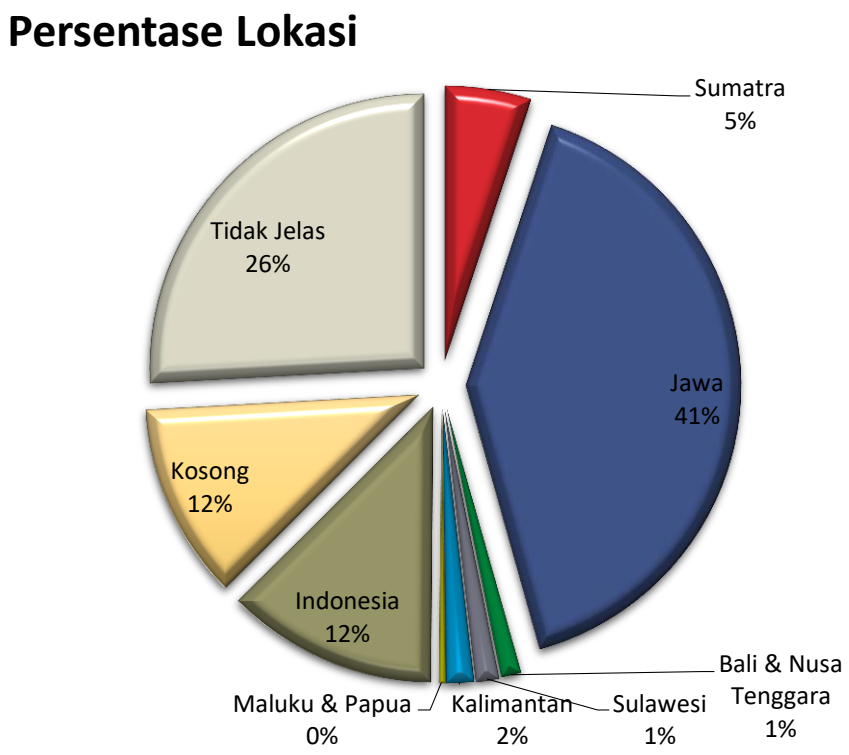


Gambar 5.2 Persentase Tweet Dan Retweet

Terlihat pada Gambar 5.2 dapat diketahui bahwa dari keseluruhan data komentar mengenai RKUHP yang didapat dari *Twitter*, 79% merupakan data *Retweet* dan 21% merupakan data *Tweet.*

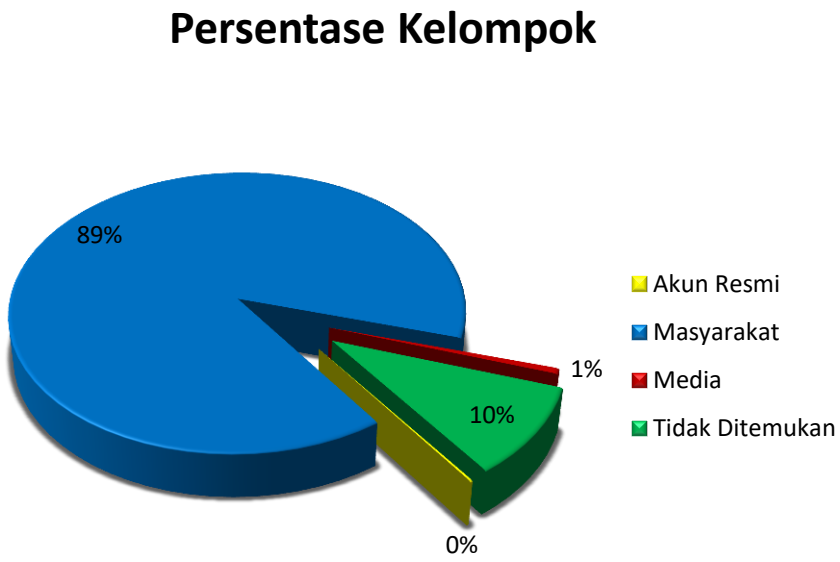
Gambar 5. 3 Persentase Sikap

Pada Gambar 5.3 dapat diketahui dari keseluruhan data terdapat 90% masyarakat yang kontra atau menolak RKUHP, merupakan masyarakat yang pro terhadap perubahan yang dilakukan pada RKUHP atau masyarakat yang membela Jokowi atas tindakan demonstrasi yang dilakukan oleh mahasiswa sebesar 4% dan masyarakat yang bersikap netral sebesar 6%.



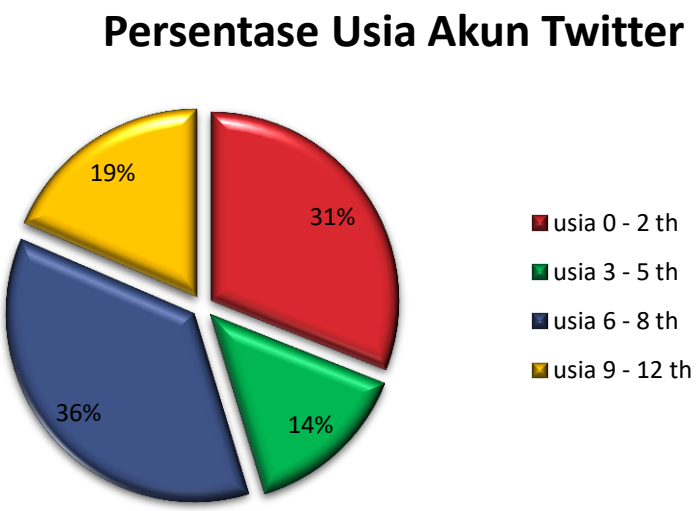
Gambar 5. 4 Persentase Lokasi

Pada Gambar 5.4 dapat diketahui persentase paling besar adalah Jawa dengan jumlah 41%, hal ini berarti masyarakat yang turut berkomentar di *twitter* mengenai RKUHP sebagian besar berada di Pulau Jawa. Persentase terbesar ke dua setelah Jawa adalah Tidak Jelas dengan persentase sebesar 26%. Pada lokasi yang Tidak Jelas ini berarti pemilik akun tidak mencantumkan lokasi yang jelas pada profil *twitter* mereka, contohnya adalah dibumi, dirumah, dikasur dan lokasi-lokasi lain yang tidak menjelaskan suatu wilayah. Pulau Maluku dan Papua menempati posisi terakhir dengan persentase sebesar 0%, yang berarti bahwa tidak banyak masyarakat yang berkomentar mengenai RKUHP di wilayah tersebut.



Gambar 5. 5 Persentase Kelompok

Pada Gambar 5.5 dapat diketahui bahwa masyarakat memiliki persentase terbesar, dengan jumlah 89%. Hal ini menunjukkan bahwa kelompok masyarakatlah yang paling banyak berkomentar di *twitter* mengenai RKUHP. Terdapat persentase sebesar 10% pada bagian Tidak Ditemukan, hal ini berarti bahwa akun-akun yang digunakan pada saat berkomentar tidak dapat ditemukan kembali pada saat proses identifikasi. Terdapat beberapa faktor bagaimana hal ini dapat terjadi salah satunya adalah pengguna sudah menghapus akun yang digunakan untuk berkomentar.



Gambar 5. 6 Persentase Usia Akun

Pada Gambar 5.6 dapat diketahui bahwa usia akun *twitter* yang digunakan untuk berkomentar mengenai RKUHP paling banyak berkisar 6 sampai 8 tahun dengan persentase sebesar 36%, hal ini tidak jauh berbeda dengan usia akun 0 sampai 2 tahun yang memiliki persentase sebesar 32%.

5.5. TF-IDF

Pembobotan pada setiap kata yang terdapat dalam dokumen harus dilakukan dalam metode klasifikasi terhadap data yang berbentuk teks yang. Teknik pembobotan kata tersebut yaitu term *frequency-inverse documen frequency* (TF IDF). Nilai dari TF IDF dipengaruhi oleh kemunculan kata dalam suatu dokumen dan jumlah kata ataupun frekuensi kata yang muncul secara keseluruhan. Berikut merupakan contoh kalimat yang telah di proses pada tahap *pre-processing* yang akan dihitung pembobotan kata "agenda pelengseran presiden mahasiswa tuntut undangundang kpk rkuhp bermasalah dibatalkan", kalimat tersebut merupakan dokumen ke-10.

Tahapan awal melakukan pembobotan menggunakan TF-IDF yaitu mencari *inverse document frequency* (idf) yaitu jumlah dokumen dibagi dengan jumlah kata yang muncul dalam keseluruhan dokumen. Berikut merupakan tabel jumlah

kata yang muncul dalam setiap dokumen berdasarkan kata diatas dan perhitungan IDF dapat dilihat pada Tabel 5.7.

Tabel 5. 7 *Contoh Perhitungan IDF*

Kata/ Dokumen	**Urutan *Tweet***					**Jumlah**	$IDF = \log\left(\frac{N}{df}\right)$
	1	**...**	**10**	**...**	**3561**		
Agenda	0	...	1	...	0	27	$\log\left(\frac{3561}{27}\right) = 2,12$
Pelengseran	0	...	1	...	0	33	$\log\left(\frac{3561}{33}\right) = 2,03$
Presiden	0	...	1	...	0	2029	$\log\left(\frac{3561}{2029}\right) = 0,24$
Mahasiswa	0	...	1	...	0	1293	$\log\left(\frac{3561}{1293}\right) = 0,44$
Tuntut	0	...	1	...	0	35	$\log\left(\frac{3561}{35}\right) = 2,00$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Undang-Undang	0	...	1	...	0	44	$\log\left(\frac{3561}{44}\right) = 1,91$
KPK	0	...	1	...	0	233	$\log\left(\frac{3561}{233}\right) = 1,18$
RKUHP	0	...	1	...	0	176	$\log\left(\frac{3561}{176}\right) = 1,31$
Bermasalah	0	...	1	...	0	71	$\log\left(\frac{3561}{71}\right) = 1,70$

Dibatalkan	0	...	1	...	0	78	$\log\left(\frac{3561}{78}\right)=1,66$

Berdasarkan Tabel 5.7 dapat diketahui bahwa jumlah kata “Agenda” pada keseluruhan dokumen adalah 27 dokumen, Sehingga didapatkan nilai IDF untuk kata “Agenda” yaitu 2,13. Perhitungan yang sama juga dilakukan untuk kata yang lain dalam dokumen. Kemudian setelah mendapatkan nilai IDF maka selanjutnya melakukan perhitungan nilai TF yaitu banyaknya kata dalam sebuah dokumen dibagi jumlah kata keseluruhan yang terdapat dalam sebuah dokumen. Dapat dilihat pada Tabel 5.8 berikut:

Tabel 5. 8 Contoh Perhitungan TF

Kata/ Dokumen	**Urutan *Tweet***				
	1	**...**	**10**	**...**	**3561**
Agenda	0	...	$\frac{1}{10}=0,1$	...	0
Pelengseran	0	...	$\frac{1}{10}=0,1$	...	0
Presiden	0	...	$\frac{1}{10}=0,1$	...	0
Mahasiswa	0	...	$\frac{1}{10}=0,1$	...	0
Tuntut	0	...	$\frac{1}{10}=0,1$	...	0
⋮	⋮	⋮	⋮	⋮	⋮
Undang-Undang	0	...	$\frac{1}{10}=0,1$	...	0
KPK	0	...	$\frac{1}{10}=0,1$	...	0
RKUHP	0	...	$\frac{1}{10}=0,1$	...	0
Bermasalah	0	...	$\frac{1}{10}=0,1$	...	0
Dibatalkan	0	...	$\frac{1}{10}=0,1$	...	0
Jumlah	0	...	10	...	0

Berdasarkan Tabel 5.8 dapat diketahui bahwa kata “Tuntut” pada dokumen 10 berjumlah 1 dan jumlah keseluruhan kata dalam dokumen 10 tersebut adalah 10 kata. Sehingga didapatkan nilai TF untuk kata “semoga” yaitu 1/10 = 0,1. Perhitungan yang sama juga dilakukan untuk kata yang lain dalam dokumen. Setelah mendapatkan nilai TF dan IDF, kemudian akan menghitung nilai TF-IDF. Nilai TF-IDF diperoleh dengan mengalikan nilai TF dengan nilai IDF. Perhitungan TF-IDF dapat dilihat pada Tabel 5.9:

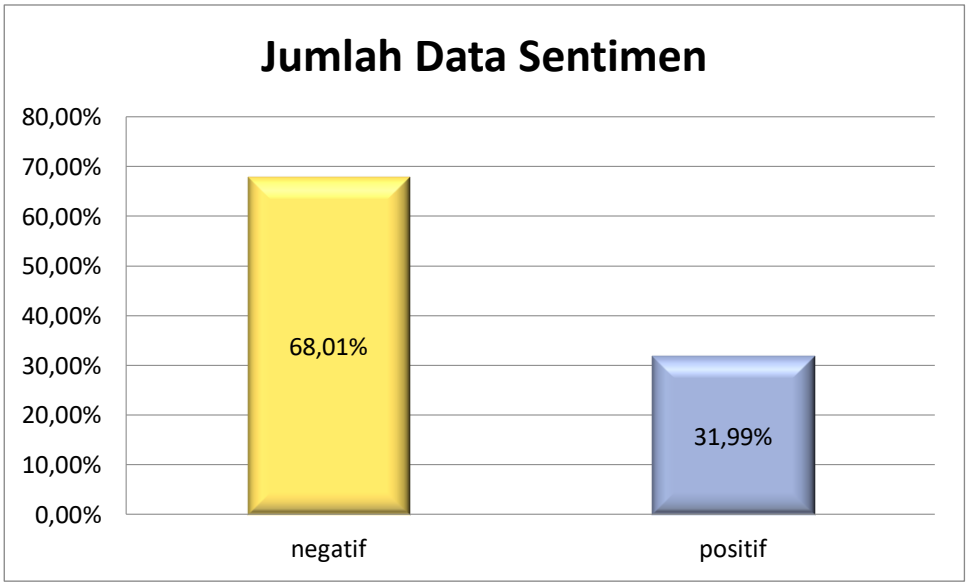
Tabel 5. 9 Contoh Perhitungan TF-IDF

Kata/ Dokumen	**Urutan *Tweet* (TF)**					**IDF**	**TF-IDF**				
	1	**...**	**10**	**...**	**3561**		**1**	**...**	**10**	**...**	**3561**
Agenda	0	...	0,1	...	0	2,12	0	...	0,21	...	0
Pelengseran	0	...	0,1	...	0	2,03	0	...	0,20	...	0
Presiden	0	...	0,1	...	0	0,24	0	...	0,02	...	0
Mahasiswa	0	...	0,1	...	0	0,44	0	...	0,04	...	0
Tuntut	0	...	0,1	...	0	2,00	0	...	0,20	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Undang-Undang	0	...	0,1	...	0	1,91	0	...	0,19	...	0
KPK	0	...	0,1	...	0	1,18	0	...	0,11	...	0
RKUHP	0	...	0,1	...	0	1,31	0	...	0,13	...	0
Bermasalah	0	...	0,1	...	0	1,70	0	...	0,17	...	0
Dibatalkan	0	...	0,1	...	0	1,66	0	...	0,16	...	0

5.6. Analisis Sentimen

Setelah proses *preprocessing*, maka dilanjutkan dengan proses pelabelan pada kelas sentimen. Pada proses pelabelan dibagi menjadi dua kelas sentimen, yaitu sentimen positif dan sentimen negatif. Penilaian dokumen yang masuk kategori kelas sentimen positif dan negatif ditentukan dengan memanfaatkan kumpulan kata bahasa Indonesia yang terdiri dari kumpulan kata-kata positif dan kumpulan

kata-kata negatif. Berdasarkan kumpulan kata berbahasa Indonesia tersebut kemudian dilakukan pelabelan otomatis menggunakan aplikasi R dengan cara menghitung skor jumlah kata positif dikurangi dengan skor jumlah kata negatif dalam suatu kalimat ulasan. Jika suatu kalimat memiliki skor > 0 akan di klasifikasikan dalam kelas positif dan jika suatu kalimat memiliki skor < 0 maka akan diklasifikasikan dalam kelas negatif. Berikut merupakan hasil perbandingan jumlah data dari pelabelan kelas sentimen pada Gambar 5.6:



Gambar 5. 7 Banyaknya Tanggapan Pengguna *Twitter*

Klasifikasi data pada penelitian ini dibagi menjadi sentimen positif dan sentimen negatif. Pada ulasan klasifikasi yang mengandung pernyataan positif seperti ungkapan terimakasih, pujian, dukungan, dan lainnya. Untuk ulasan klasifikasi yang mengandung pernyataan negatif seperti ketidakpuasaan, cacian, ketidaksetujuan, dan lainnya. Pada Gambar 5.7 didapatkan jumlah sentimen negatif sebesar 2483 atau sebesar 68,01% dan sentimen positif sebesar 1078 atau sebesar 31,99%. Dengan perolehan hasil pada Gambar 5.7 dapat dikatakan bahwa sentimen negatif lebih banyak dibandingkan sentimen positf.

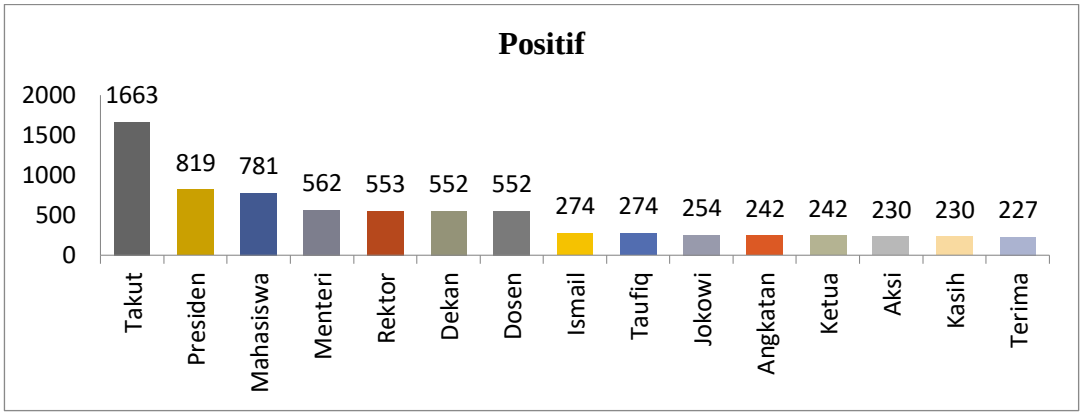
5.7. *Word Cloud*

Tahapan selanjutnya yang dilakukan adalah *word cloud*. Pada *word cloud* menggunakan library "*wordcloud*" dan "*RColorBrewer*". *Word cloud* merupakan

sebuah representasi untuk melihat visualisasi kata-kata pada data teks yang didapatkan setelah proses *cleaning*. Fungsi dari *word cloud* untuk memunculkan kata-kata yang sering digunakan sehingga memudahkan pembaca dalam memahami informasi. Hasil *word cloud* dapat dilihat pada Gambar 5.7 berikut:

Gambar 5. 8 Tampilan *Word Cloud* Sentimen Positif

Dari hasil *word cloud* pada Gambar 5.8 dapat dilihat bahwa kata yang sering atau banyak digunakan pada *tweet* sentimen positif terkait dengan RKUHP adalah takut, presiden dan mahasiswa. Semakin sering dan banyak kata digunakan maka semakin besar ukuran kata tersebut pada *word cloud*. Hasil yang menarik dari tampilan *word cloud* juga dapat memudahkan untuk membaca dan mengetahui informasi terkait.



Gambar 5. 9 Tampilan *Bar Plot* Sentimen Positif

Selain dari *word cloud* Gambar 5.8 dapat dilihat juga pada Gambar 5.9 yang merupakan *bar plot* 15 kata teratas yang sering atau banyak digunakan pada *tweet* sentimen positif terkait dengan RKUHP. Pada *bar plot* tersebut dapat diketahui jumlah kata yang paling banyak digunakan adalah kata takut dengan jumlah 1663 yang kemudian diikuti kata presiden dengan jumlah 819 dan pada urutan ke-15 adalah kata terima dengan jumlah 227.

Tabel 5. 10 Asosiasi Kata "Takut"

Kata yang berasosiasi	**Nilai**
Dekan	1.00
Dosen	1.00
Rektor	1.00
Ismail	0.99
Menteri	0.99
Taufiq	0.99
Presiden	0.84
mahasiswa	0.78

Dapat dilihat pada Gambar 5.8 dan Gambar 5.9, terdapat beberapa kata yang memiliki frekuensi cukup besar. Kata terbanyak atau terbesar pada visualisasi *word cloud* adalah kata "takut". Kata takut memiliki hubungan dengan beberapa kata seperti dekan, dosen, rektor, ismail, menteri, taufiq, presiden dan mahasiswa. Dapat dilihat pada Tabel 5.10 nilai korelasi terhadap kata "takut".

Menurut (Sarwono, 2006) pada penelitian (Burhanuddin, 2012) hasil asosiasi dapat diketahui kekuatan hubungan antar dua kata yang saling berhubungan dengan kisaran nilai -1 s/d 1. Ada beberapa kategori nilai korelasi yang digunakan sebagai berikut :

0 : Tidak ada korelasi antara dua variabel

>0 - 0,25 : Korelasi lemah

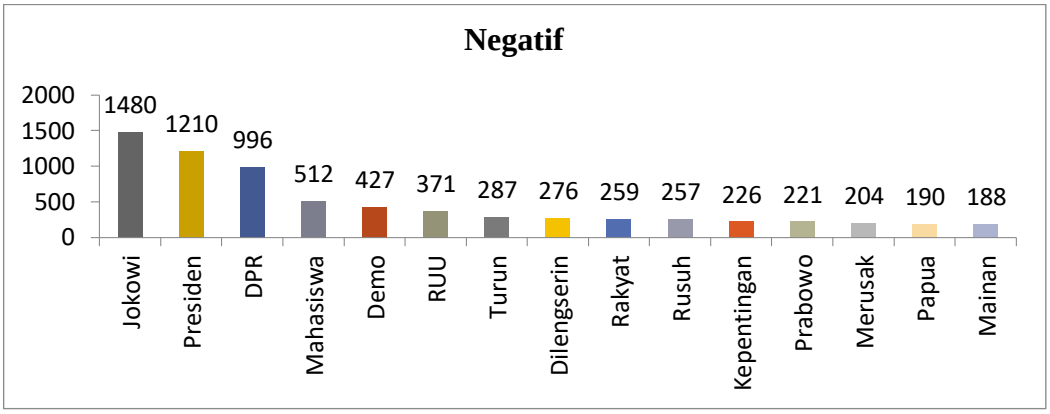
>0,25 - 0,5 : Korelasi cukup

>0,5 – 0,75 : Korelasi kuat

1 : Korelasi sangat kuat

Gambar 5. 10 Tampilan *Word Cloud* Sentimen Negatif

Dari hasil *word cloud* pada Gambar 5.10 dapat dilihat bahwa kata yang sering atau banyak digunakan pada *tweet* terkait dengan pasal-pasal yang dianggap kontroversial pada RKUHP adalah Jokowi, Presiden dan DPR. Semakin sering dan banyak kata digunakan maka semakin besar ukuran kata tersebut pada *word cloud*. Hasil yang menarik dari *word cloud* juga dapat memudahkan untuk membaca dan mengetahui informasi terkait.



Gambar 5. 11 Tampilan *Bar Plot* Sentimen Negatif

Selain dari *word cloud* Gambar 5.10 dapat dilihat juga pada Gambar 5.11 yang merupakan *bar plot* 15 kata teratas yang sering atau banyak digunakan pada *tweet* sentimen negatif terkait dengan RKUHP. Pada *bar plot* tersebut dapat diketahui jumlah kata yang paling banyak digunakan adalah kata Jokowi dengan jumlah 1480 yang kemudian diikuti kata presiden engan jumlah 1210 dan pada urutan ke-15 adalah kata mainan dengan jumlah 188.

Tabel 5. 11 Asosiasi Kata "Jokowi"

Kata yang berasosiasi	**Nilai**	**Kata yang berasosiasi**	**Nilai**	**Kata yang Berasosiasi**	**Nilai**
Prabowo	0,38	Makassar	0,25	Mayoritas	0,19
Dikritik	0,36	Serukan	0,25	Tuntut	0,18
Barusan	0,35	Kesalahan	0,25	Malu	0,18
Nyesel	0,35	Becus	0,24	Rakyat	0,17
Simple	0,35	Lengser	0,24	Krn	0,17
Track	0,35	Kritik	0,24	Dll	0,17
Khan	0,34	Dgn	0,24	Bayarin	0,16
Dituntut	0,33	Hausrakus	0,23	Door	0,16
Memimpin	0,33	Langsung	0,23	Pancasila	0,16
Record	0,33	Menurunkan	0,23	Polamp	0,16
Takut	0,31	Pelantikan	0,23	Provokatoramp	0,16
Mas	0,29	Mahasiswa	0,21	Tangkap	0,16

Sumbar	0,28	Cinta	0,21	Suara	0,16
BEM	0,27	Haus	0,21	Dipercepat	0,15
Pilih	0,27	Knp	0,21	Eeh	0,15
Ratusan	0,26	Lakukan	0,21	Logika	0,15
Berhadapan	0,26	Mengkritiknya	0,21	Makasih	0,15
Memilih	0,26	Tersandera	0,21	Mencerna	0,15
Menggagalkan	0,26	Berharap	0,20	Nyampe	0,15
Mundur	0,25	Citra	0,20	Rukuhp	0,15

Dapat dilihat pada Gambar 5.10 dan Gambar 5.11, terdapat beberapa kata yang memiliki frekuensi cukup besar. Kata terbanyak atau terbesar pada visualisasi *word cloud* sentimen negatif adalah kata "Jokowi". Kata Jokowi memiliki hubungan dengan beberapa kata seperti Prabowo, dikritik, barusan hingga kata RUKUHP, seperti yang dapat dilihat pada Tabel 5.11**.** Pada Tabel 5.11 juga dapat dilihat nilai korelasi terhadap kata "jokowi".

Menurut (Sarwono, 2006) pada penelitian (Burhanuddin, 2012) hasil asosiasi dapat diketahui kekuatan hubungan antar dua kata yang saling berhubungan dengan kisaran nilai -1 s/d 1. Ada beberapa kategori nilai korelasi yang digunakan sebagai berikut :

0 : Tidak ada korelasi antara dua variabel
>0 - 0,25 : Korelasi lemah
>0,25 - 0,5 : Korelasi cukup
>0,5 – 0,75 : Korelasi kuat
1 : Korelasi sangat kuat

5.8. Klasifikasi *Naive Bayes*

Dalam penelitian ini, pembuatan data latih sangat diperlukan. Data latih dapat mempengaruhi tingkat akurasi yang dihasilkan. Data uji merupakan data yang digunakan untuk mengujji tingkat akurasi dari model yang dibuat oleh data latih. Pembuatan data latih dilakukan dengan menentukan proporsi data yang telah melewati proses sebelumnya. Total data keseluruhan sebesar 3561 yang terdiri

dari data *tweet* dan *retweet*. Peneliti menggunakan proporsi 80% untuk data latih dan 20% untuk data uji. Dari kedua proporsi data tersebut diperoleh:

Tabel 5. 12 Pembagian Data *Training* dan *Testing*

	Data *Training*	Data *Testing*
Ratio	80%	20%
Jumlah	2849	712

Data latih = 80% x 3561

= 2849

Data uji = 20% x 3561

= 712

Dari perhitungan proporsi untuk menentukan jumlah data latih dan data uji, diperoleh jumlah data yang digunakan untuk data latih sebesar 2849 data dan untuk data uji yang diperoleh sebesar 712 data.

Analisis *Naive Bayes Classifier* dilakukan setelah menentukan data *testing* dan data *training*. Untuk pembagian data *testing* dan data *training* pada analisis NBC dapat dilihat pada Tabel 5.12. Analisis NBC menghasilkan *confusion matrix*. Klasifikasi pada analisis NBC sebelumnya harus memiliki *prior probabilities*, yang merupakan komponen utama pada konsep *Naive Bayes*. Nilai *prior probabilities* dapat dilihat pada Tabel 5.13.

Tabel 5. 13 *Pior Probabilities*

Positif	**Negatif**
0,27	0,73

Berdasarkan Tabel 5.13 telah didapatkan hasil *prior probabilities. Prior probabilities* merupakan tahapan untuk mencari nilai probabilitas pada masing-masing pengamatan yang akan menghasilkan klasifikasi, dan dapat dikatakan bahwa nilai prior adalah nilai suatu peluang kejadian. Pada Tabel 5.13 dapat dilihat bahwa nilai peluang prior pada kategori positif adalah sebesar 0,27, yang artinya terdapat peluang kejadian sebesar 27%. Nilai tersebut didapatkan dari

banyaknya jumlah pada kategori kelas positif dibagi dengan total dari seluruh data. Persamaannya dapat ditulis sebagai berikut:

$$Prior\_Probabilities\_Positif = \frac{n(Positif)}{Keseluruhan\_Data} = \frac{767}{2849} = 0,27$$

Nilai *prior probabilities* pada kategori negatif didapatkan hasil sebesar 0,73 yang artinya terdapat peluang kejadian sebesar 73%. Nilai prior tersebut didapatkan dari banyaknya jumlah pada kategori kelas negatif dibagi dengan total keseluruhan data, sehingga didapatkan persamaan sebagai berikut:

$$Prior\_Probabilities\_Negatif = \frac{n(Negatif)}{Keseluruhan\_Data} = \frac{2082}{2849} = 0,73$$

Tabel 5. 14 Perbandingan Data Aktual dan Prediksi

No.	Aktual	Prediksi
1.	Positif	Negatif
2.	Positif	Negatif
3.	Negatif	Negatif
4.	Negatif	Negatif
5.	Negatif	Negatif
6.	Positif	Negatif
7.	Positif	Positif
8.	Positif	Negatif
9.	Negatif	Negatif
10.	Negatif	Negatif
⋮	⋮	⋮
712.	Positif	Negatif

Pada Tabel 5.14 merupakan tabel perbandingan data aktual dengan data prediksi. Pada baris 1 terdapat *miss clasification* karena sentimen positif terprediksi menjadi sentimen negatif atau dapat disebut dengan *false negatif* (FN). Pada baris ke 5, memiliki prediksi yang benar, dimana sentimen negatif diprediksi menjadi sentimen negatif juga atau dapat di sebut dengan *true negatif* (TN).

Tabel 5. 15 *Confusion Matrix NBC*

Prediksi	Aktual	
	Positif	Negatif
Positif	169	4
Negatif	45	494

Dilihat dari Tabel 5.15 data aktual sentimen positif didapatkan prediksi yang tepat sesuai dengan kategorinya dari data uji tersebut adalah sebesar 169. Artinya terdapat 169 data yang diprediksi oleh mesin atau model dengan tepat dan tidak terjadi *miss clasification.* Tetapi dari total 214 data pada sentimen positif terdapat 45 data sentimen positif yang terprediksi pada sentimen negatif. Pada kategori sentimen negatif terdapat 494 data yang diprediksi sesuai dengan data aktual. Tetapi terdapat 4 data sentimen negatif yang terprediksi kedalam sentimen positif.

$$Recall = \frac{TP}{TP + FN} \times 100\% = \frac{169}{169 + 45} \times 100\% = 0,789 \times 100\% = 78,9\%$$

Recall merupakan tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi. Dari perhitungan *recall* didapatkan hasil sebesar 78,9%, hal ini berarti tingkat keberhasilan sistem cukup bagus dalam menemukan kembali sebuah informasi.

$$Precision = \frac{TP}{TP + FP} \times 100\% = \frac{169}{169 + 4} \times 100\% = 0,976 \times 100\% = 97,6\%$$

Precision (presisi) merupakan tingkat ketelitian atau ketepatan dalam klasifikasi. Pada perhitungan presisi didapatkan hasil sebesar 97,6%, hal ini berarti tingkat ketelitian atau ketepatan dalam klasifikasi dapat dikatakan bagus dan nyaris sempurna.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% = \frac{169 + 494}{169 + 4 + 494 + 45} \times 100\% = 0,931 \times 100\% = 93,1\%$$

Accuracy (akurasi) digunakan untuk mengetahui seberapa bagus model bias mengklasifikasikan data dengan benar. Pada perhitungan akurasi didapatkan nilai sebesar 93,1%, dari hasil tersebut dapat dikatakan bahwa model bias dapat mengklasifikasikan data dengan benar.

$$Spesificity = \frac{TN}{TN + FP} \times 100\% = \frac{494}{494 + 4} \times 100\% = 0,991 \times 100\% = 99,1\%$$

$$FPR = 1 - Spesificity = 1 - 0,991 = 0.009$$

Spesificity digunakan untuk mengukur proporsi negatif yang benar diidentifikasi. Dari hasil perhitungan didaptkan nilai sebesar 99,1% hal ini dapat diartikan bahwa identifikasi yang dilakukan untuk mengukur proporsi negatif sudah bagus.

$$AUC = \frac{1 + Recall - FPR}{2} = \frac{1 + 0,789 - 0,009}{2} = 0,89$$

Nilai AUC digunakan untuk mengukur kinerja deskriminatif menggunakan perkiraan probabilitas hasil dari sampel yang telah dipilih secara acak dari suatu populasi negatif dan positif. Klasifikasi dikatakan baik jika nilai AUC semakin tinggi. Dari perhitungan nilai AUC didapatkan hasil sebesar 0,89 yang berarti bahwa klasifikasi baik.